

# Actor-Observer Asymmetries in Judgments of Intentional Actions

**Sarah Wellen (swellen@andrew.cmu.edu)**

Department of Philosophy, 135 Baker Hall, Carnegie Mellon University  
Pittsburgh, PA 15213 USA

**David Danks (ddanks@cmu.edu)**

Department of Philosophy, 135 Baker Hall, Carnegie Mellon University  
Pittsburgh, PA 15213 USA; and  
Institute for Human & Machine Cognition, 40 S. Alcaniz St.  
Pensacola, FL 32502

## Abstract

Much recent empirical research has explored the influence of moral evaluations on judgments about the intentionality of foreseeable side-effects of actions. Research on this ‘Side-Effect effect’ (also called the ‘Knobe effect’) has relied almost exclusively on vignette-based surveys, which have serious limitations when used in isolation. We present a novel behavioral methodology that tests the Side-Effect effect in two previously unexamined contexts: (i) judgments of real (rather than hypothetical) actions, and (ii) judgments about one’s own actions. The results suggest that judgments about one’s own actions tend to show a reverse Side-Effect effect: actors judge that (real) positive side-effects of their own actions are intentional whereas negative ones are not. The use of non-hypothetical situations also appears to attenuate the standard Side-Effect effect, which raises interesting challenges for standard theoretical accounts. These results provide preliminary evidence that the Side-Effect effect is driven by the same mechanisms underlying other asymmetries in causal attribution.

**Keywords:** Intentional action; Social cognition; Moral psychology; Actor-observer bias; Side-effect effect; Knobe effect; Vignettes

## Introduction

Notions of ‘intentionality’ and ‘causation’ are crucial in the way we understand our own and others’ mental lives, as well as a range of social interactions (e.g. Gergely, Nadasdy, Csibra, & Biro, 1994; Gopnik, *et al.*, 2004; Leslie & Keeble, 1987; Sloman, 2005; Woodward, Sommerville, & Guajardo, 2001). There has recently been a surge in research on folk understandings of these notions, much of it by experimental philosophers (e.g. Adams & Steadman, 2004a, 2004b; Knobe, 2003, 2004, 2006; Machery, 2008; Nadelhoffer, 2006; Nado, 2008; Uttich & Lombrozo, 2010). This research has revealed that folk judgments often display some surprising patterns. Perhaps the canonical example is the so-called ‘Side-Effect effect’ (also called the ‘Knobe effect’): experimental participants judge that a foreseeable side-effect of an action is more intentional when the side-effect is morally bad than when it is morally good. This paper develops a novel experimental method to investigate the Side-Effect effect that avoids some of the limitations of previous methods. Moreover, we report findings that

suggest that the Side-Effect effect is an instance of a much more general pattern of actor-observer biases, rather than a distinct phenomenon in its own right.

## The Side-Effect Effect

Consider a company chairman who acts to maximize profit, even though that action has the (foreseeable) side-effect of harming the environment. Suppose also that the chairman does not care about the environment; the occurrence of this side-effect is irrelevant (in any way) to his decision. Knobe (2003) found that 85% of participants judged that the chairman had nonetheless *intentionally* harmed the environment. When the program had the side-effect of helping the environment, though, only 23% of participants judged that the chairman had intentionally helped the environment (Knobe, 2003). In both vignettes, the chairman’s explicit goals (i.e., his explicit intentions) and actions remain the same, yet people’s judgments about whether he intentionally brought about the side-effect change significantly depending on the valence of the side-effect. This influence of outcome valence was present cross-culturally (Knobe & Burra, 2006), in four-year-old children (Leslie, Knobe, & Cohen, 2006), and for a range of mental state ascriptions besides simply intentionality (Pettit & Knobe, 2009). While there is general agreement that the intentions of the actor should influence our moral evaluations, it is less clear why the converse should hold. Moral concerns seemingly ought not to have a pervasive impact on what appear to be straightforward judgments about an actor’s mental states (though see Bratman, 1987 and Uttich & Lombrozo, 2010, for defenses of this influence).

Two different types of theories have been offered to explain the Side-Effect effect. Side-effect-centered theories hold that the different responses arise because of properties of the side-effect: some feature of negative or positive side-effects leads the actions to be judged as intentional or unintentional (respectively), or both. For instance, perhaps negative side-effects are always thought to be brought about intentionally (Knobe, 2004, 2006; but see also Pettit & Knobe, 2009, for an updated view), or negative cases require the actor to judge trade-offs (Machery, 2008). Alternately, the fact that an action violates the generalized

norm against bringing about negative outcomes might (rationally) be informative about the actor's mental state, since norm-violation typically requires additional reasons or intentional action (Uttich & Lombrozo, 2010).

In contrast, process-centered theories hold that our judgments would be symmetric in the two conditions, except that one of the situation-types (or both) elicits a mental process that alters or biases our judgment away from our typically symmetric intuitions. Various influencing processes have been suggested: a negative emotional reaction to the chairman (Nadelhoffer, 2006); a desire to blame based on conversational pragmatics of intentional language (Adams & Steadman, 2004a, 2004b); an asymmetric responsibility judgment (Wright & Bengson, 2009); or a distinct "moral mechanism" (Nado, 2008).

The Side-Effect effect has been studied almost exclusively using vignette-based surveys, which arguably have serious limitations when used in isolation. Vignettes focus on hypothetical situations in which many important (in the real world) properties and conditions are left unspecified. Exactly what information is included in (or omitted from) a vignette has been shown to have a substantial effect on people's judgments (Gugliermo & Malle, 2010; Mele & Cushman, 2007; Phelan & Sarkissian, 2008). Moreover, vignettes describe hypothetical situations, and previous research has found that judgments can differ substantially when provided in response to real rather than hypothetical situations. For instance, experimental participants reveal different utility functions when they are given real and hypothetical scenarios (List & Gallet, 2001; Murphy, Allen, Stevens, & Weatherhead, 2005; Neill, Cummings, Ganderton, Harrison, & McGuckin, 1994). Vignettes also require (almost always) that the experimental participant be an observer, not the actor. Thus, it is an open question whether the Side-effect effect occurs when an individual judges the intentionality of the side-effects of her own actions.

### Actor-Observer Asymmetries

Actor-observer asymmetries have been found in a wide range of domains. For instance, actors and observers largely attend to different aspects of a social interaction (Malle & Pearce, 2001). Actors are willing to engage in riskier behaviors than observers will condone (Fernandez-Duque & Wifall, 2007) and we appear to hold ourselves to different moral standards than we hold others to (Nadelhoffer & Feltz, 2008). One classic asymmetry is the Actor-Observer hypothesis in causal attribution (Jones & Nisbett, 1971): we tend to emphasize internal dispositional factors (e.g., attitudes, personality traits) as the causes of other people's actions, and aspects of the external situation (e.g., social constraints, situational factors) as the causes of our own actions. Recent research suggests the classic formulation is an oversimplification; instead of a simple internal/external distinction, actors tend to emphasize their reasons for acting (the beliefs, desires, and values that contributed to their decision to act), whereas observers tend to cite the causal

history of these reasons, including the attitudes, personality traits, and upbringing of the actor (Knobe & Malle, 2002).

A recent meta-analysis found little evidence for a general Actor-Observer asymmetry, but did find evidence for one that was mediated by the valence of the action (Malle, 2006). Observers tend to use more internal explanations than actors when explaining a negative action, but *fewer* internal explanations when explaining a positive action:

Table 1: Influences on causal attribution (adapted from Malle, 2006)

	Negative Valence	Positive Valence
<b>Overall</b>	$\text{diff}_A < \text{diff}_O$ <sup>1</sup>	$\text{diff}_A > \text{diff}_O$
<b>Real Situations</b>	$\text{diff}_A < \text{diff}_O$	$\text{diff}_A > \text{diff}_O$
<b>Hypothetical Situations</b>	$\text{diff}_A < \text{diff}_O$	$\text{diff}_A \approx \text{diff}_O$

Malle (2006) found that the valence-modified actor-observer bias was magnified when participants explained real events (middle row of Table 1). For hypothetical events, however, there was no difference between actors' and observers' explanations of positive actions. This suggests that, at least when explaining our own actions, it makes a difference whether the action is real or hypothetical.

This paper presents a novel experimental method in which participants judged real actions from either an actor or an observer role. We used a  $3 \times 3$  design that varies role {Actor, Observer, Motivated Observer} between-participant and side-effect valence {Neutral, Positive, Negative} within-participant. Side-effect-centered theories largely predict that the actor-observer manipulation should make no difference, since actors and observers should share the same conceptual asymmetries. The exception is the "Rational Scientist" view of Uttich & Lombrozo (2010), which arguably predicts that there should be no Side-Effect effect in the actor condition, regardless of outcome, since one presumably already knows one's own mental state. Most process-centered theories do not make a determinate prediction, since they leave unspecified whether and how the process that causes the asymmetry applies to judgments of one's own actions. Finally, if participants exhibit the asymmetric pattern of judgments characteristic of the valence-modified Actor-Observer hypothesis, then it is likely that the Side-Effect effect is due to the same mechanisms driving asymmetries in causal attributions.

## Experiment

### Participants

46 participants from the McGill, Concordia, and Carnegie Mellon University communities were divided into the Actor

<sup>1</sup> 'diff<sub>A</sub>' and 'diff<sub>O</sub>' refer to the difference between the number of internal and external reasons given by actors and observers, respectively, in the different conditions.

(N = 16), Observer (N = 15), and Motivated Observer (N = 15) conditions. Participants were mostly undergraduates; all were fluent in English and without cognitive deficits.

## Method

The experiment involved a computer game in which the participants (in the Actor condition) generated their own actions, discovered the consequences of these actions, and judged whether they had brought about those consequences intentionally. The game involved a computer interface where each movement of a joystick (up, down, left, or right) led to one or more colored balls being displayed on the computer screen. One of these balls (the red one) was the ‘goal ball’ and each time the actors got this ball they were rewarded with 10 tokens. Participants in the Actor and Motivated Observer conditions were told that they would redeem the tokens for money at the end of the experiment (but were not provided with the ‘exchange rate’); participants in the Observer condition were simply told that collecting tokens was the goal of the game.

The experiment had four phases. In the *Practice* phase, participants in every condition (Actor/Observer/Motivated Observer) played the computer game. During this phase, a joystick movement up deterministically produced a red ball (and so 10 tokens), while movements in all other directions produced a different color ball (and so no tokens). Each action thus produced a single consequence. Participants continued until they generated the red ball on six consecutive trials. At this point, the connections were shuffled (without notice) so that each movement produced a different color ball than before, and participants continued until they again generated the red ball six consecutive times by discovering and then moving in the new rewarded direction. The Practice phase continued until the movement-ball connections had been shuffled three times.

**Actor Condition** Participants in the Actor condition then moved (without notice) to the *Neutral* phase. This phase was identical to the practice phase, except that participants now received *two* balls after each joystick move. A move in the “rewarded” direction always produced both a red ball and a white ball, and every other move resulted in a random pair of unrewarded balls (white, yellow, blue, or green). The white ball thus played the role of a foreseeable side-effect: participants quickly learned that it always appeared with the red ball, but were indifferent to its occurrence. The white ball was introduced after the Practice phase to ensure that participants were motivationally neutral towards it.

The rewarded direction was randomly selected at the start of the Neutral phase. Once participants had made six consecutive moves in that direction, the rewarded direction would change and participants would have to rediscover it. The Neutral phase ended when the participant discovered the rewarded direction after the third change of ball-movement connections.

Participants then entered either the *Negative* or *Positive* phase (order counterbalanced between participants). In these

phases, participants were informed that generating the white ball now had consequences: a randomly chosen other participant in the experiment (i.e., *not* the participant) would either gain (Positive) or lose (Negative) three tokens for each white ball that the participant generated (no others were actually helped or harmed, however). As in the Neutral phase, participants continued until there had been four different rewarded positions. After completing one of these phases, Actor participants were informed of the change in the valence of the white ball, and then performed the other phase.

Throughout all three non-practice phases (Neutral, Positive, and Negative), the following two questions were asked at regular intervals:

Action question: You moved [direction]. How intentional was this?

Color question: You got a [color] ball. How intentional was this?

The Action question was always presented directly after a joystick move, and the Color question was presented after the ball display screen. We are particularly interested in responses about the white (side-effect) ball after participants had already discovered the currently rewarded direction. In this case, the participants presumably could foresee that they would get the white ball, so the white ball is a foreseen side-effect of their action. Answers were reported by clicking on a (540-point) rating scale with left-middle-right anchor points of “not at all intentional”, “somewhat intentional”, and “completely intentional”.

**Observer Condition** In the Observer condition, participants performed the practice phase, but then watched a video of an actor playing the Neutral phase instead of playing it themselves. Participants were told that the other individual was receiving real money when she got red balls. The actor’s hand, joystick, and computer screen were shown, but not her face. Whenever the actor was asked an Action question, participants observed her answer “completely intentional.” But instead of observing the actor answer for the Color question, the video paused and participants were asked:

Color question (Observer): Julie got a [color] ball.  
How intentional was this?

Participants were able to replay the video as much as desired before providing an answer. After the Neutral phase, participants observed the same actor in the Positive and Negative phases. Participants were told the relevant valence information, and the video order of the Positive and Negative phases was counterbalanced between participants. As a comprehension check, participants were also given a follow-up questionnaire asking whether the actor was trying to get the white ball in the Positive and Negative phases.

**Motivated Observer Condition** The Motivated Observer condition was identical to the Observer condition, except that participants were informed that they were the individual who had been randomly selected as the ‘other participant’ to

gain (Positive) or lose (Negative) three tokens whenever the actor got the white ball. They were reminded that tokens were redeemable for actual money.

## Results

There were no order effects, so different orderings were pooled for further analyses. Although the questions used a 540-point rating scale, responses largely concentrated around the three labeled points, indicating that many participants interpreted the scale as categorical. We thus transformed the data into a five-point scale, where each category corresponded to an equal-length segment of the original line. This transformation captured the generally categorical use of the line while preserving information about those few participants who used between-label points in their answers. Intentionality ratings for the goal (i.e., the red ball) were all at ceiling. In contrast, mean intentionality ratings for the foreseen side-effect (i.e., the white ball) are shown in Figure 1 (error bars indicate confidence intervals).

In the Actor condition, a planned, one-way within-participants ANOVA revealed a significant effect of phase valence ( $F(2, 42)=6.09, p<.05$ ). Post hoc Tukey HSD comparisons revealed that the Positive phase ( $M=3.18, SD=1.34$ ) led to significantly higher ratings than the Neutral ( $M=2.10, SD=1.36$ ) and Negative ( $M=2.09, SD=1.25$ ) phases,  $p<.05$ ; the latter two were not significantly different. Participants in the Actor condition thus rated the foreseen side-effect as significantly more intentional when it had a positive valence than a negative one, as found in standard actor-observer biases in causal attribution. A planned, one-way within-participants ANOVA did not reveal a significant effect of valence in the Observer ( $F(2, 42)=0.36, p=.69$ ) or Motivated Observer ( $F(2, 42)=0.48, p=.62$ ) conditions. The conditions were designed to also permit a statistical comparison of ratings from different conditions. One-way between-participants ANOVAs revealed significant effects of participant role for the Neutral ( $F(2, 42)=9.33, p<.05$ ) and Negative ( $F(2, 42)=6.97, p<.05$ ) phases, but not for the Positive phase ( $F(2, 42)=1.3, p=.28$ ). Actors had significantly different ratings than both the Observers and the Motivated Observers, except when the side-effect was positive.

Since every participant experienced all three valence phases, we can also analyze the rating patterns of each individual. We define ‘POS’ participants to be those who gave significantly higher intentionality ratings for the Positive phase compared to the Negative phase. As predicted by the valence-modified actor-observer hypothesis, the frequency of POS participants was significantly greater in the Actor condition (8/16 participants) than in the Observer condition (4/15 participants),  $\chi^2(1, N=30)=.96, p<.05$ , or Motivated Observer (2/15 participants) conditions,  $\chi^2(1, N=30)=1.0, p<.05$ . We can also approximate forced-choice probes (common in Side-Effect effect experiments) by further discretizing intentionality ratings into a binary variable (“Side effect was intended” iff rating  $\geq 3$ ). For the Actor

condition, the positive side-effect was significantly ( $p<.05$ ) more likely to have been rated as intentional than the negative one (75% vs. 38%). There was no corresponding difference in the Observer (86% vs. 73%;  $p>.05$ ) or Motivated Observer (73% vs. 86%;  $p>.05$ ) conditions.

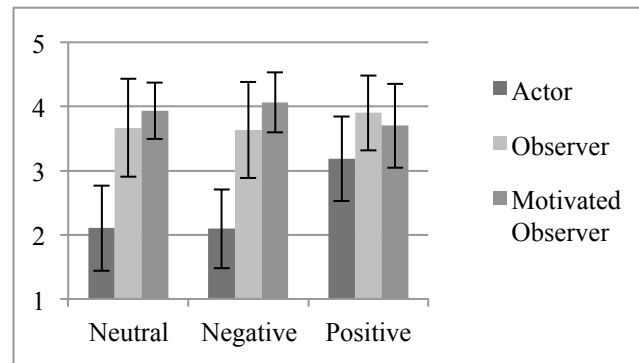


Figure 1: Mean intentionality ratings for foreseen side-effect (95% confidence intervals)

## Discussion

When asked to judge their *own* actions, participants were systematically more likely to judge those actions as intentional when the foreseen side-effect was positive rather than negative; that is, they exhibit something like a reverse Side-Effect effect. This pattern of results closely mirrors the valence-modified actor-observer hypothesis in causal attribution, where actors are more likely to cite internal dispositional factors as the causes of their positive actions than their negative ones. This suggests that the mechanisms driving the Side-Effect effect are plausibly the same ones driving asymmetries in causal attribution, although the nature of these mechanisms is still an open question.

Process-centered accounts focus primarily on the motivations of the observers. While these accounts have not made explicit predictions about the actor case, it is clear that in moral scenarios actors and observers often have different motivational pressures. Actors may have an incentive to minimize their responsibility (blameworthiness) for negative side-effects while overestimating their responsibility (praiseworthiness) for positive ones. Observers do not have these same concerns, and may even have incentives that tend in the opposite direction. However, we found that actors and observers differed in their intentionality judgments even for neutral side-effects, a result that cannot be explained by process-centered accounts. This difference is instead plausibly explained in terms of the epistemic differences between actors and observers. Actors have greater access to their own intentions, and thus are more aware that the side-effect (regardless of valence) was not among their goals. The observers do not have direct access to the mental states of the actor and thus they may rely more heavily on the behavior of the actor and the norms of the situation as a guide to her intentions.

This epistemic explanation fits nicely with the ‘Rational Scientist’ view, wherein epistemic differences between

norm-conforming and norm-violating cases drive the asymmetries in observer judgments. However, these epistemic considerations alone are unable to explain the reverse side-effect for actors, because actors have (at least perceived) direct access to their own mental states. It seems likely, then, that a combination of epistemic and motivational factors is necessary to explain the full range of results. Further research should explore this possibility.

Interestingly, we did not find a significant difference between judgments in the Positive and Negative phases in the Observer condition, as has typically been found using vignette studies. There are several possible reasons for this null result. One explanation is that we used visually presented situations involving real people rather than hypothetical vignettes. The abstract nature of vignettes potentially magnifies various distorting effects; for example, participants are perhaps less likely to believe the hypothetical chairman's reported intentions and more likely to be swayed by a moral evaluation of his actions. Another (not mutually exclusive) possibility is that the side-effect in the present experiment was not sufficiently negative or positive from the perspective of the observers. This explanation seems unlikely, however, since no significant difference was found even in the Motivated Observer condition where the side-effect directly affected the participant. Similarly, Feltz, Harris, & Perez (2010) found no significant Side-Effect effect for observer conditions using realistic settings, suggesting that our finding is not simply an artifact of our experimental design. Finally, an explanation could be given from the perspective of the 'Rational Scientist' view: the norms of computer games in psychological experiments differ from the norms of corporations affecting the environment, and it might be that the observers did not believe the actors in the experiment were violating any norms. This explanation is in some tension, however, with the fact that individuals were observed to violate the generalized norm not to harm others.

The results we report do not provide a definitive connection between the Side-Effect effect in vignette studies and judgments in realistic settings. The vignette results may plausibly generalize in cases where the side-effect valence is strongly negative or positive, when bringing about the negative side-effect involves violating a norm, or when there is greater psychological distance between the participant and the person whose actions are being judged (e.g., when reading a news article or listening to a description), although we currently have no evidence that this is the case. It is clear, however, that the Side-Effect effect should be situated within a more general class of valence-mediated actor-observer biases, and that the connections between these biases and folk psychological judgments warrant further investigation.

### Acknowledgments

Andrew Reisner, Lesley Fellows and Nathalie Camille provided valuable advice and assistance. We also thank Joshua Knobe, Thomas Nadelhoffer, and other participants

at the 2010 SPP Conference for helpful comments on an earlier version of this paper. This research was partially supported by the James S. McDonnell Foundation.

### References

- Adams, F., & Steadman, A. (2004a). Intentional action and moral considerations: Still pragmatic. *Analysis*, 64(3), 268-276.
- Adams, F., & Steadman, A. (2004b). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64(2), 173-181.
- Bratman, M. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Fernandez-Duque, D., & Wifall, T. (2007). Actor/observer asymmetry in risky decision making. *Judgment and Decision Making Journal*, 2(1), 1-8.
- Feltz, A., Harris, M., & Perez, A. (2010). Perspective in intentional action attribution: Reversing the side-effect effect. Manuscript Submitted for Publication.
- Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165-193.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 3-32.
- Gugliermo, S., & Malle, B. F. (in press). Can unintended side-effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*. In Press.
- Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelly, R. E. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190-194.
- Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, 64(2), 181-187.
- Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130(2), 203-231.
- Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, 6(1-2), 113-132.
- Knobe, J., & Malle, B. F. (2002). Self and other in the explanation of behavior. *Psychologica Belgica*, 42, 113-130.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25, 265-288.
- Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17(5), 421-427.
- List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and

- hypothetical stated values? *Environmental & Resource Economics*, 20(3), 241-254.
- Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, 23, 165-189.
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919.
- Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during social interaction: Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology*, 81, 278-294.
- Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31, 184-201.
- Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental & Resource Economics*, 30(3), 313-325.
- Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9(2), 203-220.
- Nadelhoffer, T., & Feltz, A. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics*, 1(2), 133-144.
- Nado, J. (2008). Effects of moral cognition on judgments of intentionality. *The British Journal for the Philosophy of Science*, 59, 709-731.
- Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W., & McGuckin, T. (1994). Hypothetical surveys and real economic commitments. *Land Economics*, 70(2), 145-154.
- Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgement. *Mind & Language*, 24(5), 586-604.
- Phelan, M. T. & Sarkissian, H. (2008). The folk strike back; Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291-298.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116, 87-100.
- Woodward, A., Sommerville, J., Guajardo, J. (2001). How infants make sense of intentional action. In B. F. Malle, L. J. Moses, D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge, MA: The MIT Press.
- Wright, J., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language*, 24(1), 24-50.