UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Probabilistic generative models for counterfactual reasoning and blame attribution

Permalink

https://escholarship.org/uc/item/5mj417cb

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 34(34)

ISSN

1069-7977

Authors

McCoy, John
Ullman, Tomer
Stuhlmuller, Andreas
et al.

Publication Date

2012

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Why blame Bob?
Probabilistic generative models, counterfactual reasoning, and blame attribution

John McCoy[1*] (jmccoy@mit.edu), Tomer Ullman[1*] (tomeru@mit.edu), Andreas Stuhlmüller[1] (ast@mit.edu), Tobias Gerstenberg[2] (t.gerstenberg@ucl.ac.uk) & Joshua Tenenbaum[1] (jbt@mit.edu)

[1]Department of Brain and Cognitive Sciences, MIT

[2]Cognitive, Perceptual and Brain Sciences, University College London, UK

*These authors contributed equally to the paper.

Abstract

We consider an approach to blame attribution based on counterfactual reasoning in probabilistic generative models. In this view, people intervene on each variable within their model and assign blame in proportion to how much a change to a variable would have improved the outcome. This approach raises two questions: First, what structure do people use to represent a given situation? Second, how do they choose what alternatives to consider when intervening on an event? We use a series of coin-tossing scenarios to compare empirical data to different models within the proposed framework. The results suggest that people sample their intervention values from a prior rather than deterministically switching the value of a variable. The results further suggest that people represent scenarios differently when asked to reason about their own blame attributions, compared with the blame attributions they believe others will assign.

Keywords: counterfactuals; blame attribution; probabilistic models; causal reasoning

Introduction

Alice and Bob play a coin-tossing game. If their coin tosses match, they win. Alice goes first and tosses heads, Bob goes second and tosses tails, and hence they lose. Who, if anyone, will be blamed? Counterfactually, what would have happened if Alice had tossed heads? One intuition is that how much someone will be blamed for an outcome is closely related to how strongly they affected the outcome (cf. Spellman, 1997). Through counterfactual thinking, people can reason how a change in the past would have affected the present and use such reasoning for cognitive tasks including social judgments, causal attribution, problem solving, and learning (see Roese, 1997; Byrne, 2002, for reviews). But how do people reason counterfactually? And what is the relationship between counterfactual thinking and blame attribution?

Psychological research on counterfactual reasoning has revealed factors that influence which events attract counterfactual thoughts, including unusual events (Kahneman & Miller, 1986), early events in a causal chain (Wells, Taylor, & Turtle, 1987), and late events in a temporal chain (Byrne, Segura, Culhane, Tasso, & Berrocal, 2000). There have also been formal accounts which aim to explain the empirical findings in terms of principled mental operations that do not depend on event features (Spellman, 1997; Byrne, 2002; Chockler & Halpern, 2004; Rips, 2010; Petrocelli, Percy, Sherman, & Tormala, 2011). Some of these formal models have been separately tested against empirical data (Sloman & Lagnado, 2005; Gerstenberg & Lagnado, 2010).

Kahneman and Tversky (1982) suggest that people reason counterfactually by using a "simulation heuristic", whereby they mentally alter events and run a simulation of how things would have gone otherwise given these changes. In this paper, we use a computational-level framework that formalizes the spirit of this suggestion: when attributing blame, people mentally alter each possible event in turn, consider the consequences for the outcome, and blame an event in proportion to how much the change would have improved the outcome.

We model this computation of counterfactual consequences using interventions on causal models (Pearl, 2000). We explore what causal models people use to represent the games in our experiments and how they choose alternatives when intervening on a particular event.

The plan for the paper is as follows. We first describe the formal framework this work is based on and the space of models we explore. We then report results of experiments in which we varied aspects of the coin-tossing game described above, and suggest a possible explanation for these results within our framework. We conclude by discussing implications and limitations of this account, and possibilities for future research.

Formal framework

We assume that, when reasoning counterfactually, people represent the situation they are reasoning about using a probabilistic generative framework. Probabilistic models have been used to explain many aspects of high-level cognition, including perception, prediction, decision making and social reasoning (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In this paper, we use causal Bayes nets and the functional equations they are derived from as the underlying probabilistic generative framework (Pearl, 2000). Other representations are possible—see, for example, Gerstenberg and Goodman (in prep) for an approach to counterfactual reasoning based on probabilistic programs.

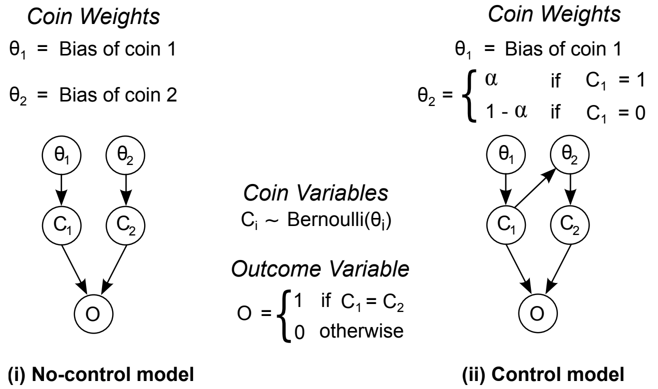
We model people's reasoning about blame as follows. First, consider each event in the situation—represented by a variable in a causal Bayes net—and intervene on it, i.e., consider a counterfactual value for this event ('do' in Pearl, 2000). Each such intervention results in a distribution over counterfactual worlds, where a counterfactual world is an assignment of values to variables. Next, compare these distributions to the actually observed world in order to assign blame. The variable being intervened on is assigned a degree of blame proportional to the difference in expected utility between the counterfactual world and the actual world. When interventions are chosen stochastically, we take the expectation over intervention values.

In our experiments, there are two possible outcomes, a win (1) and a loss (0), and the actual outcome is described as a loss. In this setting, our model of blame judgments is equivalent to assigning a degree of blame to a variable in proportion to the probability of reaching a counterfactual world in which the game is won after intervening on the variable. This combines the idea of using the 'do' operator as a psychologically plausible basis for counterfactual thinking (e.g., Sloman & Lagnado, 2005) with the idea of assigning causality by considering how each of the events in a given scenario affects the outcome (e.g., Spellman, 1997).

The framework as described presents at least two open questions about how people reason counterfactually to attribute blame.

What generative structure do people use to represent a situation? Even for simple scenarios, there exists a rich space of possible representations. Consider the coin-tossing game described in the introduction. Previous work suggests that people sometimes believe themselves or others capable of control over events such as coin tosses that are in fact random (Langer, 1975; Shafir & Tversky, 1992). Based on this work, we consider two simple generative models for the coin-tossing game, a *no-control* model that represents the coin tosses as independent, and a *control* model that represents the players as having some control over their tosses, assumes that they have knowledge of the previous player's toss, and that they try to match this toss (see Figure 1). This 'control' is captured in the following way: If the coin tossed by the first player came up 1 ('heads'), the bias for the second player's coin getting 1 is now $\alpha > 0.5$. If the first player tossed 0 ('tails'), then the bias for the second player's coin getting 0 is now $\alpha > 0.5$. When we compute counterfactuals in the 'control' setting, we use the following functional equations that reflect this idea: $u_1 = \text{Bernoulli}(\theta_1)$, $u_{2a} = \text{Bernoulli}(\alpha)$, $u_{2b} = \text{Bernoulli}(1-\alpha)$, $C_1 = u_1$, and $C_2 = u_{2a}$ if $C_1 = 1$, otherwise u_{2b}. We later discuss a perspective-dependent model, which combines the 'control' and 'no-control' models.

What new value do people assign to the intervened-upon variable? The 'do' operator produces counterfactual worlds when given a variable to intervene on and a value to set the variable to, but there is a question about what value people use. One possibility is that people set the intervened-upon variable to be different from what it was before the intervention. The idea of *only* considering alternatives to reality when reasoning counterfactually has intuitive appeal. This approach is taken by Gerstenberg, Goodman, Lagnado, and Tenenbaum (2012) in reasoning counterfactually about physical events. In the case of binary variables, as in the current study, this involves simply switching the observed value, hence we refer to this way of choosing intervention values as *intervene-switch*. The generalization to non-binary variables is not immediate and is not considered here.


Coin Weights
θ_1 = Bias of coin 1
θ_2 = Bias of coin 2
θ_1
θ_2
C_1
C_2
O
Coin Variables
$C_i \sim \text{Bernoulli}(\theta_i)$
Outcome Variable
$O = \begin{cases} 1 & \text{if } C_1 = C_2 \\ 0 & \text{otherwise} \end{cases}$
Coin Weights
θ_1 = Bias of coin 1
$\theta_2 = \begin{cases} \alpha & \text{if } C_1 = 1 \\ 1-\alpha & \text{if } C_1 = 0 \end{cases}$
θ_1
θ_2
C_1
C_2
O

(i) No-control model

(ii) Control model

Figure 1: Different generative structures for the coin-tossing game. Coin tosses are drawn from a Bernoulli distribution with the coin bias as its parameter. 'heads' and 'tails', and 'win' and 'lose' are replaced by 1 and 0. **(i)** In the 'no-control' version the bias is simply the unchanged bias of the coin. **(ii)** In the 'control' version players are represented as having some influence over the coin, as is formalized in the text.

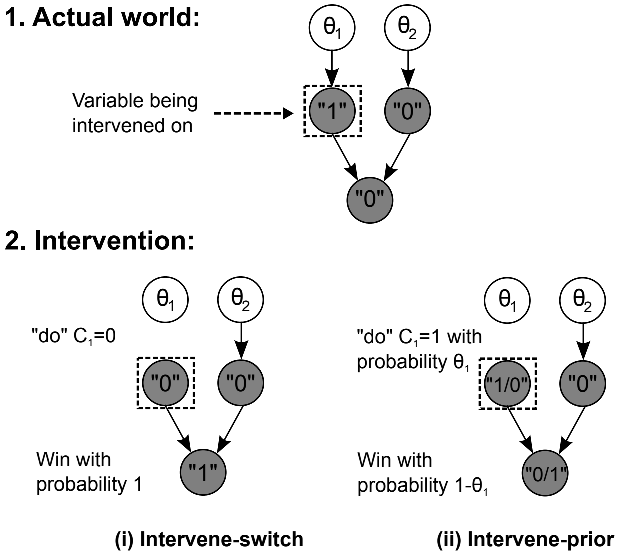
A second possibility is to draw the intervention value from the (conditional) prior over the variable being intervened upon. This allows an assessment of how unlikely the counterfactual event is, which has been stressed as an important factor by Petrocelli et al. (2011). We refer to this possibility for setting intervention values as *intervene-prior*. Figure 2 illustrates these two possibilities. A third possibility is that people choose an intervention value optimally, in such a way as to maximize the expected utility of counterfactual worlds. In our models for the coin-tossing domain, this possibility coincides with *intervene-switch*, hence we do not discuss it separately.

Each combination of these two factors—which causal structure to use, and how to choose an intervention value—results in a different model of blame attribution within our general framework. There are other factors which remain the subject of future research and which will almost certainly be necessary to predict subjects' judgments in richer situations. However, even these two factors offer a space for exploring how people reason counterfactually and assign blame.

A simple example of predicting blame judgments

Consider again the coin-tossing game with a fair coin, in which the player going first (C_1) tosses heads and the player going second (C_2) tosses tails, resulting in a loss. Who is to blame for the loss? We examine the predictions of four different models. In this scenario, some of the models make the same predictions; other scenarios used in our experiments provide additional discriminatory power.

1. *no-control/intervene-prior*: Coin tosses are independent



(i) Intervene-switch **(ii) Intervene-prior**

Figure 2: How to choose a new value for the intervened-upon variable, using the coin-tossing game as example. In the actual world the first player tossed heads (1) and the second player tails (0), and they thus lost. In intervening upon the first player's coin toss there are two possibilities: (i) *intervene-switch*: the first player's coin toss is changed to be the opposite of what it was (i.e. from 1 to 0), (ii) *intervene-prior*: intervention values are chosen from the prior and so there is a θ_1 chance of setting the first player's toss to heads and a $1-\theta_1$ chance of setting the first player's toss to tails.

and values for interventions are chosen by drawing them from the variable prior. We begin by considering an intervention to the variable C_1. Because the coin is fair, with probability 0.5 we choose the intervention 'tails', resulting in the outcome variable being assigned a 'win', as both players now tossed matching coins (counterfactually speaking). With probability 0.5 we choose the intervention 'heads', resulting in a 'loss'. So, by intervening on C_1 we improve the odds of winning from 0 (actual observation) to 0.5, and the amount of blame assigned to C_1 is 0.5. What about the second player's blame? For C_2 the process is the same as C_1, thus the blame for C_2 is also 0.5. Hence, this model predicts no difference in blame between C_1 and C_2.

2. *control/intervene-prior*: Players are represented as having some control over their coin (Figure 1(ii)) and interventions are selected in the same way as before. The intervention on C_1 works as described above, resulting in a 0.5 amount of blame. However, in intervening on C_2, we draw from a prior skewed towards 'heads' (the coin now has a bias $\alpha > 0.5$ towards heads). With probability α we choose the intervention 'heads' for C_2, and with probability $1-\alpha$ we choose 'tails'. This results in a 'win' with probability $\alpha > 0.5$, meaning C_2 will be blamed more than C_1.
3. *no-control/intervene-switch*: Coin tosses are independent and the value used for an intervention is different from the observed value of the variable. Because C_1 was observed to be 'heads', an intervention always sets it to 'tails', resulting in a 'win' with probability 1. This model also predicts no difference in blame between the players.
4. *control/intervene-switch*: Players are represented as having some control over their coin (Figure 1(ii)) and the value used for an intervention is different from the observed value of the variable. Using this model, the control players have does not make a difference, as the intervention on C_1 is always 'tails' (switching it from 'heads'), and the intervention for C_2 is always 'heads'. In a way similar to the previous variant, the probability for a 'win' resulting from intervening on either C_1 or C_2 is 1, and thus both players receive the same blame.

Experiment

Procedures and methods

One hundred subjects per condition were recruited on Mechanical Turk. Approximately twenty subjects were dropped in each condition for failing comprehension questions. We presented descriptions of simple scenarios involving blame attribution in the aforementioned coin-tossing game. All scenarios share the following:

(1) An introduction describing the game: Each person tosses a coin in turn. If all coins land the same, the players each receive \$1000, and otherwise receive \$0.

(2) Subjects were told the order of play, and the result of each coin toss (e.g., "Bob tosses his coin first. It comes up heads. Then you toss a coin. It comes up tails.").

(3) The end result was a loss.

(4) Subjects were asked, depending on condition, how much blame they attributed to the players, or how much blame they believed the players would attribute.

(5) Subjects responded using a discrete 1-7 scale, with 1 marked *minimal blame* and 7 marked *maximal blame*.

Results and discussion

We use the space of models we have discussed to examine predicted differences in blame attribution, and test these predictions using one-sided Student *t*-tests.

1. Same room, subject not involved Subjects read descriptions of Player 1 and Player 2 playing the coin game. Player 1 tosses heads, then Player 2 tosses tails. Subjects were asked how much Player 1 would blame Player 2, and how much Player 2 would blame Player 1. The results of these and subsequent experiments, as well as the model predictions, are shown in Figure 3. This first experiment replicates the temporality effect (Miller & Gunasegaram, 1990) in the sense that subjects believe that Player 1 will blame Player 2 more than Player 2 will blame Player 1 (p<0.0001, cf. Figure 3a). The only one of the four simple models under consideration which accords with these results (rows 3-6 in Figure 3) is the model which assumes causal control and draws intervention values from the prior. However, experimental results to be discussed shortly imply that the situation is more subtle. We first test the sensitivity of assumptions about causal control to knowledge about epistemic access (i.e. knowing the result of the the other player's coin toss).

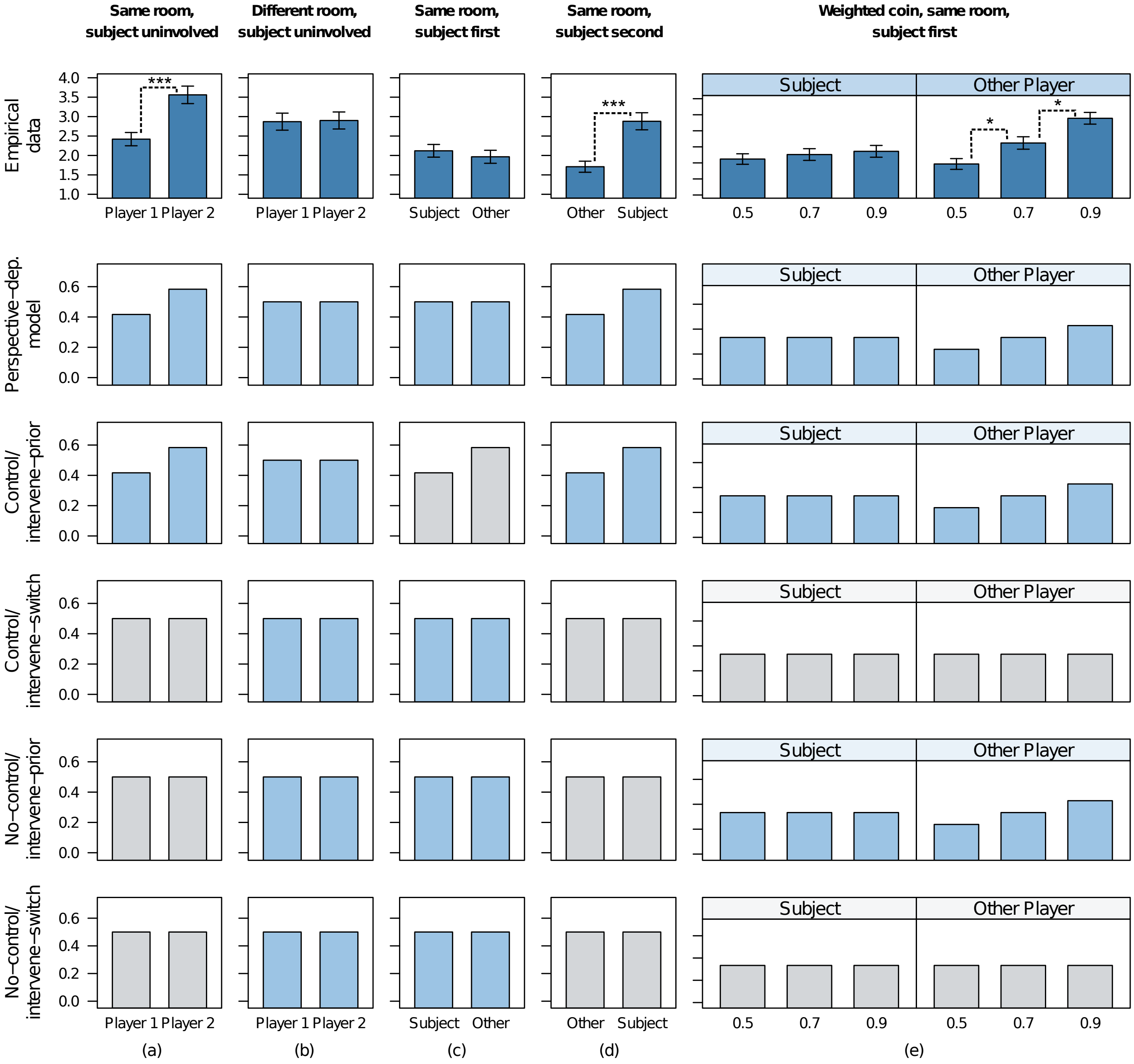


Figure 3: **Empirical results and model predictions for the coin-tossing game.** Columns represent separate experiments, with the empirical data shown in the top row. The y-axis is the mean amount of blame subjects believe will be assigned *to* the player on the x-axis, by the other player in the game. For example, in Experiment 1 subjects believe that Player 1 is assigned a 2.42 blame rating by Player 2, while Player 2 is assigned a 3.56 blame rating by Player 1. In each sub-figure, the player on the left side of the x-axis is the one going first. For example, in Experiment 1 Player 1 tosses the coin first. Model predictions are colored blue if they qualitatively match the ordinal blame judgments in the empirical results, and are grey otherwise. The amount of blame predicted by the models is normalized. We set $\alpha = 0.7$ in the above, but this makes no difference to the qualitative results.

2. Different rooms, subject not involved Subjects read descriptions of a game that only differed from the one in the previous experiment by players being in different rooms, such that Player 2 was unaware of the result of Player 1's coin toss. In this experiment, subjects predict that Player 1 will blame Player 2 significantly less than in the 'same room' scenario ($p<0.05$, cf. Figure 3b vs. a). There is no significant difference in the amount of blame attributed by the first and second player.

One explanation of these findings is that causal control is only believed possible when the second player is aware of the first player's coin toss. That is, subjects have a sophisticated model of the situation that treats the other players as agents that have the capacity for control, but that requires epistemic access for the agents to make use of this capacity.

Lack of epistemic access is not the only way to explain the results of this experiment. The two models based on setting intervention values by switching the observation (*intervene-switch*) are also consistent with these results. We will shortly provide independent evidence for sampling from the prior over switching, but first we manipulate whether subjects report their own blame judgments or their predictions about the judgments of others.

The 'classic' temporality effect replicated in Experiment 1 may strike some readers as odd. Surely one person is not to blame more than another in a purely random game? To examine this intuition, which is supported by previous research (Mandel, 2003), we now compare first-person and third-person blame judgments.

3. Same room, subject involved The game was described as in Experiment 1, but the subject was described as playing the game with another player (denoted 'Other Player' in Figure 3). One group of subjects was told that they tossed first, while another group was told that they tossed second. In both cases the player going first tosses 'heads' and the player going second tosses 'tails'. Subjects were asked how much they themselves would blame the other player, and how much they believed the other player would blame them. When subjects are asked about how much they think the other player will blame them, the temporality effect is replicated. That is, subjects believe that the other player will blame them significantly more when subjects flip second rather than first (p¡0.001 cf. blame to 'Subject' in Figure 3c and d).

Crucially, however, when comparing the amount of blame subjects attribute to the other player, we find no effect of position (cf. blame to 'Other' in Figure 3c and d). That is, the temporality effect does not exist when subjects are asked about the amount of blame they themselves attribute. As in Experiment 2, the *no-control/intervention-prior* model is consistent with the lack of a temporal effect, as are the two models in which intervention values are set by switching the observation. One hypothesis is thus that subjects do not themselves attribute causal control when assessing blame, but believe that other people do so (this possibility is represented by the 'perspective-dependent' model shown in the second row of Figure 3). It is also possible, however, that people think about choosing the intervention values differently depending on whether they are taking a first or third person perspective. That is, when considering how other people model a situation, they draw interventions from a prior, but when reasoning about their own perspective, they draw interventions by switching variables. The latter suggestion appears overly complex, but is not ruled out by the evidence presented so far.

We thus consider an experiment aimed specifically at examining whether people do set intervention values by sampling from the prior.

4. Biased coin, subject involved Subjects were told that they participated in the game with one other player. In one case, they were told that the other player used a coin biased 70% towards heads, in another case biased 90% towards heads. In both cases, subjects were told that they went first and got heads, and the other player tails. Subjects were asked how much they blamed the other player. We find that the greater the bias of the coin, the more the other player is blamed. The other player is blamed significantly more when the bias is 0.7 than when the coin is fair ($p < 0.05$, data from Experiment 3 were used for this comparison), and significantly more when the bias is 0.9 ($p < 0.05$; cf. blame to 'Other' in Figure 3e). Subjects' ratings about how they themselves will be blamed were not affected by the bias of Other's coin (cf. blame to 'Subject' in Figure 3e). For modeling this situation, we make the simplifying assumption that the second player's coin is entirely dependent on the specified bias, rather than on any causal control.

The models which are consistent with the effect of coin bias are those where values for the intervention are drawn from the prior. The results are analogous to experiments showing the tendency of people to focus on more unusual events when asked to reason counterfactually (Kahneman & Miller, 1986).

Given the small space of models we consider, one parsimonious account of the empirical data is that both when attributing blame themselves and when reasoning about how others attribute blame, people draw values for the intervention from the prior. Our results further suggest that people may assume that other people believe in causal control of random events, but that they themselves do not. This model of causal control seems to be sensitive to factors such as the other player's state of knowledge.

General discussion

We have explored aspects of a psychological framework for counterfactual reasoning, focusing in particular on its use for blame attribution. We have assumed that people represent the situation using probabilistic generative models and that they assign blame to an event by determining the counterfactual consequences of intervening on this event.

Within this framework, we have investigated what causal structure people use to represent a given situation and how people assign values to intervened-upon variables in the domain of coin-tossing games. In this domain, psychological effects such as the temporality effect and the tendency to focus on unusual events seem to arise naturally from counterfactual reasoning using probabilistic models.

We have presented data that suggests that people sample the values of their interventions from a prior rather than deterministically switching the variable to an alternate, unobserved value. Experimental evidence suggests that people may represent situations involving random events differently when reasoning about their own judgements compared to their predictions about the judgements made by others. People may model others as believing that such situations involve causal control, but they themselves may not believe in such control. One possible explanation for this perspective-dependent representation is that people model others' views in less detail than their own. For example, people may have a default assumption of causal control, but in situations such as coin games, they may be able to suppress this assumption. This suppression may take additional resources which, in general, may not be available or used when predicting how others represent the same situation.

This perspective-dependent difference in representation or computation provides an intriguing avenue for future research. For example, does there exist a similar difference in games of skill where control is the correct assumption? Beyond perspective-dependent differences, it is almost certain that different people model identically described situations in different ways, which suggests a per-subject analysis in addition to the aggregate approach taken in this paper.

There are many other ways in which our modeling and experimental results can be extended, even in the simple coin-tossing game. First, we do not explicitly take into account the prior probability of winning a game. Second, we do not explore situations in which the value of multiple variables would need to change to result in a win, for example a situation in which six people were playing the game with two people tossing tails and four people tossing heads. Third, while our model predictions are quantitative, we have restricted ourselves here to a qualitative analysis. Fourth, while we have modeled agents in some situations as having causal control, we did not give a full account of agents as having — and reasoning about—intentionality, foresight, and complex epistemic states, which are known to affect blame attribution (Lagnado & Channon, 2008). To capture the subtlety of human blame attribution and counterfactual thinking, richer models which include a more sophisticated representation of agents and their beliefs will be necessary.

Acknowledgments

This work was supported by ARO (W911NF-08-1-0242), ONR (N00014-09-0124), an NSF Graduate Fellowship (TDU) and a doctoral grant from the AXA research fund (TG).

References

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Science*, *6*(10), 426–431.

Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition*, *28*(2), 264–281.

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

Gerstenberg, T., & Goodman, N. (in prep).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, *132*(3), 419–434.

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology*, *59*(6), 1111–1118.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of personality and social psychology*, *100*(1), 30–46.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, *24*(4), 449–474.

Sloman, S., & Lagnado, D. (2005). Do we "do". *Cognitive Science*, *29*, 5–39.

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

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, *53*(3), 421–430.