Modeling Responsibility Attribution in a Group

Thesis by
Yishan Zhou

Department of Psychology
In fullfillment of B.S. in Cognitive Science

CARNEGIE MELLON UNIVERSITY
Pittsburgh, PA

2019

ACKNOWLEDGEMENTS

This work would not have been made possible without my advisor, Dr. David Danks. It was a pleasure to work with him, and I greatly appreciate his guidance and feedback on this thesis. His patience and insight encouraged me to think more critically and independently, which I am especially grateful for.

I would like to thank the Department of Psychology and the Department of Philosophy for their support, as well as all the subjects who volunteered to participate in this study.

I would also like to thank Dr. Charles Kemp for introducing me to cognitive modeling, and Dr. Kun Zhang and Dr. Clark Glymour for introducing me to causality.

Finally, I would like to thank the inhabitants of Math Lounge for inspirations and mental support.

ABSTRACT

Attributing responsibilities to a group of agents who contribute to a collective cause is common in our daily life. While we as humans could relatively efficiently assign responsibility to others using our intuition, the wide range of possible scenarios and complexity of our cognitive mechanisms should not be overlooked. Understanding of how rational responsibility judgments are made could expand our knowledge about human causal and moral reasoning. Over recent years, a few theories have been developed to account for responsibility attribution, but they have various theoretical and empirical limitations. In this thesis, an extended computational model is proposed in an effort to expand the explanatory power of previous models. We devised two sets of experiments to test the performance of this new model, especially with regards to aspects including (1) epistemic state of rater, (2) perceived autonomy of agents, and (3) value range of contributions. Our results suggest that the new model demonstrated reasonable power in predicting responsibility attribution, albeit with several unexpected inaccuracies. The close relationship among responsibility, criticality, and pivotality is nonetheless supported.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

A team of ten acquaintances is voting to elect its leader. They agreed on using the simple majority rule, and everyone has to nominate exactly one person (no abstention allowed). In the end, a team member named Aditya received the most votes and was consequently elected to be the leader. However, it later turned out that Aditya, though a nice and understanding guy, was irresponsible, which caused great difficulty for effective team collaboration. Now, are any of the team members other than Aditya responsible for this unfortunate outcome? If so, is anyone more responsible than others?

Assigning responsibility to others like this is part of our ordinary life, while such responsibility attribution can be very much non-trivial. Often we find ourselves in a group, contributing as individuals to some collective outcome (e.g. electing a leader). In return, we are held responsible for our action with respect to the actual outcome (e.g. electing an irresponsible leader). Responsibility attribution could appear intuitive in some cases. For instance, if Aditya had a marginal win, the ones (and only those) who voted Aditya may be perceived fully responsible for the teamwork disaster, since their action determined the outcome, not any other vote to the non-Aditya candidate. On the other hand, it could be the case that Aditya received way more votes than any other nominated candidates, and the result could not be any different if only one or two people voted someone else. Therefore, we might give every Aditya-supporter less responsibility, since their individual behavior has less impact on the eventual outcome. While quantitative differences may not seem to complicate the scenario much (e.g. having twenty voters as opposed to ten), introducing qualitative complexity could make responsibility attribution less straightforward - for example, when every team member could nominate up to three candidates. Additionally, differences in mindfulness can also complicate responsibility attribution (e.g. one voter decides to mindlessly copy another person's vote, versus according to her own judgment, both leading to the same result). The complexity of different components in responsibility attribution reflects the elaborate cognitive mechanisms we use in a seemingly simple decision-making task.

There are multiple implications of understanding responsibility attribution: first,

causal reasoning has been found to play a key role in responsibility judgment (D. A. Lagnado & Gerstenberg, 2016). It is rational for people to make reference to the effect of actions and cause of an outcome to determine who made relevant contributions to bring about a certain result. A valid causal link between the action of some agent and an outcome provides justification for assigning her higher **(causal) responsibility**. For example, rubbing a match against a rough surface is perceived causally responsible for lighting the match, and someone who put pesticide in food should be held causally responsible for food poisoning of those who ate the contaminated food. Additionally, responsibility judgment involves moral reasoning, especially when contributing agents are human, who often bear **moral responsibilities**. An agent may be judged more responsible if her action is "wrong," compared to if her behavior is more or less neutral but still leads to the same result (Reuter et al., 2014). Note that moral reasoning is not independent of causal reasoning; rather, it has an augmentative effect on the final responsibility judgment (Sytsma, Livengood, & Rose, 2012).[1] Finally, perception of autonomy is also crucial for responsibility judgment. Granted, autonomy and responsibility are two distinct concepts, whose relationship is still under debate (Fischer, 2010). Nonetheless, being able to carry out actions according to one's own motives has certain implications on intuitive responsibility judgment. For instance, we tend to give lower responsibility rating to agents who accidentally brings about a bad outcome than those who intentionally do so. Understanding responsibility attribution to agents of different levels of autonomy could, therefore, help us gain insight into how perceived autonomy and responsibility perception interact with each other.

We will now introduce a few terms for the purpose of this thesis. First, a *game* is defined as a situation in which two or more independent agents/players make individual contributions that collectively result in an outcome according to predefined rules. It will be used interchangeably with "scenario" and "story" throughout this paper in reference to the responsibility attribution tasks. In addition, a *positive contributor* is defined as the agent who makes contributions that facilitates the occurrence of a certain outcome. e.g. if a group of ten elected Aditya as their leader, the positive contributors are (only) the ones who voted for Aditya.

To systematically understand how people attribute responsibility to positive contributors in a game, several computational models have been proposed. Lagnado,

[1] In addition to causal and moral responsibility, role responsibility is another consideration. However, since roles of agents are harder to generalize, for the purpose of this paper, role responsibility is not of primary concern.

Gerstenberg & Zultan (2013)'s Criticality-Pivotality Model (CPM) had high accuracy (correlation strength = 0.90) in predicting and explaining people's responsibility judgment for scenarios where all team members were perceived to contribute simultaneously towards some binary-valued outcome (e.g. win/loss). A few earlier models also attempted at explaining how 1people assign responsibility scores to team members, including Crediting Causality Model (CCM) (Spellman, 1997), and the Structural Model (Chockler & Halpern, 2004).

Nonetheless, all previous models related to responsibility attribution have limited explanatory power considering both their theoretical and empirical shortcomings. CCM has failed to explain team contribution scenarios it theoretically could account for (Reuter et al., 2014). It could not explain why people attribute more responsibility to later agents, whose action only changes the probability of an outcome by little, when the model predicts smaller responsibility in line with the smaller probability change (Mandel, 2003). The scenarios CPM could verifiably explain was only a small subset of real-world examples (e.g. it was not tested against tasks with more than two possible collective outcomes). Additionally, it directly incorporates the Structural Model, which arguably oversimplifies the structure of team contributions. While predictions made by the Structural Model alone share a reasonable correlation with actual human responsibility judgment (correlation strength = 0.77; D. Lagnado, Gerstenberg, & Zultan, 2013), they make strict assumptions about unit of change represented in the original scenario—that is, the amount of change an individual agent could make about her contribution can be quantified using some standard measurement. However, this is not always the case in reality. A more detailed discussion can be found in the Literature Review section below.

Given the complexity of components involved in responsibility attribution tasks, it would be desirable to have a more general computational model that captures some important aspects of relevant reasoning and strategies people typically use in order to have a more comprehensive account of responsibility attribution. The model could then be used to predict and explain a wider range of relevant scenarios, and hopefully shed more light into complicated reasoning processes behind decision-making tasks like responsibility attribution. In particular, this paper intends to focus on a few specific questions while trying to generalize the model:

(1) **To what extent does the epistemic state—namely, whether the rater knows about the final outcome or not—of the responsibility rater affect their responsibility judgment?**

(2) **How are responsibility judgments about human agents different from those about non-human agents, if any?**

(3) **How does the range of possible individual contributions affect people's responsibility judgment?**

Epistemic state refers to the amount of knowledge an agent possesses at a given point in time. In the case of evaluating individual responsibilities in a team, epistemic state of the evaluator is related to the amount of information the evaluator has about each individual contribution (also the expectation/belief the evaluator has about the contributions unknown to her). To manipulate the epistemic state of the responsibility rater, we could reveal some but not all information about actual contributions, and ask the rater to assign responsibility to agents who they have information on before revealing the rest information and asking for a second round of responsibility rating on the same set of agents.

The third question is of particular interest here: previous models have been developed for scenarios with binary-valued contributions (e.g. win vs. loss, voting one candidate vs. voting another), which is not always the case in general responsibility attribution tasks. More importantly, they don't offer reasonable prediction or explanation for scenarios where contributions are associated with integer values, or even real values (a detailed discussion can be found in the Literature Review section below). Being able to capture the relation between responsibility rating and different ranges of possible individual contributions would thus help generalize the model.

The main focus of this thesis is to **(1) propose a more general model to account for responsibility attribution for a wider range of games, and (2) test performance of such model against actual human behavior**. In particular, the generalized model will be based on CPM, but with more generalized definitions of its components. As we will show in later sections, the model was able to maintain the explanatory power of CPM while providing additional insight into more complex games. Before diving into the model proposal, a literature review of some of the most popular models relevant for responsibility attribution will follow.

1.1 Example Game: Aditya Problem

We will use this example scenario throughout the rest of the paper, mainly to illustrate the limitations of existing models for responsibility attribution. Consider the following game, similar to the election scenario described at the beginning

of Introduction: four acquaintances (Aditya, Bart, Ciriaco, Doreen) were voting for their leader using the majority rule. Every person was allowed to vote for at most three candidates among themselves (abstention allowed), and they make their decisions independently (i.e. they can't look at others' vote and change theirs). Aditya voted himself and Bart, Bart voted Aditya and Doreen, Ciriaco voted Aditya, and Doreen abstained. Therefore, Aditya became the leader. However, it later turned out that Aditya, though a nice and understanding guy, was irresponsible, which caused great difficulty for effective team collaboration.

Chapter 2

LITERATURE REVIEW: EXISTING MODELS OF RESPONSIBILITY ATTRIBUTION AND/OR CAUSAL SELECTION

We will review five models relevant for responsibility attributions, including Counterfactual Model, Structural Model, Criticality-Pivotality Model (CPM), Crediting Causality Model (CCM), and Norm Violation Account (NVA). There will be specific emphasis on CPM and CCM, since these two are relatively more comprehensive models of responsibility attribution, and will be extensively compared against our new generalized model.

Counterfactual Model & Structural Model

Based partially on Lewis' work (1973), the Counterfactual Model (as the name suggests) explains the difference in responsibility ratings of different agents in a game using a pure counterfactual strategy. It first identifies the actions that directly caused the outcome using the counterfactual theory of causality (e.g. if some action A alone did not happen, the outcome *would have* been different), and consequently assigns full responsibility to the agents carrying out the said actions. This simple model clearly works in certain cases. For instance, if assassins A and B both shot at a target C, where A missed but B hit C, then B would get full responsibility for C's gunshot, since if her bullet missed C as well, C would not have been injured.

On the other hand, the Counterfactual Model also has obvious limitations. First, it only captures a limited sense of responsibility, since it either assigns full responsibility or none at all. However, we clearly perceive responsibility less as a binary concept. One implication is that, *without any further extension to the Counterfactual Model*, the Counterfactual Model is not able to correctly explain causal overdetermination. Consider the Aditya Problem: since Aditya got three votes, while Bart and Doreen each got one, he could have still won the election and become the leader with two votes, had exactly one of Bart, Ciriaco, or himself not voted for Aditya. Since undoing any single vote for Aditya would not have affected the outcome, the Counterfactual Model would assign zero responsibility to all three Aditya supporters. However, we as humans would at least attribute some degree of responsibility

to the three. The simple Counterfactual Model thus fails at explaining games with **overdetermination**.

Furthermore, the Counterfactual Model does not fully capture the nuance in different degrees of responsibility. In a different example, consider if assassin A's bullet hit target C's head and assassin B's bullet hit C's arm, which collectively led to the death of C, but individual wounds would not (i.e. suffering from only one wound would not have caused C to die). The Counterfactual Model suggests assigning full responsibility of C's death to both A and B. However, intuitively, A should bear more responsibility (or, alternatively, B should be attributed less responsibility) since her action caused more severe damage than B. Additionally, the Counterfactual Model does not consider intention, goal-directedness, or control of agents over their action.

The Structural Model proposed by Chockler & Halpern (2004) is a generalization of the Counterfactual Model to address the issue with causal overdetermination. Specifically, given some action I, the Structural Model assigns responsibility rating to I equal of $\frac{1}{N+1}$, where N is the minimal number of changes that need to be made to the original scenario in order for the outcome of the modified (alternative) scenario to be counterfactually dependent on I. For example, in the overdetermined example given in the discussion of the Counterfactual Mode, Doreen would receive a responsibility rating of $\frac{1}{3}$ for her action because it takes two changes to the original scenario (i.e. two people out of Aditya, Bart, and Ciriaco do not vote Aditya) for Doreen's action to solely determine the outcome of this election. In contrast, Aditya, Bart, and Ciriaco would each receive a responsibility rating of $\frac{1}{2}$, because it only takes one change to the original scenario for their individual action to solely determine the election result. This intuitively makes sense, as Doreen did not vote for Aditya to start with, while the other three did. Thus the Structural Model is able to explain the typical responsibility judgment with respect to overdetermination. However, it does not provide a satisfactory account of scenarios with some perceivable ordering of events. If, say, person A loads bullet into a gun, and B picks up the gun not knowing it's loaded, and B shoots C dead, people tend to find B more responsible for C's death (see a similar example involving an overloaded mainframe by Sytsma, Livengood, & Rose, 2012), even though the outcome *would have* been different had A not loaded the bullet. In contrast, the Structural Model gives A and B the same responsibility rating.

In summary, both the Counterfactual Model and Structural Model provide good primitive explanations for responsibility attribution. However, the basic Counter-

factual Model may be overly simplistic, and cannot fully represent the different degrees of responsibility. The Structural Model, which is essentially a form of generalized Counterfactual Model, fails to explain responsibility judgments made when contributions are represented with some explicit ordering.

Criticality-Pivotality Model (CPM)

Lagnado, Gerstenberg & Zultan (2013) proposed the Criticality-Pivotality Model (CPM) for group responsibility attribution. CPM expects the responsibility rating of an arbitrary agent to be a linear combination of **heuristic criticality** and **structural pivotality** scores of her contribution, plus a constant term.

The heuristic criticality of player A is $\frac{P(\texttt{win} \mid \texttt{player A succeeds}) - P(\texttt{win} \mid \texttt{player A fails})}{P(\texttt{win} \mid \texttt{player A succeeds})}$, or equivalently, $1 - \frac{P(\texttt{win} \mid \texttt{player A fails})}{P(\texttt{win} \mid \texttt{player A succeeds})}$ when both player A's contribution and the team outcome are binary-valued. It is the relative decrease in the probability of team winning when a given player fails, normalized against the probability of the team winning when the player succeeds at contributing positively to the winning outcome. Therefore, heuristic criticality provides a way to quantify the importance of some player's positive contribution for achieving a collective goal. Note that this definition assumes that failure at the level of an individual player decreases the probability of the entire team winning, which is usually the case in real life. For example, in the Aditya Problem, say each person was equally likely to vote Aditya 50% of the time, in addition to 100% probability of whoever they would vote (i.e. Bart for Aditya, Doreen for Bart). Then the heuristic criticality of Bart, with respect to Aditya winning the election, is $1 - \frac{0.5}{0.875} \approx 0.429$.

Structural pivotality is adapted directly from the Structural Model (Chockler & Halpern, 2004), and so the structural pivotality of player A is equivalent to $\frac{1}{N+1}$, where N is the minimal number of changes that need to be made to the original scenario in order for the outcome of the modified (alternative) scenario to be counterfactually dependent on A. For instance, in the Aditya Problem, assume six people voted Aditya, and five voted Doreen. If Bart voted Aditya, his structural pivotality is equal to $\frac{1}{2}$, as it only requires one other Aditya-supporter to change their vote from Aditya to Doreen (while others' vote(s) stay the same) for the outcome to be purely dependent on Bart's decision.

Also, notice that heuristic criticality and structural pivotality correspond nicely to conditions of necessity and sufficiency respectively. Having full heuristic critical-

ity (i.e., equal to 1) implies that the agent's success is necessary for the team to win, as the probability of collective win given failure on the agent's part is zero ($P(\texttt{win} \mid \texttt{player A fails}) = 0$). Having full structural pivotality (i.e., equal to 1) implies that the agent's positive contribution would be sufficient for the team to win, since given all other agent's contribution remain the same ($N = 0$), her action would fully determine the outcome.

CPM has several limitations. The most obvious one is the limited set of game contribution scenarios it can explain with heuristic criticality and structural pivotality. For instance, imagine assassin A and B both stabbed a common target C, and C survived. A stabbed C once while B stabbed C twice. Each of them also carried a loaded gun at the time. The probability of either one stabbing C is 0.5, where the number of stabs follows an exponential distribution, and the probability of using a gun is 0.5. Additionally, say if A and B stab a total number of four times, or if either assassin shoots C with a gun, C *would have* died. First of all, it is unclear how success on the level of an individual player in this scenario should be represented. A stabbing C once is "successful" because it led to C's survival *in this particular scenario*, but it could also be "unsuccessful" had B stabbed C three times instead of two. In other words, heuristic criticality as it is defined does not naturally extend to situations where individual contributions cannot be represented by binary values. Similarly, we cannot divide criticality equally between A and B by considering their contributions in a disjunctive manner (e.g. "if A or B makes a certain contribution, C *would have* survived").

Furthermore, the above assassin example poses a challenge for structural pivotality, which does not handle cases where contributions are not associated with binary values. B could have killed C with one more stab, or used a gun instead of a knife, either of which involves one change to the original scenario. However, the two changes are qualitatively different, and structural pivotality does not capture such a distinction.

Moreover, CPM does not take intention into account, when goal-directedness has a non-trivial effect on people's view on the causal link between an action and an outcome, and consequently their responsibility attribution as well (Alicke, 1992; Spellman, 1997). For instance, say in the Aditya Problem, if Ciriaco didn't know Aditya well and voted him by randomly picking a candidate, his contribution *would have* been accidental. On the contrary, had he know Aditya fairly well and still voted Aditya, his contribution would be deemed intentional. Previous similar studies

have shown that people tend to attribute less responsibility when the agent made a mistake by accident rather than on purpose, such as if a driver got stopped by a police officer for speeding, as opposed to being mistakenly stopped (Spellman, 1997). However, in either case (Ciriaco voting Aditya by accident vs. on purpose), Ciriaco's contribution *would have* remained quantitatively the same, leading to the same responsibility rating under CPM.

Despite its constraints on the type of games (e.g. binary-valued contributions only), especially with regards to values of contributions and eventual outcome, CPM has reasonable explanatory power for the subset of problems it addresses. It is also the main motivation for this thesis.

Crediting Causality Model (CCM)

CCM (Spellman, 1997) makes the explicit link between responsibility rating and the change in probability of outcome due to a particular action. Denote the probability of the actual outcome before a contributing cause i as p_i and the probability of the actual outcome after the contributing cause i happened as $p_i^{'}$. CCM essentially attributes responsibility to agent A and B with contribution i and j respectively by comparing $p_i^{'} - p_i$ and $p_j^{'} - p_j$. Note, however, that there does not need to be a direct linear mapping between the probability changes and responsibility. The changes do not have to be independent, either.

CCM has been shown better at explaining (and predicting) certain responsibility attribution behavior in humans with its explicit representation of the sequence of contributions than some previous responsibility attribution models. Behavioral results from Spellman (1997)'s paper suggest that people's responsibility ratings were consistent with CCM's prediction on games like Coin Toss (Miller & Gunasegaram, 1990) and Multiple-choice Game (Vinokur & Ajzen, 1982), which previous accounts have failed at explaining (Miller & Gunasegaram, 1990).

Nonetheless, people tend to attribute more responsibility (and blame) to later contributions in a temporal (not causal) sequence that brought the overall probability of outcome to one, even when the outcome is pretty much determined before the last action (Mandel, 2003). If a victim who has been fatally poisoned is killed in a car crash, people attributed responsibility of the victim's death to the car rather than the poison. Interestingly, this effect is reportedly only present when the actions involved are physical, so when the actions are all brought about by humans, all of

them tend to be judged as the actual causes (McClure et al., 2007). This discrepancy thus suggests the possible existence of two modes in causal selection for responsibility attribution tasks, corresponding to teleological vs. mechanistic explanations (Lombrozo, 2010) as different types of causal information becomes available. On the other hand, the distinction could also be due to moral considerations (Knobe, 2010). Furthermore, similar to CPM, an agent's intention is not modeled by CCM.

To sum up, CCM could reliably predict human responsibility judgement given sequential contributions to some extent. On the other hand, CCM may be insensitive to information about the way actions are brought about, which poses a non-trivial limitation on the model's performance on games with human players involved.

Norm Violation Account (NVA)

Causal selection involves choosing one or more "true" causes of a certain outcome out of all possible actual causes. Importantly, given a specific context, "true" cause usually refers to a cause of higher salience. This has several implications on the relationship between causal selection and responsibility attribution: since "true" causes are essentially causal factors, they are clearly relevant for causal responsibility attribution. As Bernstein (2017) points out, it is generally true that an agent's moral responsibility with respect to an outcome is also proportional to her actual causal contribution. In addition, salience of a cause is directly related to its perceived significance with respect to the outcome. Perceived importance in turn influences the perceived responsibility of the cause. An important causal factor is more likely to receive higher responsibility rating than other actual causes that are not "true" causes. Therefore, understanding causal selection provides indirect but valuable insight into factors key to responsibility attribution.

As a model of causal selection, NVA (Hitchcock & Knobe, 2009) suggests that only moral judgments of norm violation influence causal selection. In other words, a rational agent would identify actions that violate some form of norm to be the cause of a consequent outcome. A classic example *Pen Case* that NVA can successfully explain is as follows: imagine there are a number of pens in the receptionist's drawer. The rule is that administrative staff can take the pens, but the faculty members should bring their own pen. One day an administrative staff and a professor each took a pen from the drawer. Later that day, the receptionist needs to take an important message but found no pens in her drawer. Typically people found the professor to be responsible for the unfortunate outcome but not the administrative staff, as

the professor clearly violated the norm (i.e. "faculty member should not take pens from the receptionist's drawer") while the administrative staff bears minimal moral burden from taking the pen. Nonetheless, a simple twist of the *Pen Case* reveals the scope of scenarios that NVA could explain: if the professor is a new faculty member in the department and wasn't aware of the rule, people tend to assign her much less responsibility for taking the pen (Reuter et al., 2014), whereas NVA would still give the same prediction as the original *Pen Case* (i.e. the new faculty member is responsible, but not the administrative staff). Thus NVA is insensitive to the way actions are brought about, such as the intention of an agent.

Furthermore, NVA does not directly predict responsibility attribution. In fact, it does not specify how quantitative responsibility ratings should be generated, since NVA is after all a model of causal selection. A natural way to use NVA for responsibility attribution would be to extend it with a responsibility attribution theory, like the Structural Model mentioned above (e.g. assign all "true" causes with a responsibility rating proportional to its Structural Model score). Additionally, as one of its theoretical limitations, NVA does not account for games where no norm is violated.

Although NVA is not specifically developed for responsibility attribution, because of the close connection between responsibility judgment and causal selection, NVA may be incorporated into existing models for consideration of moral responsibility. NVA on its own, as we have discussed above, may be insufficient, since it does not produce any quantitative predictions, not can it be reliably applied to games with nor norm violation.

In conclusion, each of the models discussed in this section provides unique, valuable insight into what underlies human responsibility attribution. They also pose questions about a more comprehensive, general view of responsibility attribution: How are responsibility assignments for humans different from those for non-human objects? Can we remove some of the constraints or assumptions some of these models have, and still more or less maintain its ability to predict and explain responsibility attribution by humans? For instance, CPM assumes that the games all have binary-valued contributions and final outcomes. Given its established accuracy, it would be interesting to see if it could be extended to account for games that do not satisfy such assumption.

Chapter 3

MODEL PROPOSAL

The model we propose here is a natural extension of CPM. In particular, this Extended CPM generalizes some of CPM's assumptions, including the range of contribution values (CPM assumes binary-valued contributions on level of each individual player, either success or failure), the range of final outcome of the team (CPM assumes binary-valued outcome on level of the team, either success or failure), and prior probability of individual contributions. We propose generalizations of the two main quantitative components of CPM—Pivotality and Criticality—as well as introduce a new qualitative component, Intention, in Extended CPM. As discussed in Introduction, intention could make a significant difference for human responsibility judgment even when criticality and pivotality remain the same (e.g. *Pen Case*). Therefore, this extension of CPM also attempts to take intention into account, *but only qualitatively*. We will not show how to do so in a formal, mathematical fashion.

3.1 Generalized Heuristic Pivotality

Heuristic Pivotality (D. Lagnado, Gerstenberg, & Zultan, 2013) of a player refers to the normalized probability difference of a fixed collective outcome solely due to the occurrence of some particular action taken by the player.

Note that there are two sets of probability values involved in Heuristic Pivotality: specifically, the probability of a particular outcome for the team, and the probability of a particular action for some individual player. To differentiate the two, we will denote the former as $P_O(R = r)$, where R is a random variable for collective outcome, and r a particular collective outcome. Similarly, denote the latter as $P^i_A(X_i = a)$, where X_i is a random variable for player i's action, and a a particular action of player i.

Both P_O and P^i_A only reflect estimates of the real probability distributions if no reliable information about the true underlying distribution is given. For example, $P^i_A(X_i = a)$ indicates the responsibility rater's *perceived* probability for player i to carry out action a if the rater knows nothing about player i's strategy. However, if the rater does have an accurate belief about player i's action, $P^i_A(X_i = a)$ should

be the same as the real probability distribution. This concept of estimation is important when predicting heuristic pivotality in prospective versus retrospective responsibility attribution, as more information becomes available and the epistemic state of the responsibility rater changes.

Define A_i to be the action space, or the set of values of player i's actions so that the value of player i's any action corresponds to exactly one element in A_i. In other words, we assume an arbitrary action a player may carry out and all other possible actions of their own are mutually exclusive (`action exclusivity assumption`). Let $|A_i|$ be the cardinality of such set (the number of possible actions). Moreover, if $|A_i| = n$, then we can write A_i as

$$A_i = \{a_1, ..., a_n\}$$

where $a_1, ..., a_n$ represent the n distinct, independent values player i's action could take.

Fix player i with some action $a^* \in A_i$, and let a denote an arbitrary action in the set A_i. The probability for the group to receive a particular outcome r^*, given that player i's action has value a^*, is therefore

$$P_O(R = r^* | X_i = a^*)$$

On the other hand, if player i's action has any value other than a^*, the probability of $R = r^*$ becomes $P_O(R = r^* | X_i \in A_i \setminus \{a^*\})$. Assuming we have access to the prior distribution over A_i (i.e. the probability distribution over player i's all possible actions), whether real or estimated, we then know the values of $P^i_A(X_i = a')$ for all $a' \in A_i$, $a' \neq a^*$.

By definition of Heuristic Pivotality,

$$Pivotality(i) = 1 - \frac{P_O(R = r^* | X_i \in A_i \setminus \{a^*\})}{P_O(R = r^* | X_i = a^*)}$$

where $P_O(R = r^* | X_i \in A_i \setminus \{a^*\})$ denotes the probability that the team achieves result r^* given player i's action is different from a^*, and $P_O(R = r^* | X_i = a^*)$ refers to

the the probability that the team achieves result r^* given player i's action is exactly a^*.

Using Bayes' Rule,

$$\begin{aligned}
&P_O(R = r^* | X_i \in A_i \setminus \{a^*\}) \\
&= \frac{P_O(R = r^*, X_i \in A_i \setminus \{a^*\})}{P^i_A(X_i \in A_i \setminus \{a^*\})} \\
&= \frac{\int_{a \in A_i \setminus \{a^*\}} P_O(R = r^*, X_i = a)}{\int_{a \in A_i \setminus \{a^*\}} P^i_A(X_i = a)} \quad \texttt{(action exclusivity assumption)} \\
&= \frac{\int_{a \in A_i \setminus \{a^*\}} P_O(R = r^* | X_i = a) P^i_A(X_i = a)}{\int_{a \in A_i \setminus \{a^*\}} P^i_A(X_i = a)}
\end{aligned}$$

Thus the generalized heuristic pivotality is

$$Pivotality(i) = 1 - \frac{\int_{a \in A_i \setminus \{a^*\}} P_O(R = r^* | X_i = a) P^i_A(X_i = a)}{P_O(R = r^* | X_i = a^*) \int_{a \in A_i \setminus \{a^*\}} P^i_A(X_i = a)}$$

When $A_i = \{\texttt{succeed}, \texttt{fail}\}$ and $r^* = \texttt{win}$, the generalized heuristic pivotality term is reduced to the simplified version of heuristic pivotality introduced in Lagnado et al.'s paper, namely

$$Pivotality(i) = 1 - \frac{P(\texttt{win} \mid \texttt{player } i \texttt{ fails})}{P(\texttt{win} \mid \texttt{player } i \texttt{ succeeds})}$$

3.2 Generalized Structural Criticality

Structural Criticality (D. Lagnado, Gerstenberg, & Zultan, 2013) of a player is inversely proportional to the amount of counterfactual changes to the actual scenario required for the outcome to be entirely dependent on the player's action. i.e.

$$Criticality(i) = \frac{1}{N + 1}$$

where N is the minimum number of changes required for the collective outcome to be counterfactually dependent on player i's action.

Let C_i be the set of actions made by all n individual contributors **except** player i in the *actual* scenario. i.e. $C_i = \{a_1, ..., a_{i-1}, a_{i+1}, ..., a_n\}$. Define $\tilde{C}_i$ as follows:

$$\tilde{C}_i = \{\tilde{a_1}, ..., \tilde{a_{i-1}}, \tilde{a_{i+1}}, ..., \tilde{a_n} | \forall j . (1 \leq j \leq n \wedge j \neq i)\ \tilde{a_j} \in A_j \wedge \exists a_j \in A_j . a_j \in C_i\}$$

In other words, $\tilde{C}_i$ is an alternative set of actions that could have been made by the other $n-1$ players, or C_i itself, if $a_j = \tilde{a}_j$ for all j.

Furthermore, define

$$|\tilde{C}_i| = \sum_{j \neq i, \tilde{a}_j \in \tilde{C}_i} d_j(a_j, \tilde{a}_j)$$

where $d_j(a_j, \tilde{a}_j)$ computes the value difference between a_j and $\tilde{a}_j$. This difference function should be defined based on the specific action space of player j, as different players could potentially have different action spaces (i.e. there are n such value difference functions, corresponding to each individual contributor).

Criticality is closely related to *influence of action*. Structural Criticality essentially finds the alternative game in which a player's action is influential (i.e. fully determine the outcome), and computes criticality score based on the similarity between such alternative game and the original game. A player is deemed "fully critical" with respect to the original game setting (criticality score equals to 1 under Structural Criticality) if, fixing other players' contribution in the original setting, her action alone fully determines the outcome of the entire team. More critical players take more influential actions; that is, their action could bring out the target outcome with higher probability. For example, under Structural Criticality, a player's criticality score is *higher* iff her action can fully determine the outcome of the entire team with *fewer* counterfactual alterations to other players' action outcome (i.e. smaller difference between the alternative set of actions $\tilde{C}$ and the actual one C). Equivalently, if an otherwise critical player fails to carry out an influential action, the team is more likely to end up with a different, possibly worse outcome. With binary-valued contributions, the range of possible changes in the final outcome due to change to exactly one player's action is limited. The influence of one player thus has a limited range of values, making it simpler to characterize games in which a player's action is influential. However, since our generalized definition of Criticality should be applicable to contributions of binary as well as non-binary values, there could be a much larger - even infinite - number of possible alterations to the final outcome because of exactly one player. We will thus need a new way to find the alternative games in which an agent's action can be regarded influential.

We will use $M(\tilde{C}, a)$ to denote the likelihood that an agent manages to bring about a target outcome in setting $\tilde{C}$ through an action that is *not* a. For a critical agent

who performed a in setting $\tilde{C}$, $M(\tilde{C}, a)$ is probably small. Recall that if an agent has high criticality score, her action has a high influence on the team outcome - that is, a relatively small difference in her action has a big impact on whether a target outcome could be achieved or not. If she does something other than a, the group outcome is unlikely to be the same as when she does a; otherwise she could alter her action to some extent (e.g. slacking off) and still keep the outcome the same. $M(\tilde{C}, a)$ will be formally defined later in this section.

Let R be the random variable for the collective outcome and r^* be the actual collective outcome. Fix player i with some action $a^* \in A_i$ in her action space A_i. In that case:

$$P(R = r^* | \tilde{C}_i, X_i = a^*) - M(\tilde{C}_i, a^*)$$

measures the *influence of action* a^* with respect to fixed outcome r^* and setting $\tilde{C}_i$ based on the difference this particular action makes compared to alternative possible actions of player i. For a "fully critical" agent in setting $\tilde{C}_i$, the term

$$P(R = r^* | \tilde{C}_i, X_i = a^*)$$

is equal to 1 while $M(\tilde{C}_i, a^*)$ is fairly small. Thus the entire expression evaluates to something close to 1, implying higher *influence of action*.

Now we will formally define $M(\tilde{C}_i, a^*)$, the likelihood for player i to do something different than a^* yet still cause outcome r^* to happen in some setting $\tilde{C}_i$. In Extended CPM, $M(\tilde{C}_i, a^*)$ is either *Worst Consequence*:

$$\arg\min_{a \in A_i \setminus a^*} P(R = r^* | \tilde{C}_i, X_i = a)$$

or *Weighted Total Consequence*:

$$\frac{\sum_{a \in A_i \setminus a^*} P(R = r^* | \tilde{C}_i, X_i = a) P(X_i = a | \tilde{C}_i)}{\sum_{a \in A_i \setminus a*} P(X_i = a | \tilde{C}_i)}$$

While *Worst Consequence* only considers the "worst" alternative action that player i *could* have done in terms of maintaining the same outcome, *Weighted Total Consequence* takes all other alternatives player i *would* have chosen. Since both are plausible ways to quantify what humans perceive as influence of an action, we will use each definition to generate predictions, and see which offers a better explanation. Note that the two definitions are equivalent when X_i and R can only take binary

values.

Additionally, there are two possible ways to utilize *influence of action* to compute $SI(a^*, r^*, \tilde{C}_i)$ (SI stands for Sufficiently influential), the condition which should only be satisfied by (possibly alternative) settings $\tilde{C}_i$ that are used for criticality calculation. Such $\tilde{C}_i$ would make player i with action a^* sufficiently influential and critical while maintaining closest distance to the *actual scenario*. The first way is to add an additional free variable ϵ as the threshold of influence:

$$SI(a^*, r^*, \tilde{C}_i) = P(R = r^* | \tilde{C}_i, X_i = a^*) - M(\tilde{C}_i, a^*) > \epsilon$$

Therefore, the influence of a player in a particular context can be characterized in a binary fashion: it either exceeds a predetermined threshold of importance, or falls below said threshold. It does not distinguish between alternative games, even if they require different numbers of counterfactual alterations from the original game, as long as they all fall above or below the same threshold.

Another way is to look for the alternative game setting $\tilde{C}_i$ such that the influence is maximized:

$$SI(a^*, r^*, \tilde{C}_i) = \arg\max_{\tilde{C}_i} P(R = r^* | \tilde{C}_i, X_i = a^*) - M(\tilde{C}_i, a^*)$$

From this perspective, the influence of any player can be characterized by referencing the maximal difference between the *worst* alternative game and the original game in terms of bringing out outcome r^*. This definition of action influence thus does not distinguish between alternative games who do not maximize the difference a^* makes between the original and the alternative game.

While the two definitions are both sensible approaches, for sake of simplicity, the first definition using threshold will be used in the rest of this paper.

The specific $\tilde{C}_i$ we are interested in is the one most similar to the actual setting C_i—that is, $|\tilde{C}_i|$ is minimum—from among all $\tilde{C}_i$ still under consideration:

$$C_{i_{min}} = \texttt{min}(\{\tilde{C}_i | SI(a^*, r^*, \tilde{C}_i)\})$$

Additionally, such $\tilde{C}_i$ allows a^* to be as influential as possible; given the IM function we chose, this means *influence of action* is above a predetermined influence threshold ϵ. Large $|\tilde{C}_i|$ implies that player i's action is only maximally critical with lots of

counterfactual changes to the original scenario. If *action exclusivity assumption* is true, it's unlikely such counterfactual scenario would happen. Thus player i's action would be deemed less critical if it only has a strong influence on the final outcome in some alternative game distant from the original one.

Finally, generalized criticality can be expressed as

$$Criticality(i) = \frac{1}{|C_{i_{min}}| + 1}$$

To see that the generalized criticality can be reduced to (specific) Structural Criticality, WLOG let all action space, as well as the set of collective outcomes, be the same binary set $\{0, 1\}$. Let $r^* = 1$ and $a^* = 1$. Moreover, since both action and outcome are binary-valued, the two implementations of $SI(a^*, r^*, \tilde{C}_i)$ turn out to be equivalent. It follows that

$$C_{i_{min}} = \texttt{min}(\{\tilde{C}_i | \arg\max_{\tilde{C}_i}(P(R = 1|\tilde{C}_i, X_i = 1) - M(\tilde{C}_i, 1))\})$$

$C_{i_{min}}$ is therefore equivalent to the $\tilde{C}_i$ minimally different from C_i such that flipping X_i's value flips the collective outcome as well. Consequently, the difference between C_i and $C_{i_{min}}$ is the minimum number of counterfactual "flips" required on other player's action for the collective outcome to be dependent on X_i, the exact same defintion as N. Thus

$$Criticality(i) = \frac{1}{N + 1}$$

3.3 Intention/Goal-Directedness & Control

There has been little progress on a mathematical formalization of intention, goal-directed behavior, or control, and it is a problem on its own that transcends the scope of this thesis. However, the role that these factors play is clear: as mentioned before, in the *Pen Case*, people tended to judge professor who knew well that they were not supposed to take pens from the drawer for being responsible of pens running out, whereas they gave the new professor who took pens without realizing the internal rules much less responsibility (Reuter et al., 2014).

Additionally, it is unclear whether the effect of intention on responsibility rating is additive or multiplicative. The model thus should take intention into account, though the specific mechanism it does so so far remains unknown.

Chapter 4

EXPERIMENT

Experimental Design & Hypotheses

Non-binary Contributions

One of the main extensions of Extended CPM from CPM is related to the possible range of contributions, relevant to calculations of both heuristic criticality and structural pivotality. While CPM only applies to binary-valued contributions, Extended CPM could be used on games where contributions don't necessarily have binary values. Comparing actual human responsibility judgments to predictions by Extended CPM could therefore provide us with additional insight if the general idea of heuristic criticality and structural pivotality is still viable with games CPM was not tested against (i.e. those where agents' contributions cannot be simply classified as "success" or "failure").

Belief Update & Epistemic State Manipulation

The extended definition of heuristic criticality also allows us to include the influence of prior beliefs about an agent's likely contribution. Lagnado, Gerstenberg & Zultan (2013) tested CPM using games where individual contributions were not made in any perceivable order. Because subjects acquired all knowledge about one game at once, it is hard to see what difference epistemic state could make. However, since prior distribution is explicitly included in the extended heuristic criticality, it is potentially possible to model how prior belief gets updated as information is gradually acquired. To do so, the experiments will be designed to reveal only some information about individual contributions at a time. Participants will make responsibility judgments based on different partial information as well. This way, we can also examine CCM, which relies on contributions represented in some perceivable order to assign responsibility, and see if it offers a reasonable prediction.

We hypothesize that retrospective ratings for non-critical agents will be higher. Prospective ratings for these agents would be based on incomplete information, which could lead to underestimation of the effect of a contribution due to uncertainty (Gillett, 1985). As a result, the agents would be perceived as less pivotal,

as the probability of the outcome (i.e. P_O) will be estimated less than it actually is. Additionally, the action of these agents will be seen as less influential given the uncertainty in outcome, meaning they should also receive lower criticality score. For instance, if we only know that Aditya voted for himself in the Aditya Problem before everyone had voted (without knowing he won eventually), it is expected for people to underestimate how critical and pivotal his vote is. It would be hard to predict the outcome with the majority of decisions (i.e. other 9 votes) completely unknown. However, in the *Retrospective Rating Phase*, subjects will likely revise their decision based on complete information of the game, including a higher probability of outcome (and thus higher pivotality) based on subsequent contributions (e.g. five other people voted for Aditya and four voted Bart, so his vote would be necessary), and/or higher *influence of action* (and thus higher criticality).

Intention/Goal-directedness

Furthermore, the explicit inclusion of goal-directedness as a factor allows us to compare games with human agents versus non-human agents. The paired experiments will be mathematically equivalent but differ in the type of agents. We further hypothesize that, since human agents are more likely to be perceived intentional and goal-directed which justifies a higher rate of moral responsibility (Shultz & Wright, 1985), they will receive higher responsibility ratings compared to their non-agent counterpart.

Additionally, the experiments were designed to verify some previous theories. Specifically, we are interested in testing the order (attenuation) effect reported by Gerstenberg & Lagnado (2012), which states that contributions made after an outcome has become certain gets discredited in terms of responsibility compared to the responsibility rating they would otherwise receive if they were to appear when the outcome is still uncertain.

Participants 44 subjects (20 male, 24 females) were recruited using flyers. The mean age was 20.6.

Materials There were six Responsibility Attribution Tasks. Each Responsibility Attribution Task involves a different written story about some team contribution scenario, where several players individually contributed to a common goal follow-

ing some explicit order, giving rise to a certain collective outcome (see Appendix A for details). There are two sets of tasks: the first set (Task 1 & 2) is a paired task with mathematically equivalent setup, but different types of agents (human voters in Task 1 vs. non-human machine reels in Task 2); the second set (Task 3–6) consists of four related tasks, where three kids scored points and their total number of points determined whether they could receive a treat or not. Each subject completed all six tasks. During each task, subjects engaged in the *Prospective Rating Phase* first, followed by the *Retrospective Rating Phase*. All subjects completed the tasks in the same order.

Prospective Rating Phase

For each cause described in each task, subjects were asked to rate

(1) *estimated outcome probability*: the probability of each possible collective outcome for the team given all known individual contributions so far, on a 7-point scale (for Task 1 & 2 with 7 players) or a 5-point scale (for Task 3–6 with 3 players), and

(2) *responsibility rating*: the responsibility of the acting agent (player) with respect to all possible collective outcomes of the team given all known individual contributions so far, on a 5-point scale

Subjects were prompted to report the ratings immediately after reading about each agent's action, but before revealing what the next agent in the sequence did. For example, on the first Responsibility Attribution Task with 7 voters, 2 candidates (Andre and Brenda), and majority rule (no abstention allowed), subjects read about the first voter's choice (and no one else's), and were asked to respond to "At this point, who do you think will win this election?" on a 7-point scale (1 = "Definitely Andre", 7 = "Definitely Brenda"). They were also asked to rate "If Andre ended up winning this election, how much do you think the first voter is responsible for Andre's win?" and "If Brenda ended up winning this election, how much do you think the first voter is responsible for Brenda's win?" on a 5-point scale respectively. After submitting their ratings, the second voter's choice was shown, and the same process was repeated.

Retrospective Rating Phase

For each task, after viewing all players (and so learning the final outcome), subjects were asked to rate responsibility for each contributing agent involved in the game described in the task on a 5-point scale. By this point, they have complete knowledge about all contributions.

Procedure All subjects were informed about the general structure of the experiment, consisting of six short written stories, each of which constituted a discrete Responsibility Attribution Task. They were also told about the response format (mostly multiple choice questions, with one short answer question for each story). After signing a digital consent form, the subjects then completed the six Responsibility Attribution Tasks in a predetermined order (i.e. the six tasks were presented in the same order for everyone). Subject responses to the tasks were recorded using a computer keyboard. They had the option to take a break between tasks.

Design The study consists of two sets of experiments, Tasks 1 & 2 and Tasks 3–6. The first set of experiments (Tasks 1 & 2) used a within-subject design. The independent variable was the type of acting agents (human vs. machine). The second set of experiments (Tasks 3–6) used a 2x2 within-subject design. The independent variables were the collective outcome (success vs. failure) and the relative performance of the second player (no worse than any other player vs. worse than some other player). For both sets of experiments, the dependent variables were probability judgments and responsibility ratings of each agent with respect to all possible collective outcomes. The hypotheses were examined using paired sample t-test, multivariate correlation, and repeated measures ANOVA.

Chapter 5

RESULTS

All 44 subjects submitted valid responses during the *Prospective Rating Phase*. However, because retrospective ratings were reported in the form of short-answers rather than multiple choices (as in the *Prospective Rating Phase*; see Appendix A), 7 subjects submitted at least one invalid response during the *Retrospective Rating Phase* for all tasks. The responses were invalid because they contained out-of-range numerical values (e.g. 0 on a 5-point scale, where the minimum value is 1), or missing responsibility rating for at least one positive contributor. Retrospective rating data from these 7 subjects were therefore excluded from subsequent data analysis. Everyone else submitted valid responses during the the *Retrospective Rating Phase* for every task. Analysis of retrospectively made responses was therefore based on data from 37 subjects.

Sensitivity of Responsibility Rating to Perceived Contributing Order

The first set of tasks (Tasks 1 & 2) tested whether responsibility rating was sensitive to the order in which contributions were made (i.e. if people attribute responsibility differently to agents solely because the agents were perceived to have acted at different times), as well as possible influences of the type of agent (human vs. machine). Both tasks involved a game where agents (voters or reels) carried out actions in some fixed sequential order. It is worth noting that the subjects were told that all agents in the scenarios made independent contributions. Additionally, in Task 1, the number of votes either candidate gets was not disclosed at any point during the election, and the results were made public to all voters at once after the last voter cast their vote. By making these assumptions clear, any difference in responsibility ratings for two arbitrary contributors was less likely due to subjects' belief that the two had some sort of "collusive" behavior, or were engaged in strategic contributing as opposed to acting purely on themselves. In other words, the task description attempted to establish the independence of decisions, so as to isolate the variance of each responsibility rating to aspects only relevant to the corresponding agent and nothing else.

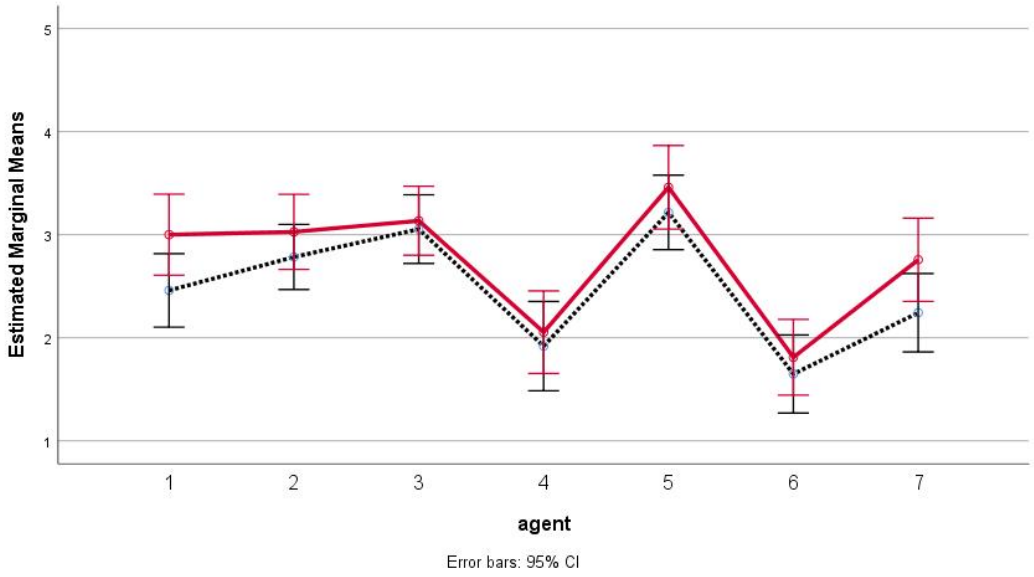


Figure 5.1: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 1 (seven voters with overdetermination). The dotted black line corresponds to prospective rating, and the solid red line corresponds to retrospective rating. Agent 1, 2, 3, 5, and 7 voted for the winning candidate, while agent 4 and 6 voted against the winning candidate.

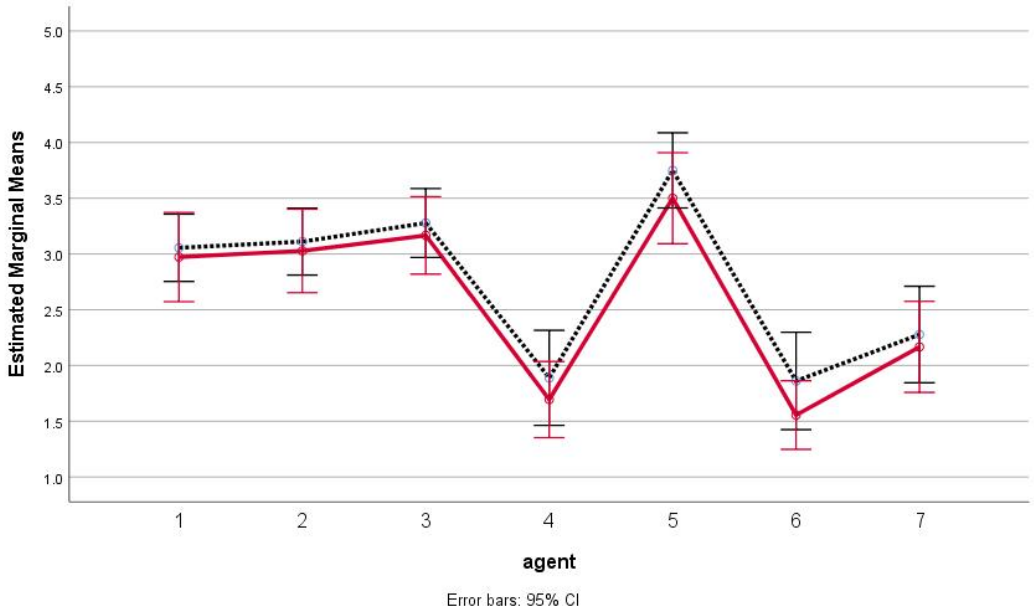


Figure 5.2: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 2 (seven reels with overdetermination). The dotted black line corresponds to prospective rating, and the solid red line corresponds to retrospective rating. Agent 1, 2, 3, 5, and 7 contributed positively to the eventual outcome, while agent 4 and 6 did not contribute to the outcome.

Indeed, subjects' responsibility ratings of positive contributors in Task 1 exhibited strong order effect, in both prospective rating (i.e. subjects rate responsibility of each agent as the story unfolds, without information about contributions down the line), $F(4, 40) = 13.852$, $p < 0.001$, $\eta_p^2 = 0.244$, as well as retrospective rating (i.e. subjects rate responsibility of each agent with knowledge about every contribution), $F(4, 33) = 6.949$, $p = 0.002$, $\eta_p^2 = 0.162$. More importantly, there was a clear trend of increasing responsibility rating for positive contributors as the story progressed until the outcome had been determined in both prospective and retrospective rating tasks (see Figure 5.1), with a reduced magnitude in responsibility rating when judgments were made retrospectively compared to prospectively.

The same order effect in responsibility judgment was also observed in Task 2 (machine reels with overdetermination; see Figure 5.2), in both prospective rating, $F(4, 40) = 22.611$, $p < 0.001$, $\eta_p^2 = 0.345$, and retrospective rating, $F(4, 32) = 15.260$, $p < 0.001$, $\eta_p^2 = 0.304$. In parallel with Task 1, subjects assigned agents (i.e. reels) that contributed later in the sequence of actions of higher responsibility. Again, the magnitude of such trend was attenuated (i.e. the difference in rating for the first few agents) in retrospective rating compared to prospective rating.

Gerstenberg & Lagnado (2012) reported attenuation effect in responsibility attribution when later contributors are aware of earlier contributions: in cases where the outcome was already determined, contributions were discounted. This was also (partially) predicted by CCM: since the outcome was certain, the probability of any outcome remained constant, regardless of actions of the remaining contributors. Any further contribution would cause no change in the probability, and thus received a score of zero. However, CCM fell short in predicting the sensitivity of responsibility rating to performance even after the outcome was certain (Gerstenberg & D. Lagnado, 2012).

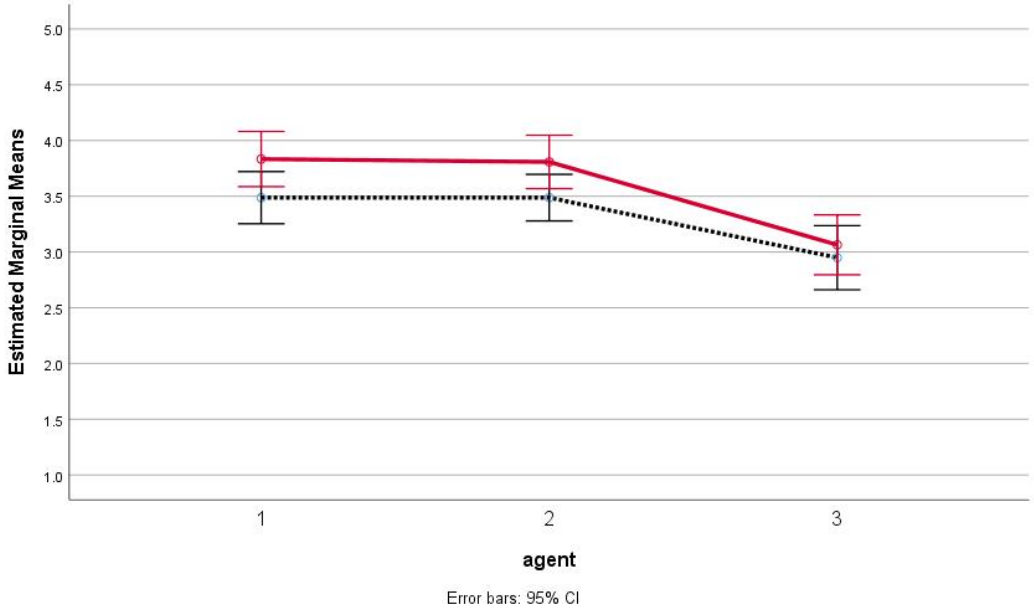


Figure 5.3: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 3 (three kids, scoring 4 (James), 4 (Albus), 3 (Lily) in sequence). The dotted black line corresponds to prospective rating, and the solid red line corresponds to retrospective rating.

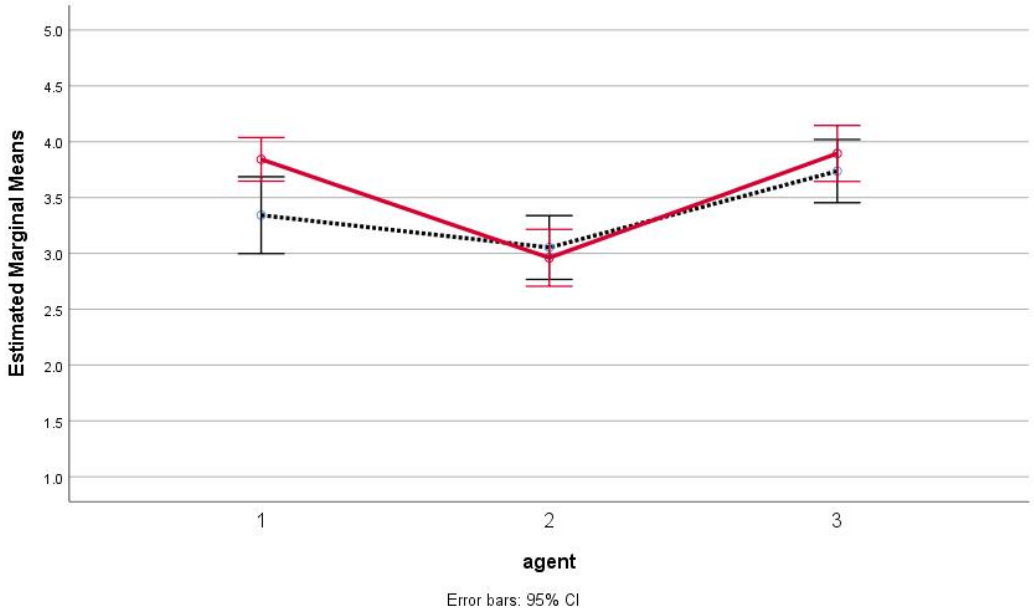


Figure 5.4: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 5 (three kids, scoring 4 (James), 3 (Lily), 4 (Albus) in sequence). The dotted black line corresponds to prospective rating, and the solid red line corresponds to retrospective rating.

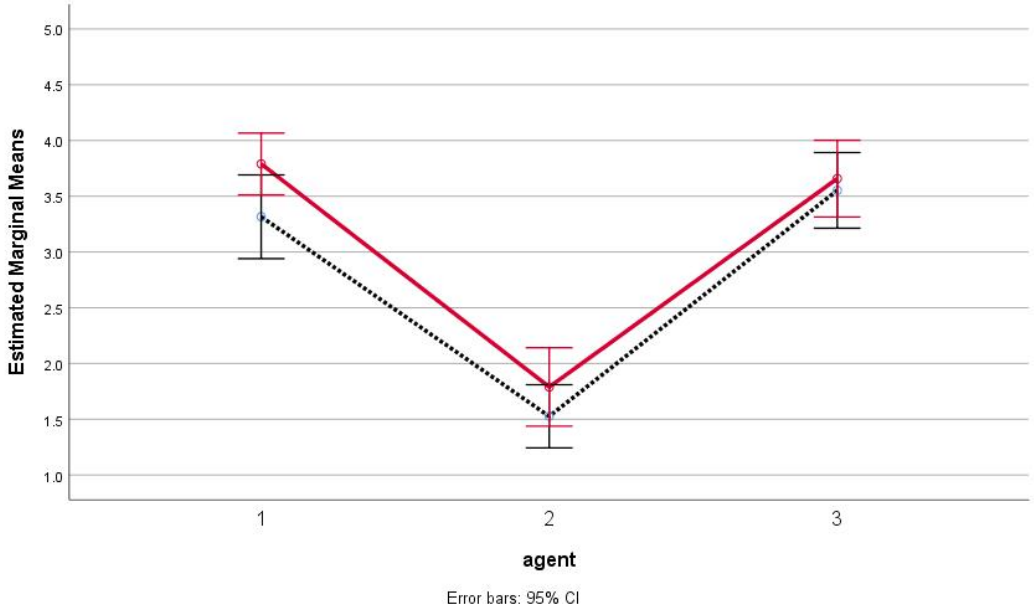


Figure 5.5: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 4 (three kids, scoring 1 (James), 4 (Albus), 1 (Lily) in sequence). The dotted black line corresponds to prospective rating, and solid red line corresponds to retrospective rating.

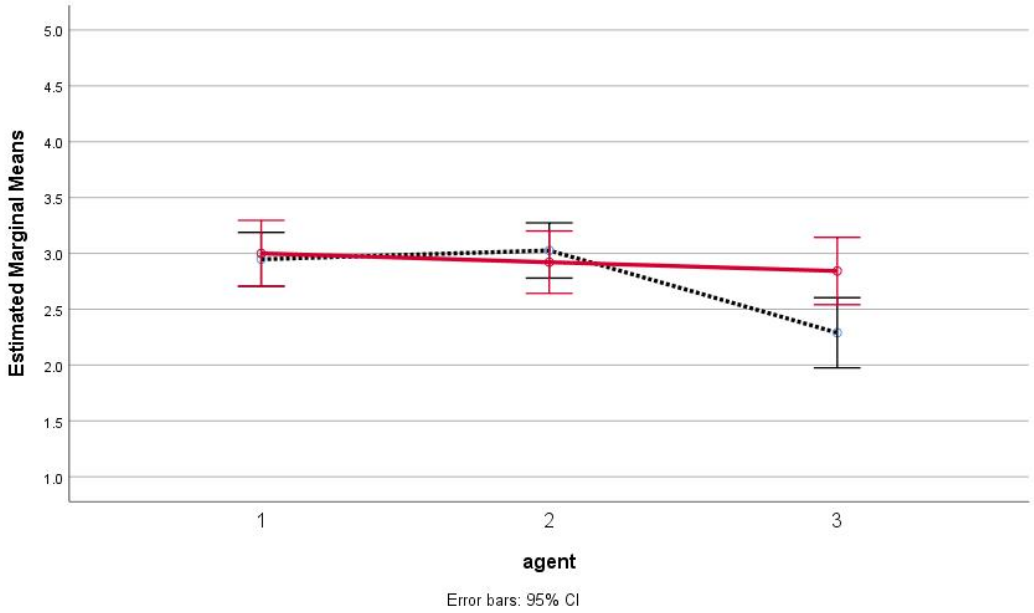


Figure 5.6: Average responsibility rating (y-axis) for each agent (x-axis, in order of contribution) on Task 6 (three kids, scoring 3 (James), 3 (Lily), 3 (Albus) in sequence). The dotted black line corresponds to prospective rating, and solid red line corresponds to retrospective rating.

By design of the experiments, Task 3 and Task 5 were essentially the same, except the last two contributions were revealed in reversed order: subjects rated responsibility of James, Albus, and Lily, in that order, after reading about one of the kid's contribution, but before reading about the rest contributions for Task 3. During Task 5, they rated the responsibility of James, Lily, and Albus, in that order, again after reading about one of the kid's contribution, but before reading about the rest contributions. In both cases, the contributions were exactly the same, and the subjects were told that the contributions were made independently. Additionally, the outcome only became certain on the last contribution for both tasks. Therefore, the order in which contributions were revealed does not have an impact on the quality of the contributions. By comparing subjects' responsibility rating for the last two agents in Task 3 (see Figure 5.3) and Task 5 (see Figure 5.4), the aforementioned attenuation effect was observed during both *Prospective Rating Phase* ($p = 0.009$) and *Retrospective Rating Phase* ($p < 0.001$).

Predictability of Probabilistic Change on Responsibility Rating

Probabilistic estimates are key to CCM, which predicts responsibility judgment based on the change in probability of outcome due to an agent's contribution. Specifically, CCM uses the change in probability associated with a certain outcome before and after an agent's contribution and assigns responsibility based on how each change compares to others. Thus if CCM does indeed capture human responsibility attribution, we would expect to see a correlation between change in reported *estimated outcome probability* and responsibility rating for the agent that directly caused the change in perceived outcome likelihood. CPM, however, does not incorporate CCM.

The probabilistic change used by CCM is in fact an accurate predictor of subjects' responsibility rating regardless whether agents were humans or machine reels. We found strong, statistically significant correlation between changes in subjects' reported probability of outcome (*outcome probability estimate*) and prospective or retrospective responsibility rating of subsequent contributing agents. In other words, the difference between estimated probabilities of outcome that were reported before and after agent i contributed was strongly correlated with subjects' responsibility assignment. In Task 1 (seven voters with overdetermination), changes in *outcome probability estimate* were reliable predictors of prospective (correlation strength = 0.835, $p = 0.039$) but not retrospective responsibility ratings (correlation strength

= 0.782, $p = 0.066$). In Task 2 (seven reels with overdetermination), changes in *outcome probability estimate* were reliable predictors of prospective (correlation strength = 0.954, $p = 0.0031$) as well as retrospective responsibility ratings (correlation strength = 0.930, $p = 0.0073$).

Since there were only two data points for each correlation analysis between *outcome probability estimate* and prospective or retrospective ratings for Task 3–6, no meaningful correlation can be concluded.

Prospective vs. Retrospective Responsibility Attribution

Our results of Task 1 (voting) suggested that prospective and retrospective responsibility ratings were not different except for the first ($p = 0.009$) and the last voter ($p = 0.007$). In particular, subjects tended to give the first voter more responsibility in retrospect (0.56 more on average), thus assigning the first three voters (i.e. the ones whose voted did not determine the election result in the order the votes were cast) more or less the same. Subjects also gave the last voter about 0.53 more responsibility on a 5-point scale in retrospect than in prospect.

In contrast, responsibility attributed to all agents in Task 2 (random machine) remained the same through *Prospective* and *Retrospective Rating Phase*, although retrospective ratings were consistently slightly lower than their corresponding prospective ratings.

For tasks with continuous-valued contributions, prospective and retrospective responsibility ratings for the first agent (James) were different when the contributing pattern was 443 ($p = 0.008$), 141 ($p = 0.011$), and 434 ($p = 0.003$). The ratings also differ for the second agent for contributing patterns 443 (Albus, $p = 0.009$) and 141 (Albus, $p = 0.039$). Additionally, the ratings differed on the last agent of contributing pattern 333 (Albus, $p = 0.006$).

Responsibility Attribution on Human vs. Non-human agents

Responsibility ratings and probabilistic estimates were compared across Task 1 and 2, which were mathematically equivalent, but were primarily different because of the agents' level of (perceived) goal-directed behavior. Since Task 1 involved actual human voters, as opposed to machine reels in Task 2, it is more likely for subjects to see agents in Task 1 having more control and motivation behind their action.

Paired t-tests revealed significantly higher responsibility ratings on agent 1 ($p < 0.001$), 2 ($p = 0.012$), and 5 ($p < 0.001$) only in Task 2 than Task 1. Responsibility attributed to agent 3 was marginally different ($p = 0.058$). Additionally, the probability estimates were very different across two tasks after the first agent contributed ($p < 0.001$ for all), where Task 2 received consistently higher probability estimates than Task 1.

Chapter 6

DISCUSSION

Order Effects in Responsibility Attribution

Results of the first set of experiments (Task 1 & 2) showed that subjects exhibited very similar order effects in both their prospective and retrospective responsibility attribution to agents who contributed to a collective cause sequentially. Specifically, agent contributions before the outcome becomes certain received higher responsibility score if their contribution was perceived to have brought the team closer to the "critical point" of the outcome. i.e. responsibility rating was inversely proportional to the distance from the current state of the game to the expected final state of the game in terms of the contribution sequence. The patterns in responsibility judgment were consistent across participants. Intuitively, as the game progresses, the probability of alternative outcomes decreases, and the result could be determined by very few future contributions *in this particular scenario*. Although earlier contributions could very well steer the outcome, temporal order likely overrides such consideration by encouraging heavier weights on later contributions, whose temporal distance to the determined outcome is shorter. Therefore, people tend to treat later contributors as more critical when an explicit temporal order of events is provided (Miller & Gunasegaram, 1990), assigning them higher responsibility.

Furthermore, attenuation effect reported by Gerstenberg & Lagnado (2012) was present in the first set of experiments with binary contribution values: the last agent contributed positively to the outcome, but received the lowest average responsibility score among all five positive contributors, as the outcome was already fully determined by the time they contributed. However, note that this responsibility is not zero, which CCM would predict, as the last contribution did not cause any change probability of the outcome. Therefore, people do reason about both the scenario they were presented with as well as alternative scenarios, where the last contribution *would have* mattered.

The result of the second set of experiments (Task 3-6, but Task 3 and Task 5 in particular) further supported the observation of attenuation effects reported by Gerstenberg & Lagnado (2012). There was strong evidence that, if an agent's positive contribution was perceived to have shown up as or after the collective outcome

becomes certain, it would be treated as if it is less important than it *would have* been considered had it shown up *before* the outcome is determined. For instance, subjects attributed 0.27 (out of 5) more responsibility on average to Lily when her contribution was revealed when the outcome was uncertain (Task 5) than when her contribution was revealed as the outcome becomes certain (Task 3).

Role of Probabilistic Reasoning in Responsibility Attribution

We were able to collect participants' probability estimates about the occurrence of certain outcome based on disclosed information of individual contributions at different points in the game/scenario and compare them with corresponding responsibility scores. Strong correlations of prospective responsibility ratings and probability estimates were observed regardless of the humanness of agents. Additionally, there was a strong correlation between retrospective responsibility ratings and probability estimates when the agents were non-human, but not when they were humans. Since CCM is constructed specifically to explain individual contribution events unfolding in a perceivable sequence with no constraint on whether such sequence reflects the actual ordering or not (Spellman, 1997), these results are consistent with what CCM promises.

However, note that the correlation between retrospective responsibility ratings and probability estimates for human agents (i.e. Task 1) was not significant, even though there was a strong correlation with prospective ratings. Since we prompted subjects to report their probability estimates of outcome up to the point when they learned about an agent's action before assigning responsibility to said agent, they might be more aware of the potential causal relationship between the probability of outcome and responsibility. On the other hand, they were not asked to report probability estimate again during *Retrospective Rating Phase*. Therefore, participants were either (1) more aware that the ordering is arbitrary, as an agent's contribution could be rendered more or less critical depending on the specific contribution sequence, or (2) more conscious of the intention and/or autonomy of agents. The first explanation was in line with Gerstenberg & Lagnado (2012), who suggested that CCM could be extended to also account for the possibility that an agent's contribution could make a difference in other similar games. However, considering the fact that probability estimates were correlated with both prospective and retrospective rating for non-human agents (Task 2), the second explanation may be more plausible. Since Gerstenberg & Lagnado (2012) did not study games with non-human objects, it

would be of future study's interest to look into how CCM could be extended to account for retrospective ratings.

In conclusion, our experimental results suggest that probability reasoning played a significant role in subjects' responsibility attribution decisions. This implies that CCM did offer accurate explanations of responsibility attribution under certain constraint, namely that responsibility judgments are made given contributions explicitly represented in a sequence. Without such constraint, CCM still produced accurate predictions if the game involved non-human agents exclusively, but not when it involved human agents, consistent with Gerstenberg & Lagnado (2012)'s finding.

Effect of Epistemic State on Responsibility Attribution

The main difference between prospective and retrospective responsibility attribution was people's epistemic state: prospective responsibility attribution, as opposed to retrospective responsibility attribution, inevitably involves uncertainty about future contributions and thus the outcome. We hypothesized that retrospective responsibility rating would be higher for all agents who were not "critical" in the given sequential scenario. This hypothesis turned out to be partially true: when the agents were humans, the first and last positive contributor received significantly higher responsibility ratings in retrospect. This is expected because these contributions were more likely discredited, as they were made either when there was a lot of uncertainty left in the game, or the particular outcome determination rendered the contribution inconsequential. However, in retrospect, subjects likely realized that in some alternative scenario, earlier contributions and "excessive" contributions *would have* been critical. Therefore, they revised their responsibility rating in retrospect, giving more credit to the overlooked contributions. Indeed, there was not any difference among retrospective responsibility ratings for agents other than the "critical" one, which suggests that complete knowledge facilitates counterfactual reasoning of different, but similar scenarios.

On the other hand, prospective and retrospective judgments about the critical agent (who received highest responsibility rating on average prospectively) were indifferent. This may be because the critical agent was treated as the reference point for other non-critical actors during retrospective attribution, as she unmistakably contributed positively to the actual outcome. Out of 37 pairs of valid prospective and retrospective responsibility ratings for the critical agent (voter 5/reel 5), only 5 subjects raised their retrospective responsibility, while the majority of the rest

kept their prospective judgment, or even revised to something lower. During *Retrospective Rating Phase*, 78% of subjects did not distinguish non-critical but positive contributing agents from the critical one. Furthermore, their justification demonstrated that they were aware of the arbitrariness of the contributing order, which was cited as the reason why they assigned all positive contributing agents responsibility similar to the critical agent in the specified sequence.

Effect of Intention and Goal-directed Autonomy on Responsibility Attribution

Comparing tasks in the first set of experiments led to a surprising revelation: nonhuman agents were perceived as more responsible for their contribution than human agents. This result is in stark contrast with our intuitive hypothesis that people would see human agents as more responsible for their action due to factors like motivation and goal-directed behavior. In other words, humans have more control over their individual contribution, and thus they should be held responsible for their behavior, especially when they act autonomously.

There are several possible explanations: first, although Task 1 and 2 were mathematically equivalent, the stake associated with the outcome was different. Task 1 with voters has outcomes (1) Andre winning and (2) Brenda winning. Since there was no further information given about the qualification or performance of either candidate, it is unlikely people would prefer one candidate over another. However, Task 2 with random reels has outcomes (1) Andrew getting $10 and (2) Andrew getting $0. Naturally, people would prefer getting a positive amount of money when the alternative is getting none. Therefore, it is reasonable for them to attribute more responsibility to nonhuman agents (reels) to account for the appeal of a certain outcome. Therefore, a possible improvement to responsibility attribution models like CPM might be to factor in different degrees of stake associated with different outcomes.

Another explanation for this counterintuitive result is that people may perceive a stronger link between the nonhuman object and its contribution. As Lombrozo (2010) pointed out, people seek mechanistic explanations when teleological explanations are unavailable. For human agents, teleological explanations like “voter 1 voted A because she preferred A” makes sense. Yet in the case where all agents were machine parts like reels, a similar causal explanation making reference to the agents’ intention is unavailable. Ceteris paribus, when intention as a factor is missing in a game, mechanistic explanations become more accessible and thus favored over

teleological ones. As a result, people's understanding of the game is largely based on the mechanical properties of the actors (e.g. which symbol a certain reel is made more likely to show). The perceived deterministic nature of mechanics, especially in contrast to possibly arbitrary intentions, could have led to the (wrong) belief that the contributions are more deterministic than they really are. Determinism, in turn, rendered the causal link between the (non-human) agent and its contribution to the eventual outcome.

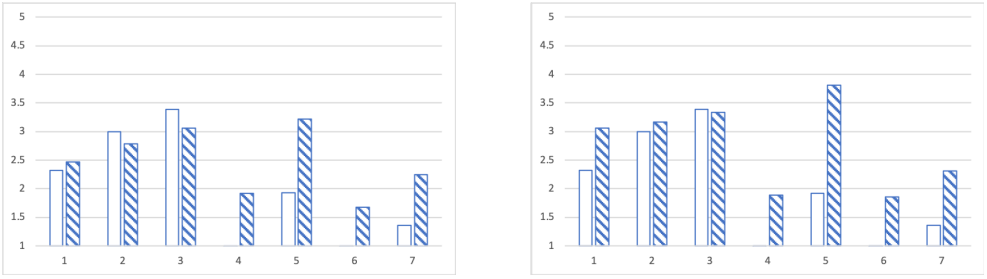
In fact, the core difference may be less about non-human versus human agent, but rather the available mode of explanation. Lombrozo (2010) showed that, even when agents are human, if specific conditions of the game do not naturally lead to teleological explanations, people's explanation of the game follows the mechanistic mode. Therefore, future studies should investigate how the difference in available modes of explanation may result in a difference in responsibility attribution behavior.

Chapter 7

MODEL PREDICTION RESULTS

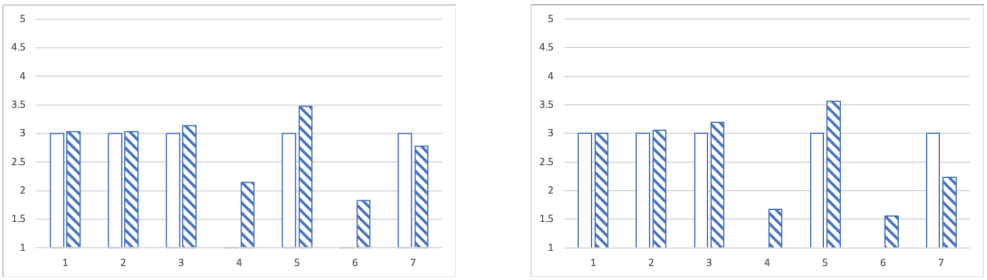
Recall that the two definitions of $M(\tilde{C}_i, a^*)$, the likelihood for player i to do something different than a^* and still have r^* as the group outcome, are equivalent when the action space A_i is binary-valued. Therefore, predictions produced by the Extended CPM using *Worst Consequence* or *Weighted Total Consequence* are exactly the same for Task 1 and 2.

Task 1 & 2



(a) Task 1 (positive contributors: 1,2,3,5,7) *(b) Task 2 (positive contributors: 1,2,3,5,7)*

Figure 7.1: Average prospective responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 1 and Task 2.



(a) Task 1 (positive contributors: 1,2,3,5,7) *(b) Task 2 (positive contributors: 1,2,3,5,7)*

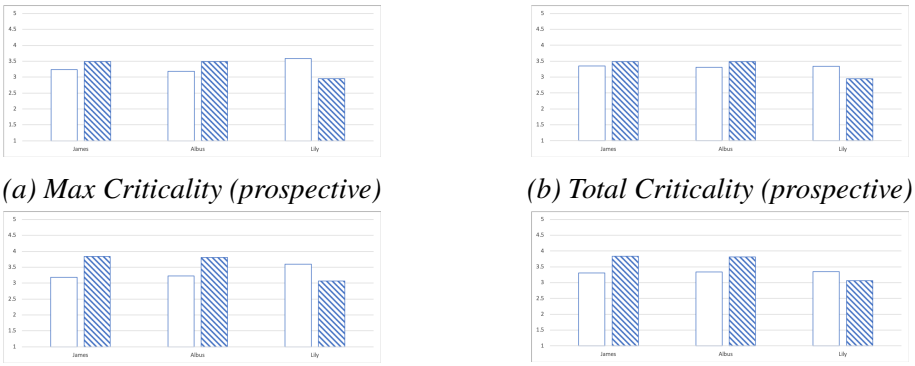
Figure 7.2: Average retrospective responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 1 and Task 2.

Consistent with Lagnado, Gerstenberg, & Zultan's finding (correlation = 0.90, 2013), the retrospective ratings produced by Extended CPM were highly correlated with the retrospective ratings reported by subjects: on Task 1, correlation strength was about 0.9241 ($p = 0.0029$); on Task 2, correlation strength was about 0.8628 ($p = 0.012$).

Extended CPM was able to capture some important trends shown in the behavioral data: first, it predicted the increasing amount of responsibility attributed to the first three positive contributors in Task 1 and 2 due to growing heuristic criticality (see Figure 7.1). It also predicted the attenuated, but positive responsibility rating assigned to the last "extra" contributor, as pivotality of the last contribution is much less than the previous ones that determined the outcome before the last contribution was made (see Figure 7.1). However, note that Extended CPM did not predict the increasing trend in prospective responsibility rating up until the deciding contribution (i.e. the fourth voter/reel). In fact, the responsibility rating suddenly dropped.

Additionally, Extended CPM predicted that retrospective ratings for the first three agents in Task 1 and 2 (i.e. all positive but non-critical contributors) are not different (Task 1: $p = 0.285$; Task 2: $p = 0.085$) (see Figure 7.2). Also, Extended CPM assigned the agent 5—the critical contributor—the same amount of responsibility as other positive contributors, which was not the case in our behavioral results. The model prediction for agent 7—the "excessive" contributor—was also the same as other positive contributors, but subjects gave a much lower retrospective responsibility rating.

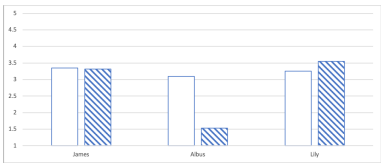
Task 3 – 6



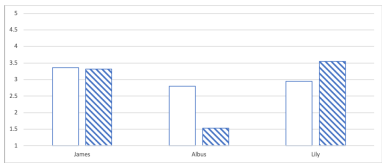
(a) Max Criticality (prospective) *(b) Total Criticality (prospective)*

(c) Max Criticality (retrospective) *(d) Total Criticality (retrospective)*

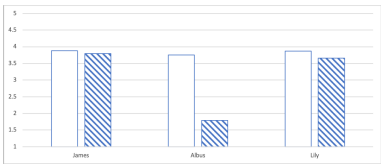
Figure 7.3: Average responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 3 (James - 4, Albus - 4, Lily - 3, in order).



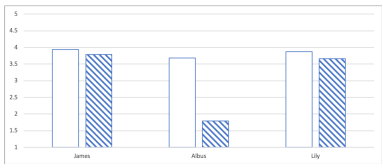
(a) Max Criticality (prospective)



(b) Total Criticality (prospective)



(c) Max Criticality (retrospective)



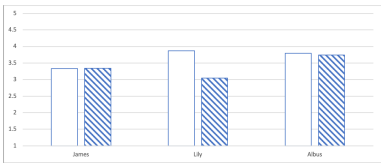
(d) Total Criticality (retrospective)

Figure 7.4: Average responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 4 (James - 1, Albus - 4, Lily - 1, in order).

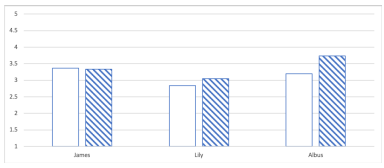
There are two sets of predictions made by Extended CPM using the two definitions of $M(\tilde{C}_i, a^*)$ respectively (i.e. *Worst Consequence* and *Weighted Total Consequence*). We did not find any statistically significant correlation between responsibility rating made by human subjects and Extended CPM on Task 3, 4, 5, and 6 in most cases, regardless whether it was prospective or retrospective. The only significant correlation found was between **prospective** responsibility ratings of human subjects and model predictions using *Worst Consequence* implementation of $M(\tilde{C}_i, a^*)$ for Task 4 (James - 1, Albus - 4, Lily - 1, in order; $p = 0.037$).

Albeit a lack of correlation between model predicted responsibility rating (using either definition of $M(\tilde{C}_i, a^*)$) and actual experimental data exists, Extended CPM generally captures the noticeable trends in responsibility attribution to some varying extent. For instance, on Task 3, Extended CPM with both *Worst Consequence* and *Weighted Total Consequence* predicted that the first two agents were assigned the same responsibility (see Figure 7.3), considering their very similar criticality and pivotality. In addition, Extended CPM gave better results on retrospective rating using either definition of $M(\tilde{C}_i, a^*)$: patterns in actual responsibility assignment were observed on Task 4 (Albus < Lily < James in terms of responsibility; see Figure 7.4), Task 5 (Lily < Albus, James in terms of responsibility; see Figure 7.5), Task 6 (Albus < James, Lily in terms of responsibility; see Figure 7.6), although the relative magnitude may be inaccurate.

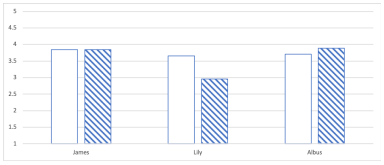
In terms of prospective rating, Extended CPM with *Weighted Total Consequence* was better than Extended CPM with *Worst Consequence*. It produced fewer erroneous patterns on Task 3 (see Figure 7.3 (a) vs. (b)) and Task 5 (see Figure 7.5 (a) vs. (b)).



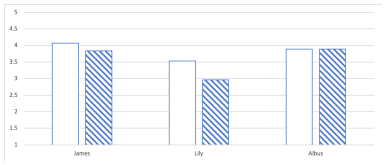
(a) Max Criticality (prospective)



(b) Total Criticality (prospective)

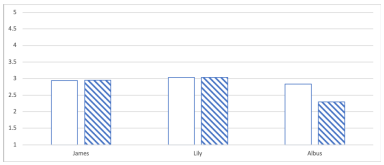


(c) Max Criticality (retrospective)

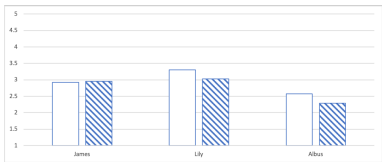


(d) Total Criticality (retrospective)

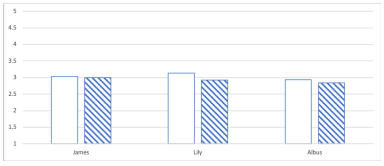
Figure 7.5: Average responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 5 (James - 4, Lily - 3, Albus - 4, in order).



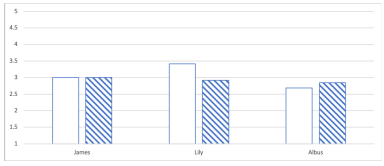
(a) Max Criticality (prospective)



(b) Total Criticality (prospective)



(c) Max Criticality (retrospective)



(d) Total Criticality (retrospective)

Figure 7.6: Average responsibility rating (shaded bars) with predictions of the Extended CPM model (white bars) on Task 6 (James - 3, Lily - 3, Albus - 3, in order).

Chapter 8

GENERAL DISCUSSION

Task 1 & 2

One of the most obvious fails of Extended CPM was regarding why subjects found agent 5, whose vote fully determined the outcome of the game given the particular order of votes, the most responsible for the final outcome. While participants unanimously assigned agent 5 the highest responsibility among all 7 agents, Extended CPM predicted attenuated responsibility, only slightly higher than the "excessive" agent 7. This prediction is because, even if agent 5 did not make a positive contribution, based on previous votes, it is very likely that someone after her would make a positive contribution anyways, leading to the same outcome. The spike in subjects' responsibility rating on the agent with the deciding vote may be due to either inflated pivotality or higher weight on pivotality in this special case. It would be of interest for future studies to investigate if people weigh criticality and pivotality differently under different degrees of uncertainty about the outcome.

Another problem seems to be the fact that Extended CPM assigned all positive contributing agents (i.e. agent 1, 2, 3, 5, 7) the same retrospective responsibility, which was not consistent with actual ratings by subjects (see Figure 7.2). However, it is worth noting that 24 out of 37 subjects (64.9%) with valid retrospective rating in Task 1 gave the same retrospective rating to *all* positive contributors (i.e. voter 1, 2, 3, 5, and 7), and 19 out of 37 subjects (51.4%) with valid retrospective rating in Task 2 gave the same retrospective rating to *all* positive contributors (i.e. reel 1, 2, 3, 5, and 7). Therefore, it is possible that subjects used different decision-making strategy in retrospective rating. Although all positive contributors share the same criticality and pivotality, it is possible that some participants still treated the critical agent *in this particular scenario* as a special case, and emphasized the importance of her contribution, either by amplifying criticality and/or pivotality or changing the weighting of the two measures. Similarly, some may have also treated the "excessive" agent as a special case *in this particular scenario*, overlooking the fact that agent 7 could very well be important in *some other scenario*.

Task 3 – 6

Because beliefs about individual agents are explicitly represented as statistical distributions in the model, which are updated immediately as new information about each agent comes in, model predictions are sensitive to order information and prior knowledge about the performance of each agent. For example, although James and Lily scored the same (-1 point) in Task 4, Lily was attributed lower responsibility by both human subjects and the Extended CPM model (see Figure 7.3 (b)), as the state of the game immediately before she contributed makes it unlikely for her to make a devastating mistake for the entire team. On the other hand, James could be seen as "setting the ground" for the team's eventual success as his contribution was revealed first. This aspect of the model is particularly interesting, as the original CPM does not apply to games without binary-valued contributions.

The reason why using *Weighted Total Consequence* in Extended CPM turned out to be slightly better than *Worst Consequence* may also be attributed to how the two use new information about agents differently. For all tasks involved in this study, $\arg\min_{a \in A_i \setminus a^*} P(R = r^* | \tilde{C}_i, X_i = a)$ happens to be 0 for all $\tilde{C}_i$ such that $\tilde{C}_i = C_{i_{min}}$ (i.e. the counterfactual game closest to the original game). Thus model prediction using *Worst Consequence* did not effectively use all information available corresponding to different epistemic state of the raters/subjects *for this particular study*. On the other hand, beliefs about individual agents and thus distributions of probabilities of the form $P(X_i = a | \tilde{C}_i)$ (where $a \in A_i \setminus a^*$ for fixed agent i and action a^*) are updated as more information about the agents become available. As the result, *Weighted Total Consequence*, or $\sum_{a \in A_i \setminus a^*} P(R = r^* | \tilde{C}_i, X_i = a) P(X_i = a | \tilde{C}_i)$, probably reflects a more accurate representation of what the responsibility judger perceive as the "consequence" or "influence" of an action. However, given that almost no correlation between model predictions and human ratings were found, as well as the cost of such computation, *Weighted Total Consequence* may be approximated, rather than meticulously calculated as the underlying cognitive mechanism.

Extended CPM, especially with *Weighted Total Consequence*, also predicted the trend in responsibility rating better when the "worst" player (in this case, Albus) is worse than others, but not by much (i.e. Task 3, Task 5; see Figure 7.3 (b)(d) and Figure 7.5 (b)(d)). In a way, observing the "worst" player performing somewhat worse confirmed the rater's prior belief that the "worst" player did constantly perform worse than everyone else. If the "worst" player clearly stood out from the rest (Task 4; see Figure 7.4) or performs better than usual (Task 6; see Figure 7.6), people may be inclined to attribute more moral responsibility to the player, thereby

exaggerating the responsibility rating in either way (i.e. significantly higher if the player is doing much worse than others, and significantly lower if the player performs better). Indeed, quite some subjects demonstrated this kind of justification for their responsibility rating during the *Retrospective Rating Phase*. The fact that Extended CPM primarily relies on causal responsibility (pivotality, criticality) for its prediction may also account for the fact that Extended CPM was clearly off on responsibility assigned to the worst player in Task 4 (i.e. Figure 7.4), where the "worst" player did much worse than others.

Chapter 9

CONCLUSION

This paper has explored a structural model to account for responsibility attribution, namely the Extended CPM. The model is a natural extension of the Structural Model proposed by Lagnado, Gerstenberg, & Zultan (2013). Specifically, Extended CPM extended the notion of heuristic criticality and structural pivotality included in CPM so they are compatible with contributions and outcomes that are not necessarily binary-valued. As a side effect, Extended CPM allows explicit representation of epistemic state of the rater as prior distributions. An additional factor of intention of contributing agents was added to Extended CPM as well, in hope to differentiate scenarios with human agents and with non-human agents.

A series of experiment were performed to test the explanatory power of Extended CPM, especially against scenarios where (1) contributing agents are non-human, or (2) contributions are non-binary values. Moreover, epistemic state of the rater was manipulated by revealing individual contributions in some arbitrary order. Extended CPM had reasonable performance on these experiments, in terms of predicting relative magnitude of responsibility rating as well as significant order effect. However, it possibly underestimates the degree to which perceived goal-directness influences responsibility attribution. It also produced the opposite prediction for human versus non-human agents, although it could be due to different stakes the outcomes entailed in the two mathematical equivalent scenarios. Future studies should therefore look into the roles perceived goal-directness of agents and desirability of outcome have on responsibility attribution.

BIBLIOGRAPHY

Alicke, M. D. (1992). "Culpable causation." In: *Journal of Personality and Social Psychology* 63.3, pp. 368–378. DOI: 10.1037/0022-3514.63.3.368.

Bernstein, S. (2017). "Causal Proportions and Moral Responsibility." In: *Oxford Studies in Agency and Responsibility* 4. Ed. by D. Shoemaker, pp. 165–182.

Chockler, H. & Halpern, J. (2004). "Responsibility and blame: A structural-model approach." In: *Journal Of Artificial Intelligence Research* 22, pp. 93–115.

Fischer, J. M. (2010). "Responsibility and Autonomy." In: *A Companion to the Philosophy of Action*. Ed. by C. O'Connor T. & Sandis, pp. 309–316. DOI: 10.1002/9781444323528.ch39.

Gerstenberg, T. & Lagnado, D. (2012). "When contributions make a difference: Explaining order effects in responsibility attribution." In: *Psychonomic Bulletin Review* 19.4, pp. 729–736. DOI: 10.3758/s13423-012-0256-4.

Gillett, R. (1985). "Nominal scale response agreement and rater uncertainty." In: *British Journal of Mathematical and Statistical Psychology* 38, pp. 58–66. DOI: 10.1111/j.2044-8317.1985.tb00816.x.

Hitchcock, C. & Knobe, J. (2009). "Cause and Norm." In: *Journal of Philosophy* 106.11, pp. 587–612. DOI: 10.5840/jphil20091061128.

Knobe, J. (2010). "Action trees and moral judgment." In: *Topics in Cognitive Science* 2.3, pp. 555–578. DOI: 10.1111/j.1756-8765.2010.01093.x.

Lagnado, D. A. & Gerstenberg, T. (2016). "Causation in legal and moral reasoning." In: *The Oxford handbook of causal reasoning*. Ed. by M. R. Waldmann.

Lagnado, D., Gerstenberg, T., & Zultan, R. (2013). "Causal Responsibility and Counterfactuals." In: *Cognitive Science* 37.6, pp. 1036–1073. DOI: 10.1111/cogs.12054.

Lewis, D. (1973). "Causation." In: *Journal of Philosophy* 70.17, pp. 556–567. DOI: 10.2307/2025310.

Lombrozo, T. (2010). "Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions." In: *Cognitive Psychology* 61.4, pp. 303–332. DOI: 10.1016/j.cogpsych.2010.05.002.

Mandel, D. (2003). "Judgment dissociation theory: an analysis of differences in causal, counterfactual, and covariational reasoning." In: *Journal of Experimental Psychology: General* 132.3, pp. 419–434. DOI: 10.1037/0096-3445.132.3.419.

McClure, S. et al. (2007). "Time discounting for primary rewards." In: *Journal of Neuroscience* 27.21, pp. 5796–5804. DOI: 10.1523/JNEUROSCI.4246-06.2007.

Miller, D. T. & Gunasegaram, S. (1990). "Temporal order and the perceived mutability of events: Implications for blame assignment." In: *Journal of Personality and Social Psychology* 59.6, pp. 1111–1118. DOI: 10.1037/0022-3514.59.6.1111.

Reuter, K. et al. (2014). "The good, the bad, and the timely: How temporal order and moral judgment influence causal selection." In: *Frontiers in Psychology* 5, p. 1136. DOI: 10.3389/fpsyg.2014.01336.

Shultz, T. & Wright, K. (1985). "Concepts of Negligence and Intention in the Assignment of Moral Responsibility." In: *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement* 17.2, pp. 97–108. DOI: 10.1037/h0080138.

Spellman, B. (1997). "Crediting causality. (attribution of the outcome of an event)." In: *Journal of Experimental Psychology: General* 126.4, pp. 323–348. DOI: 10.1037/0096-3445.126.4.323.

Sytsma, J., Livengood, J., & Rose, D. (2012). "Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions." In: *Studies in History and Philosophy of Biol Biomed Sci* 43.4, pp. 814–820. DOI: 10.1016/j.shpsc.2012.05.009.

Vinokur, A. & Ajzen, I. (1982). "Relative importance of prior and immediate events: A causal primacy effect." In: *Journal of Personality and Social Psychology* 42.5, pp. 820–829. DOI: 10.1037/0022-3514.42.5.820.

Appendix A

EXPERIMENT SURVEY

Blame! (or praise)

Thanks for your curiosity! This is a short survey/experiment about how you blame/praise people for what they have done. Basically, I will walk you through three very short stories, and ask you to make some judgement about people involved in those three stories along the way. It's mostly going to take you about 30 minutes, or probably even less, but take your time to think things through if you'd like to. Sounds good?

* Required

1. **Before we move on to the actual experiment, I would like to ask you NOT to look ahead the questions. Additionally, please DO NOT go back and change your answer. Trust your intuition.** *
   Mark only one oval.

   ◯ Got it

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The seven voters lined up, and the first voter walked up to the voting booth.

2. **The first voter voted Andre (A). At this point, who do you think will win this election? (middle = neutral/hesitant)** *
   Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

3. **If Andre ended up winning this election, how much do you think the first voter is responsible for Andre's win?** *
   Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

4. **If Brenda ended up winning this election, how much do you think the first voter is responsible for Brenda's win?** *
   Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The second voter walked up to the voting booth.

5. **The second voter voted Andre (A). At this point, who do you think will win this election? (reminder: votes so far = AA)** *
   Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

6. **If Andre ended up winning this election, how much do you think the second voter is responsible for Andre's win?** *
   Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

7. **If Brenda ended up winning this election, how much do you think the second voter is responsible for Brenda's win?** *
   Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The third voter walked up to the voting booth.

8. **The third voter voted Andre (A). At this point, who do you think will win this election? (reminder: votes so far = AAA)** *
   Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

9. **If Andre ended up winning this election, how much do you think the third voter is responsible for Andre's win?** *
   Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

10. **If Brenda ended up winning this election, how much do you think the third voter is responsible for Brenda's win?** *
    Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The fourth voter walked up to the voting booth.

11. **The fourth voter voted Brenda (B). At this point, who do you think will win this election? (reminder: votes so far = AAAB)** *
    Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

12. **If Andre ended up winning the election, how much do you think the fourth voter is responsible for Andre's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

13. **If Brenda ended up winning the election, how much do you think the fourth voter is responsible for Brenda's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The fifth voter walked up to the voting booth.

14. **The fifth voter voted Andre (A). At this point, who do you think will win this election? (reminder: votes so far = AAABA)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

15. **If Andre ended up winning the election, how much do you think the fifth voter is responsible for Andre's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

16. **If Brenda ended up winning the election, how much do you think the fifth voter is responsible for Brenda's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The sixth voter walked up to the voting booth.

17. **The sixth voter voted Brenda (B). At this point, who do you think will win this election? (reminder: votes so far = AAABAB)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

18. **Now that you know as a fact Andre wins this election, as he already gets at least 4 votes, how much do you think the sixth voter is responsible for Andre's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

Story 1: Voting

Seven people are getting ready to participate in an election. There are two candidates: Andre (A) and Brenda (B). Before they vote, they have enough time to research about the candidates and learn about their qualification. They do not know who other people will vote for, and they wouldn't know either, after they themselves vote. Once they voted, they can't go back and change their vote. *****The candidate with the majority of the votes (i.e. >= 4 votes) wins the election.*****

The last voter walked up to the voting booth.

19. **The seventh/last voter voted Andre (A). At this point, who do you think will win this election? (reminder: votes so far = AAABABA)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely Andre	◯	◯	◯	◯	◯	◯	◯	Definitely Brenda

20. **Now that you know as a fact Andre wins this election, as he already gets at least 4 votes, how much do you think the last voter is responsible for Andre's win?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full responsibility

21. **In retrospect, how would you rate the responsibility of each of the seven voters on a 5-point scale? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: voting result = AAABABA) Please write down seven (7) numbers, and/or a couple of sentences explaining your decision.** *

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He pushes the start button, and the reels begin to roll...

22. **The first reel stopped. It shows a cranberry. At this point, do you think Andrew will get the cash reward?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

23. **If Andrew ended up getting the cash reward, how much would you attribute the win to the first reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

24. **If Andrew doesn't get the cash reward, how much would you attribute the loss to the first reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

25. **The second reel stopped. It shows a cranberry. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CC, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

26. **If Andrew ended up getting the cash reward, how much would you attribute the win to the second reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

27. **If Andrew doesn't get the cash reward, how much would you attribute the loss to the second reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

28. **The third reel stopped. It shows a cranberry. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CCC, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

29. **If Andrew ended up getting the cash reward, how much would you attribute the win to the third reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

30. **If Andrew doesn't get the cash reward, how much would you attribute the loss to the third reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

31. **The fourth reel stopped. It shows a melon. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CCCM, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

32. **If Andrew ended up getting the cash reward, how much would you attribute the win to the fourth reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

33. **If Andrew doesn't get the cash reward, how much would you attribute the loss to the fourth reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

34. **The fifth reel stopped. It shows a cranberry. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CCCMC, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

35. **If Andrew ended up getting the cash reward, how much would you attribute the win to the fifth reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

36. **If Andrew doesn't get the cash reward, how much would you attribute the loss to the fifth reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

37. **The sixth reel stopped. It shows a melon. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CCCMCM, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

38. **Now that you know Andrew is getting the cash reward, how much would you attribute the win to the sixth reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

Story 2: Random Machine

A new machine has been invented. It has 7 (seven) reels, each of which has only 2 (two) symbols, namely cranberry and melon. All reels rotate independently of each other. The objective is to win reward from the machine. For this particular kind, the winning combinarions of symbols are those that have ***** at least four cranberries *****, and the order doesn't matter. If the player gets the winning combination, they will be rewarded with $10 in cash. The player can only play one hand at a time.

Andrew has been hired to test play the new machine described above. He has pushed the start button.

39. **The seventh (last) reel stopped. It shows a cranberry. At this point, do you think Andrew will get the cash reward? (reminder: symbols so far = CCCMCMC, C = cranberry, M = melon)** *
Mark only one oval.

	1	2	3	4	5	6	7	
Definitely no	◯	◯	◯	◯	◯	◯	◯	Definitely yes

40. **Now that you know Andrew is getting the cash reward, how much would you attribute the win to the seventh reel?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Full/Near full credit

41. **In retrospect, how would you attribute the win to each of the seven reels on a 5-point scale? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: symbols overall = CCCMCMC) Please write down seven (7) numbers, and/or a couple of sentences explaining your decision.**

Story 3: Trip or (No) Treat

Mr. Potter has three kids, and he would like to take them to a Halloween trip. Recently, his wife Ginny introduced a new reward system for their kids. The system works like this: The kids get a total of 10 points among them at the beginning of every week. Every time someone breaks something in the household, they get one point off (-1); every time someone cleans a room in their house, they get one point (+1). Ginny suggests that they take a look at how many points in total their kids have got over the past week, and if there is at least a total of 0 point among all three kids, they can go on a Halloween trip. Otherwise the kids have to stay at home for Halloween (no Trick-or-Treat!).

Most of the time, each of the three kids loses zero points over a week, because they generally break one thing in a room at a time. When they break something in a room, they will clean that room, hoping their parents wouldn't realize what they have done (which is never the case). Otherwise, they don't clean any room in the house. Therefore, most of the time, the kids have a total of 10 points among them by the end of the week.

Mr. Potter agrees with Ginny. He first goes into James' room to see how many points he's got.

42. **James got -4 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 6 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

43. **If you ended up not able to take your kids to this Halloween trip you really want to go to, how much would you (secretly) blame James?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 3: Trip or (No) Treat

Mr. Potter then goes into Albus' room to see how many points he's got.

44. **Albus got -4 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 2 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

45. **If you ended up not able to take your kids to this Halloween trip you really want to go to, how much would you (secretly) blame Albus?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 3: Trip or (No) Treat

Finally, Mr. Potter goes into Lily's room to see how many points she's got.

46. **Lily got -3 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have a total of -1 point)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

47. **Now that you know for a fact that you can't take your kids to the Halloween trip, how much would you (secretly) blame Lily?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

48. **In retrospect, how would you rate the responsibility of each of the three kids on a 5-point scale ? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: the scores were -4 (James), -4 (Albus), and -3 (Lily)) Please write down three (3) numbers, and/or a couple of sentences explaining your decision.**

Story 4: RE: Trip or (No) Treat

Mr. Potter is disappointed. The kids just happened to misbehave a lot over the past week, which really is very unusual for them (maybe they are just getting super hyped for Halloween?)

Fortunately, there is still time till Halloween. Mr. Potter talked to Ginny, and they agreed to observe the kids for one more week, and then use the same criteria (i.e. if the three kids get a total of 0 point or above, they will go on a Halloween trip) to decide their plan for Halloween. The points accumulated from the previous week are cleared, and the kids start with 10 points among them again.

One week later, Mr. Potter walks into James' room again to find out how many points he's got over this week.

49. **James got -1 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 9 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

50. **If you ended up taking your kids to this Halloween trip you really want to go to, how much would you (secretly) PRAISE James?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 4: RE: Trip or (No) Treat

Mr. Potter then goes into Albus' room to see how many points he's got.

51. **Albus got -4 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 5 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

52. **If you ended up taking your kids to this Halloween trip you really want to go to, how much would you (secretly) PRAISE Albus?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 4: RE: Trip or (No) Treat

Finally, Mr. Potter goes into Lily' room to see how many points she's got.

53. **Lily got -1 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have a total of 4 points)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

54. **Now that you know for a fact that you can take your kids to the Halloween trip, how much would you (secretly) PRAISE Lily?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

55. **In retrospect, how would you rate the responsibility of each of the three kids on a 5-point scale ? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: the scores were -1 (James), -4 (Albus), and -1 (Lily)) Please write down three (3) numbers, and/or a couple of sentences explaining your decision.**

Story 5: RE: Trip or (No) Treat

Mr. Potter has three kids, and he would like to take them to a Halloween trip. Recently, his wife Ginny introduced a new reward system for their kids. The system works like this: The kids get a total of 10 points among them at the beginning of every week. Every time someone breaks something in the household, they get one point off (-1); every time someone cleans a room in their house, they get one point (+1). Ginny suggests that they take a look at how many points in total their kids have got over the past week, and if there is at least a total of 0 point among all three kids, they can go on a Halloween trip. Otherwise the kids have to stay at home for Halloween (no Trick-or-Treat!).

Most of the time, each of the three kids loses zero points over a week, because they generally break one thing in a room at a time. When they break something in a room, they will clean that room, hoping their parents wouldn't realize what they have done (which is never the case). Otherwise, they don't clean any room in the house. Therefore, most of the time, the kids have a total of 10 points among them by the end of the week.

Mr. Potter agrees with Ginny. He first goes into James' room to see how many points he's got.

56. **James got -4 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 6 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

57. **If you ended up taking your kids to this Halloween trip you really want to go to, how much would you (secretly) BLAME James?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 5: RE: Trip or (No) Treat

Mr. Potter then goes into Lily's room to see how many points she's got.

58. **Lily got -3 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 3 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

59. **If you ended up not able to take your kids to this Halloween trip you really want to go to, how much would you (secretly) BLAME Lily?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 5: RE: Trip or (No) Treat

Finally, Mr. Potter goes into Albus' room to see how many points he's got.

60. **Albus got -4 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have a total of -1 point)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

61. **Now that you know for a fact that you can't take your kids to the Halloween trip, how much would you (secretly) BLAME Albus?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

62. **In retrospect, how would you rate the responsibility of each of the three kids on a 5-point scale ? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: the scores were -4 (James), -3 (Lily), and -4 (Albus)) Please write down three (3) numbers, and/or a couple of sentences explaining your decision.**

Story 6: RE: Trip or (No) Treat

Mr. Potter has three kids, and he would like to take them to a Halloween trip. Recently, his wife Ginny introduced a new reward system for their kids. The system works like this: The kids get a total of 10 points among them at the beginning of every week. Every time someone breaks something in the household, they get one point off (-1); every time someone cleans a room in their house, they get one point (+1). Ginny suggests that they take a look at how many points in total their kids have got over the past week, and if there is at least a total of 0 point among all three kids, they can go on a Halloween trip. Otherwise the kids have to stay at home for Halloween (no Trick-or-Treat!).

Most of the time, each of the three kids loses zero points over a week, because they generally break one thing in a room at a time. When they break something in a room, they will clean that room, hoping their parents wouldn't realize what they have done (which is never the case). Otherwise, they don't clean any room in the house. Therefore, most of the time, the kids have a total of 10 points among them by the end of the week.

Mr. Potter agrees with Ginny. He first goes into James' room to see how many points he's got.

63. **James got -3 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 7 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

64. **If you ended up not able to take your kids to this Halloween trip you really want to go to, how much would you (secretly) blame James?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 6: RE: Trip or (No) Treat

Mr. Potter then goes into Lily's room to see how many points she's got.

65. **Lily got -3 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have at most 4 points now)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

66. **If you ended up not able to take your kids to this Halloween trip you really want to go to, how much would you (secretly) blame Lily?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

Story 6: RE: Trip or (No) Treat

Finally, Mr. Potter goes into Albus' room to see how many points he's got.

67. **Albus got -3 points. If you were Mr. Potter, how likely is it that you will be able to go on a trip with your kids for Halloween? (reminder: since they started with 10 points total, the kids have a total of 1 point)** *
Mark only one oval.

	1	2	3	4	5	
Impossible	◯	◯	◯	◯	◯	Definite

68. **Now that you know for a fact that you are able to take your kids to the Halloween trip, how much would you (secretly) blame Albus?** *
Mark only one oval.

	1	2	3	4	5	
None/Very little	◯	◯	◯	◯	◯	All of it

69. **In retrospect, how would you rate the responsibility of each of the three kids on a 5-point scale ? Similar to above, 1 means "not at all" and 5 means "full/near full responsibility." (reminder: the scores were -3 (James), -3 (Lily), and -3 (Albus)) Please write down three (3) numbers, and/or a couple of sentences explaining your decision.**

That's the end of the experiment.

Thank you so much for your time! Let me know if you have any questions/concerns :D

Appendix B

CONSENT FORM

Responsibility Attribution Study Online Consent Form

This task is part of a research study conducted by Corey Zhou at Carnegie Mellon University and is funded by SURG from Undergraduate Research Office.

* Required

Summary

You will be asked to read several written short stories describing generic group activities such as voting and competition, and assign responsibility to individual players involved in each story with respect to some collective outcome on a numeric scale.

Purpose

The purpose of the research is to understand how people assign responsibility to individual contributors in a group activity, including how people make sense of responsibility, and how they decide what amount of responsibility is appropriate in different scenarios.

Procedures

You will be asked to read several short stories on a computer monitor. These stories involve everyday scenarios such as voting and competition, and are entirely fictional. As you read the story, you will be prompted by the computer to make judgments about individual contributors in that story. You will register your response using a mouse or computer keyboard.

You will participate in several tasks. During each task, you will be asked to read and evaluate one short written story, and make judgments about individual contributors described in that story.

You can take a short break in between the tasks if you would like to.

The anticipated time this study would take is about 30-60 minutes.

Participant Requirements

Participation in this study is limited to individuals age 18 and older.

Risks

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities. The only part of this experiment that you may find unpleasant is the repetitive nature of the tasks. You may become bored. Please do your best to stay alert and to process the reading materials as best you can. To mitigate this risk of boredom, you are always take a short break in between tasks if you would like to.

Benefits

There may be no personal benefit from your participation in the study, but the knowledge received may be of value to our knowledge of informal reasoning, and humanity as a whole.

Compensation & Costs

In appreciation for your time, you will be paid at $10.00 per hour for up to an hour.

There will be no cost to you if you participate in this study.

Future Use of Information

In the future, once we have removed all identifiable information from your data, we may use the data for our future research studies related to understanding human decision making behind responsibility attribution, or we may distribute the data to other researchers for their research studies on responsibility attribution in humans. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

The data captured for the research does not include any personally identifiable information about you. Your IP address will not be captured, either.

Your data and consent form will be kept separate. Your consent form will be stored digitally in a Google Drive associated with a CMU account, and will not be disclosed to third parties. The only people that have access to research records are the PI/experimenter and the Faculty Advisor. By participating, you understand and agree that the data and information gathered during this study may be used by Carnegie Mellon and published and/or disclosed by Carnegie Mellon to others outside of Carnegie Mellon. However, your name, address, contact information and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by Carnegie Mellon. Note that per regulation all research data must be kept for a minimum of 3 years.

Right to Ask Questions & Contact Information

If you have any questions about this study, you should feel free to ask them by contacting the Principal Investigator at yishanz@andrew.cmu.edu, or call/text (+1) 412-277-6924. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by phone or e-mail in accordance with the contact information listed above.

If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the Office of Research integrity and Compliance at Carnegie Mellon University. Email: irb-review@andrew.cmu.edu . Phone: 412-268-1901 or 412-268-5460.

Voluntary Participation

Your participation in this research is voluntary. You may discontinue participation at any time during the research activity. You may print a copy of this consent form for your records.

Age Confirmation

1. **I am age 18 or older.** *
   Mark only one oval.

   ◯ Yes
   ◯ No

Consent

2. **I have read and understand the information above.** *
   Mark only one oval.
   - ◯ Yes
   - ◯ No

3. **I want to participate in this research and continue with the tasks.** *
   Mark only one oval.
   - ◯ Yes
   - ◯ No

Thank you for filling out the consent form.

The experiment administrator will now give you further instructions on how to complete the experiment.