Available online at www.sciencedirect.com
ScienceDirect
Cognition 108 (2008) 353–380
COGNITION
www.elsevier.com/locate/COGNIT

Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment

Fiery Cushman

Department of Psychology, Harvard University, 1120 William James Hall, 33 Kirkland Street, Cambridge, MA 02138, United States

Received 9 March 2007; revised 29 February 2008; accepted 4 March 2008

Abstract

Recent research in moral psychology has attempted to characterize patterns of moral judgments of actions in terms of the causal and intentional properties of those actions. The present study directly compares the roles of consequence, causation, belief and desire in determining moral judgments. Judgments of the wrongness or permissibility of action were found to rely principally on the mental states of an agent, while judgments of blame and punishment are found to rely jointly on mental states and the causal connection of an agent to a harmful consequence. Also, selectively for judgments of punishment and blame, people who attempt but fail to cause harm more are judged more leniently if the harm occurs by independent means than if the harm does not occur at all. An account of these phenomena is proposed that distinguishes two processes of moral judgment: one which begins with harmful consequences and seeks a causally responsible agent, and the other which begins with an action and analyzes the mental states responsible for that action.

Keywords: Morality; Moral psychology; Punishment; Attribution theory; Intentional action; Theory of mind; Causation

E-mail address: cushman@wjh.harvard.edu

0010-0277/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2008.03.006

1. Introduction

On a snowy January Sunday, Hal and Peter watch football and share beers at a local bar. Both drive away intoxicated, and both lose control of their cars on the slick roads. Hal collides with a neighbor's tree, but Peter collides with a young girl playing in the snow. In the state of Massachusetts, Hal can expect a $250 fine for driving under the influence of alcohol. Peter faces a minimum of 2.5 years in prison – and up to 15 years – for vehicular manslaughter.

Cases like this have long puzzled philosophers and legal scholars (Hall, 1947; Hart & Honore, 1959; McLaughlin, 1925; Nagel, 1979; Williams, 1981), and it is easy to see why. Hal and Peter seem to have engaged in equally wrongful behavior, but poor luck in the case of Peter leads to a punishment that is radically more severe. Yet while we might be tempted to propose equal punishments for both drivers, it doesn't seem right to let Peter off with a $250 ticket for killing a girl, or to send Hal to prison for 2.5–15 years for hitting a tree. What cases like this reveal is a complex interaction of two different factors in our moral judgments: the assessment of *causal responsibility* for harm versus the assessment of *intent* to harm (along with related mental states such as beliefs, desires, and negligence). In particular, our judgments of the moral wrongness of a behavior seem to rely principally on an agent's mental state, while our judgments of deserved punishment show greater sensitivity to the harms actually caused by the agent.

This conceptual distinction between causal and intentional factors is recapitulated in the psychological literature on moral judgment. Researchers in the cognitive development tradition of moral psychology pioneered by Piaget have emphasized the dominance of intentional factors over causal factors in adult moral judgment (e.g. Hebble, 1971; Baron & Ritov, 2004; Shultz, Wright, & Schleifer, 1986; Yuill & Perner, 1988; Zelazo, Helwig, & Lau, 1996). The adult pattern of judgments is the outcome of a developmental shift: the moral judgments of young children are dominated by a causal analysis of harmful actions, while the moral judgments of older children and adults focus instead on the intention to produce harm. Critically, harmful intentions alone are found to be sufficient to warrant moral reprobation in mature children and adults, even in the absence of any harmful consequence.

This characterization of mature moral judgment stands in contrast to the basic model proposed in attribution theory, which owes greatly to a foundational work of Heider (1958). Attribution theorists have typically suggested that adult moral judgment begins by analyzing causal responsibility, only subsequently proceeding to an analysis of intention (Darley & Shultz, 1990; Fincham & Roberts, 1985; Heider, 1958; Shaver, 1985; Shultz, Schleifer, & Altman, 1981; Weiner, 1995). A central claim of this tradition is that, absent harmful consequence, malicious intentions are insufficient to trigger moral judgments of moral responsibility, blame and punishment in adults (reviewed in Weiner, 1995). As Darley and Shultz (1990) write, "judgments of moral responsibility presuppose those of causation. If the protagonist is judged not to have caused the harm, then there is no need to consider whether he is morally responsible for it."

Notably, the cognitive development and attribution literatures have adopted different dependent measures as their primary focus. Research in the attribution tradition usually asks subjects to evaluate the level of blame, moral responsibility and punishment deserved by an agent (e.g. Darley, Klosson, & Zanna, 1978; Fincham & Jaspers, 1979; Fincham & Shultz, 1981; Shultz et al., 1981). By contrast, research in the Piagetian tradition is more likely to ask subjects to judge whether an agent has behaved badly, wrongly, or naughtily (e.g. Hebble, 1971; Imamoglu, 1975; Nelson Le Gall, 1985; Piaget, 1965/1932; Wellman, Cross, & Bartsch, 1986). Of course, this parallels our intuitions in the cases of Hal and Peter: each seems to have acted equally wrongly, but on the basis of consequences their behaviors merit different punishments.

Experiments 1 and 2 of the present study build on this background, directly testing whether judgments of blame and punishment show enhanced sensitivity to the harm that an agent *causes*, while judgments of wrongness and permissibility show greater sensitivity to the harm that an agent *intends*. Such a result would challenge the commonsense notion that punishment is warranted if and only if an agent has behaved wrongfully, would explain a longstanding divergence between models of moral judgment developed in the cognitive development and attribution literatures, and would provide an empirical validation of the apparently divergent intuitions about wrongness and punishment generated in cases like those of Hal and Peter.

At the most general level, such a result would also raise important questions about how causal and intentional properties of an agent's behavior contribute to our moral judgments.[1] Recent research in moral psychology places this question at the center of inquiry (Alicke, 2000; Baron & Ritov, 2004; Cushman, Young, & Hauser, 2006; Hauser, 2006; Mikhail, 2000; Pizarro, Uhlmann, & Bloom, 2003). Decomposing moral judgment into causal and intentional analysis has also been the ambition of several decades of research in cognitive development and attribution theory (Baird & Astington, 2004; Heider, 1958; Karniol, 1978; Piaget, 1954; Shaver, 1985; Weiner, 1995). A frequent assumption of this research program, sometimes explicit but often unstated, is that moral judgment can be adequately described as a single process that integrates information about causal and intentional properties of harmful behavior. Experiments 3 and 4 challenge this assumption, asking whether causal and intentional information are evaluated in distinct processes of moral evaluation that act competitively to determine our judgments of wrongness, permissibility, blame and punishment.

[1] Throughout this essay the term "moral judgment" is used to refer globally to a broad class of evaluations that include wrongness, permissibility, punishment, blame and many other specific types of judgment. Part of the purpose of this essay is to demonstrate that these specific types of judgment are not identical, but there is still a useful commonsense notion according to which they all constitute moral judgments.

2. General methods

The following general methods were employed in Experiments 1–3.

2.1. Data collection

Subjects voluntarily logged on to the Moral Sense Test website (www.wjh.harvard.edu), which has been used in previous studies of moral psychology (Cushman et al., 2006, Cushman, Knobe, & Sinnott-Armstrong, 2008; Hauser, Cushman, Young, Jin, & Mikhail, 2007). In each experiment, a $2 \times 2 \times 2$ design (either belief $\times$ desire $\times$ consequence, or belief $\times$ desire $\times$ cause) generated eight unique combinations of factors. This $2 \times 2 \times 2$ design was applied to each of eight scenario contexts, such as a carnival, a sculpture class and a dentists office, yielding a total of 64 scenarios. Subjects were randomly assigned to eight scenarios from this set of 64 such that each subject viewed each of eight unique scenario contexts and each of eight unique combination of factors only once. The order of scenario presentation was random for each subjects. The full set of scenarios is available for download at moral.wjh.harvard.edu/methods.html.

Consistent with previous research (Cushman et al., 2006; Hauser et al., 2007), data was discarded from subjects who completed any question in fewer than four seconds, deemed to be the minimum possible comprehension and response time, as well as from subjects who provided clearly inaccurate demographic information (e.g. a 5-year-old with a post-graduate degree). For each condition of each experiment 160 non-discarded subjects were tested, totaling 1120 subjects across the seven conditions of the first three experiments. The methods used are all in accordance with the regulations of the institutional review board at Harvard University.

2.2. Statistical methods

In order to ensure that the observed effects generalized across scenario contexts, the present study treats the mean judgment of an individual scenario as the unit of analysis. This statistical approach has been employed in previous research into moral judgment (Cushman et al., 2006). For each combination of scenario context and factors – a total of 64 unique combinations – data was averaged across individual subjects' trials. A minimum of 10 trials were included in each average. A repeated-measures ANOVA was then performed treating the each scenario context as a single case, and the belief, desire and consequence variations as the repeated measures for each case. In addition to computing the F statistic, p-value and effect size as partial η^2 for each main effect and interaction, the sum of squares for each within-context main effect and interaction was calculated as a proportion of the total variability (the total within-context sum of squares, including within-context sum of squares error). Putting this statistical method in plain terms, the question asked was: what is the relative contribution of belief, desire and consequence/cause in shaping the average moral judgments of different situations, such as burning a persons hand, breaking their ankle, punching their nose, poisoning them, and so

forth? That is, what proportion of the total variability within a context does each factor explain?

As noted in the results reported below, when comparing between conditions this scenario-based analysis was supplemented with a second repeated-measures ANOVA that treated individual subjects as the unit of analysis. In each case this supplementary trail-based analysis confirmed the patterns of significance and non-significance observed in the scenario-based analysis.

3. Experiment 1

The first experiment explored whether judgments of wrongness and blame differ in their reliance on information about beliefs, desires and consequences. It was predicted that judgments of wrongness would show relatively greater sensitivity to mental state information, while judgments of blame would show relatively greater sensitivity to consequential information.

The decision to cross consequences with desire and belief independently – as opposed to confounding belief and desire into a single mental state factor – merits a brief explanation. Although the role of intentional attribution in moral judgment has been extensively investigated, a precise definition of intentional action is frequently absent. A standard definition in the philosophical literature holds that an action is intentional when its outcome is both desired and foreseen (i.e. "believed") (Forguson, 1989). An alternative account of intentionality in philosophy emphasizes the importance of a particular plan involving means to an end (Bratman, 1989). That is, an act is intentional only in the case that it brings about foreseen and desired outcome by the means that the agent planned. There is some evidence that moral judgments are sensitive to this additional criterion (Pizarro & Bloom, 2003). Adding further complexity the issue of intentionality in moral judgment is a recent and burgeoning literature suggesting that the folk concept of intentionality is dependent upon moral judgments (Knobe, 2003). That is, people appear to consider belief to be a sufficient criterion for intentional attribution in the case of morally blameworthy actions, but not in the case of morally praiseworthy actions.

Because of the confusion surrounding the proper definition of intentional action, the agents' belief and desire states are manipulated independently in the present study. This allows us to asses whether one factor contributes more strongly than the other, or whether they are each necessary conditions. At times, however, it will be most convenient to consider cases where beliefs and desires coincide, and in such cases the agent's behavior will be referred to as "intentional" (both belief and desire) or "unintentional" (neither belief nor desire).

3.1. Methods

Subjects were presented with eight moral scenarios that manipulated belief, desire and consequence in a $2 \times 2 \times 2$ design. The agents described in each scenario were explicitly stated to either believe or not believe that their action would cause a harm,

to either desire or not desire that their action would cause harm, and to either cause harm or cause no harm. In cases in which the agent did not cause harm, no harm occurred. An example of the parametric variation in a single scenario context follows:

Background

Jenny is taking a class in sculpture. She is assigned to work with a partner to weld together pieces of metal.

Desire

Jenny wants to burn her partner's hand.
or:
Jenny does not want to burn her partner's hand. Jenny only wants to weld together the metal.

Belief

Jenny thinks that if she welds a piece of metal that her partner is holding the heat will travel down the metal and burn her partner's hand.
or:
Jenny does not think that if she welds a piece of metal that her partner is holding the heat will travel down the metal and burn her partner's hand. Jenny thinks that the metal will weld without causing her partner any injury at all.

Consequence:

Jenny welds the metal, and her partner's hand is burned.
or:
Jenny welds the metal, but her partner happens to let go and is not burned at all.
The wrongness and blame questions were tested using a between-subjects design. Subjects in the wrongness condition were asked "How wrong was [agent]'s behavior?" and were given a seven-point response scale anchored at 1 with "Not at all", at 4 with "Somewhat", and at 7 with "Very much". Subjects in the blame condition were asked "How much blame does [agent] deserve?" and were given a seven-point response scale anchored at 1 with "None at all", at 4 with "Some", and at 7 with "Very much".

3.2. Results

For judgments of wrongness, the belief factor accounted for 62% of the variability, the desire factor for 21% of the variability, and the consequence factor for only 3% of the variability (Fig. 1a). All three factors contributed significantly to the model

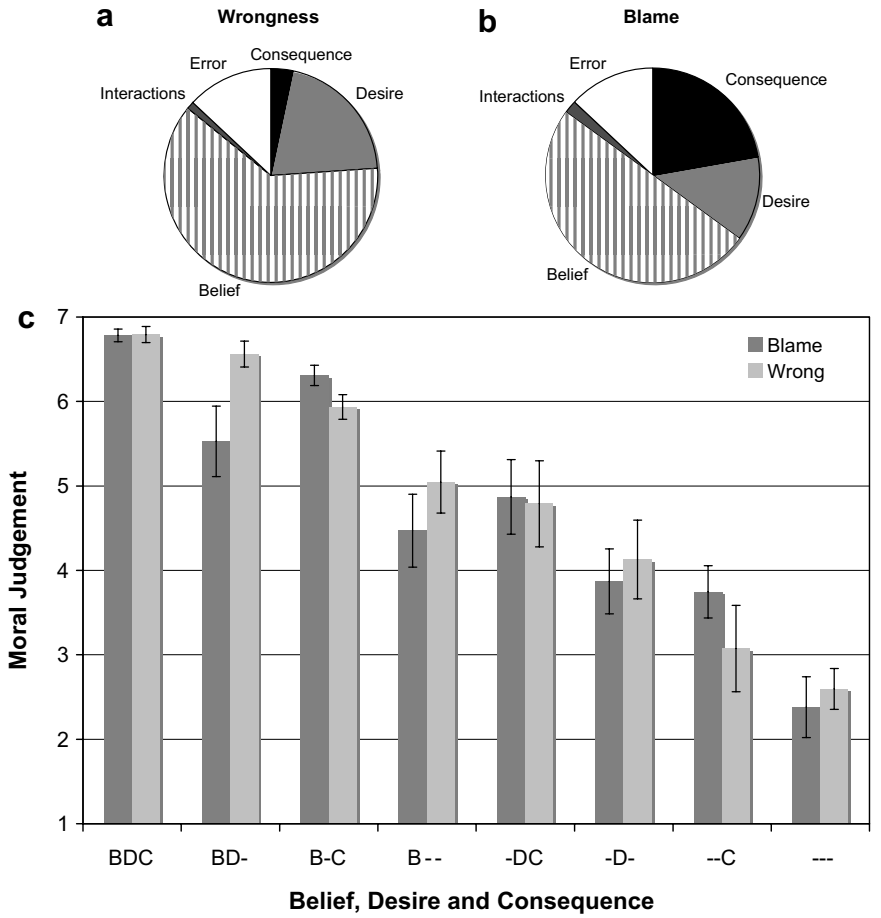

Fig. 1. Results of Experiment 1. (a) Proportion of within-context variability explained by each factor for the wrongness condition. (b) Proportion of within-context variability explained by each factor in the blame condition. (c) Direct comparison of mean moral judgment for each scenario context grouped by combination of belief, desire and consequence.

(Belief: $F(1,7) = 109.12$, $p < .001$, partial $\eta^2 = .94$; Desire: $F(1,7) = 137.39$, $p < .001$, partial $\eta^2 = .95$; Consequences: $F(1,7) = 8.09$, $p = .025$, partial $\eta^2 = .54$). The remaining 14% of variability was accounted for by interactions and error, although none of the interactions reached statistical significance at $p < .05$.

For judgments of blame, consequences played a much larger role in determining subject's judgments, accounting for 22% of the variability as compared to 3% for blame. The belief factor accounted for 50% of the variability, and the desire factor for 13% of the variability (Fig. 1b). All three factors contributed significantly to the model (Belief: $(F(1,7) = 67.61$, $p < .001$, partial $\eta^2 = .95$; Desire: $F(1,7) = 17.16$, $p < .001$, partial $\eta^2 = .88$; Consequences: $F(1,7) = 29.84$, $p = .001$, partial $\eta^2 = .79$). The remaining 15% of variability was accounted for by interactions

and error. Interactions between desire and belief and between desire and consequence both achieved significance, but in each case accounted for less than 0.1% of the within-context variability.

The wrongness and blame conditions were compared by combining data sets and including condition as an additional within-context factor. Significant condition-by-factor interactions were obtained for belief ($F(1,7) = 7.18$, $p = .032$, partial $\eta^2 = .51$), desire ($F(1,7) = 6.81$, $p < .035$, partial $\eta^2 = .49$), and consequence ($F(1,7) = 16.20$, $p = .005$, partial $\eta^2 = .70$). These results indicate that subjects used information about beliefs, desires and consequences differently when evaluating wrongness and blame. There was no significant main effect of condition ($F(1,7) = 1.85$, $p = .22$, partial $\eta^2 = .21$). This pattern of significant differences between the wrongness and blame conditions was confirmed by conducting a second repeated-measures ANOVA that treated individual responses to single trails, rather than mean responses to individual scenarios, as the unit of analysis.

Fig. 1c plots the average judgments in the wrongness and blame conditions for each combination of belief, desire and consequence. The largest difference between conditions was obtained in cases when the agent undertook an action believing that he would harm the victim and desiring to harm the victim, but in which no harmful consequence occurred. Comparing these cases to cases with identical belief and desire states but in which the harmful consequence did occur, there was a significant condition-by-consequence interaction ($F(1,7) = 10.44$, $p = .014$, partial $\eta^2 = .60$). When an agent intends harm, blame judgments are mitigated significantly more than wrongness judgments by the accidental failure to produce the harm.

A similar analysis was conducted on the mirror-image cases; that is, those cases in which an agent neither believes he will nor desires to cause a harm, but accidentally produces the harm. Comparing such cases to those with the identical belief and desire states, but in which the harmful consequence did not occur, there was a significant condition-by-consequence interaction ($F(1,7) = 7.95$, $p = .026$, partial $\eta^2 = .53$). When an agent intends no harm, blame judgments are enhanced significantly more than wrongness judgments by the accidental production of the harm.

3.3. Discussion

The results of Experiment 1 indicate that people rely on information about beliefs, desires and consequences differently when making judgments of wrongness and blame. Judgments about wrongness depended heavily on the belief state of the agent, but far less on the consequences of the agent's behavior. By contrast, judgments about blame depended substantially on both factors. These results suggest that superficially common dependent measures give rise to systematically different patterns of moral judgments.

For both the wrongness and blame conditions, the effects of belief, desire and consequence on judgments of blame were largely additive. This unexpected result is particularly notable in the blame condition. According to the standard attribution theory model of blame, both a causal role in bringing about harm and a intention

to produce harm are necessary conditions for blame, predicting large interaction terms in the model tested (reviewed in Weiner, 1995). Rather than integrating causal and intentional factors, however, subjects appeared to treat them as independent contributors to judgments of blame. Furthermore, in cases where the agent accidentally produced harm there was a significantly greater enhancement of blame than wrongness, as compared to cases where harm was neither intended nor occurred. This suggests that intention is not a necessary component for blame attribution—people will blame an agent (albeit slightly) for unintended consequences, and do so more readily than they will declare the same agent to have acted wrongly.

Finally, the results of Experiment 1 suggest that beliefs are relatively more important in determining moral wrongness and blame than are desires. This finding has considerable intuitive appeal: if I feed my guest a cake that I believe is poisonous I have done something wrong, even if I have no desire to harm my guest. On the other hand, if I feed my guest a cake that hope is poisonous I haven't done anything wrong, so long as I don't have the slightest belief that the cake actually *is* poisonous.

4. Experiment 2

The results of Experiment 1 suggest that judgments of wrongness and blame differ in their use of the information about the intentions and consequences attributable to an agent. But are these judgments truly representative of distinct and stable classes of moral judgment? This question is the focus of Experiment 2. If wrongness and blame exemplify distinct classes of moral judgment then the unique patterns of dependence on belief, desire and consequence observed in Experiment 1 should be replicable using alternative questions. Experiment 2 therefore tested the scenarios from Experiment 1 using two new questions: permissibility and punishment. It was predicted that the permissibility judgments would exhibit the signature pattern of dependency on belief found for wrongness judgments, while the punishment judgments would exhibit the signature pattern of increased sensitivity to consequences found for blame judgments. Subject's punishment judgments are of particular interest because several prominent theories in the attribution literature claim that punishment follows from the attribution of blame or moral responsibility to an agent (Fincham & Roberts, 1985; Shultz et al., 1981; Weiner, 1995).

4.1. Methods

Methods for Experiment 2 were identical to Experiment 1 except that subjects were tested using one of two different questions, permissibility or punishment. Subjects in the permissibility condition were asked to complete the statement "[Agent]'s behavior was:" and were given a seven-point response scale anchored at 1 with "Permissible" and at 7 with "Forbidden". Subjects in the punishment condition were asked "How much should [agent] be punished?" and were given a seven-point response scale anchored at 1 with "None at all", at 4 with "Some", and at 7 with "Very much".

4.2. Results

The proportions of variability explained by each factor in the permissibility condition of Experiment 2 were virtually identical to those observed in the wrongness condition of Experiment 1. The main effect of belief accounted for 63% of the within-context variability (compared to 62% in the wrongness condition), the main effect of desire accounted for 21% of the within-context variability (identical to the wrongness condition), and the main effect of consequences accounted for only 3% of the within context variability (also identical to the wrongness condition). All three factors contributed significantly to the model (Belief: $F(1,7) = 159.17$, $p < .001$, partial $\eta^2 = .96$; Desire: $F(1,7) = 71.014$, $p < .001$, partial $\eta^2 = .91$; Consequences: $F(1,7) = 17.66$, $p = .004$, partial $\eta^2 = .72$). The remaining 13% of variability was accounted for by interactions and error. The only interaction to reach significance was between desire and consequence; this interaction accounted for less than 1% of the variance in the model. These results suggest that the use of mental state versus consequence information was highly comparable across the permissibility and wrongness conditions.

Judgments of punishment exhibited a pattern of reliance on consequences far more similar to judgments of blame than either wrongness or permissibility (Fig. 2b). The main effect of consequence accounted for 20% of the variance in punishment, compared to 22% for blame, but only 3% for wrongness and permissibility. The main effect of belief accounted for 38% of the within-context variability, while the main effect of desire accounted for 30% of the within-context variability. All three factors contributed significantly to the model (Belief: $F(1,7) = 360.95$, $p < .001$, partial $\eta^2 = .98$; Desire: $F(1,7) = 88.53$, $p < .001$, partial $\eta^2 = .93$; Consequences: $F(1,7) = 113.73$, $p = .001$, partial $\eta^2 = .94$). The remaining 10% of variability was accounted for by interactions and error. None of the interactions between belief, desire and consequence was significant at $p < .05$.

Comparing cases where a harm was intended (believed + desired) and either the harm did occur or the harm did not occur, there was a significant condition-by-con-

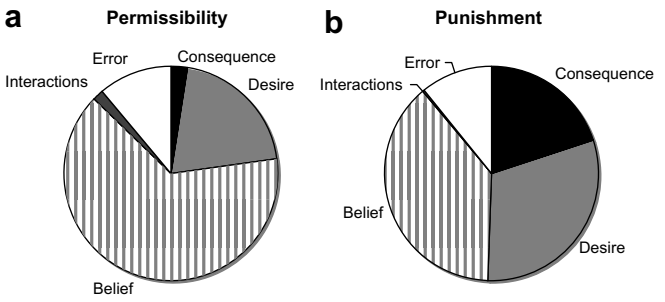

Fig. 2. Results of Experiment 2. (a) Proportion of within-context variability explained by each factor for the permissibility condition. (b) Proportion of within-context variability explained by each factor in the punishment condition.

sequence interaction for wrongness and punishment ($F(1,7) = 20.61$, $p = .003$, partial $\eta^2 = .75$), paralleling the identical result for wrongness and blame. When an agent intends harm, punishment judgments are mitigated significantly more than wrongness judgments by the accidental failure to produce the harm. A similar analysis was conducted on the mirror-image cases; that is, those cases in which an agent neither believes he will nor desires to cause a harm but accidentally produces it nevertheless. Comparing such cases to those with the identical belief and desire states, but in which the harmful consequence did not occur, there was a significant condition-by-consequence interaction for wrongness and punishment judgments ($F(1,7) = 6.01$, $p = .044$, partial $\eta^2 = .46$), again paralleling the comparison of wrongness and blame. When an agent intends no harm, punishment judgments are enhanced significantly more than wrongness judgments by the accidental production of harm.

4.3. Discussion

Experiment 2 tested whether the patterns of reliance on mental states versus consequences observed in Experiment 1 for wrongness and blame would extend to judgments of permissibility and punishment. Ratings of permissibility aligned nearly perfectly with ratings of wrongness in the proportion of variability explained by each of the three factors. In both conditions, only 3% of variability was accounted for by the consequence factor, while a large proportion of variability was accounted for by the belief factor (63–64%) and the desire factor (21%). This pattern of results suggests that subjects when people make judgments about whether a behavior was wrong, as well as when they make judgments about whether a behavior was permissible, they consider whether or not agent believed he would cause a harm and desired to cause the harm to be much more important than whether the harm actually occurred.

By contrast, subject's judgments were significantly more reliant upon consequences in the blame and punishment conditions, where the consequence factor explained 22% and 20% of the variability, respectively. Compared to wrongness, judgments punishment for agents with intent to harm are mitigated significantly more by the failure to produce harm, and judgments of punishment for agents with no intent to harm are enhanced significantly more by the accidental production of harm. The results of Experiment 2 provide further support for the claim that punishment arises from the attribution of blame, a common feature of attribution models of moral judgment (Fincham & Roberts, 1985; Shultz et al., 1981; Weiner, 1995).

The results of Experiment 2 also provide important evidence against a possible alternative account of the results of Experiment 1: that judgments of blame were influenced by consequences because of ambiguity in the meaning of the word blame. Blame often carries a moral connotation, but it is also possible to use the word blame to describe a causal relationship in the absence of a moral judgment – for instance, when we blame a cancelled picnic on the rain. Perhaps, then, the observed role of consequences in judgments of blame was driven by subjects who employed a purely causal, non-moral interpretation of the blame probe. The results of the punishment condition alleviate this concern; judgments of punishment were observed to substan-

tially on consequences, yet there is no natural interpretation of the punishment probe that picks out a causal but non-moral meaning.

The results of Experiments 1 and 2 echo a salient feature of moral development, fundamental to the stage theories of Piaget and Kohlberg. The morality of young children is marked by (1) the judgment of moral transgressions according to the consequences of behavior, and (2) a conception of morality as a system of punishments and rewards handed down by authority. Over the course of development, children undergo change on both of these fronts, (1) judging moral transgressions according to the intentions underlying behavior, and (2) conceiving of morality as a system of intrinsically valuable duties and constraints. Piaget and Kohlberg argued that the later-emerging moral theory replaced the early theory, but evidence from Experiments 1 and 2 suggests an alternative conclusion: there is a special connection between the assessment of consequences and the assignment of punishment, and this connection persists into adulthood.

This perspective motivates a division between two psychological processes: one emerges early in development, analyses causal responsibility (i.e. blame) for harmful outcomes, and supports judgments of deserved punishment, while the other emerges later in development, analyzes mental culpability and supports judgments of moral wrongness. These processes have unique inputs, rely on distinct analyses, and contribute to different classes of moral judgment. This later-emerging mental state analysis comes to constrain the punishment judgments of the earlier-emerging process of blame assignment – as is evident in the reliance on belief and desire information in the blame and punishment conditions of Experiments 1 and 2 – but never fully replaces it (Fig. 3).

The critical distinction between a single-process model and a two-process model is whether moral judgments arise only after causal and intentional information has been integrated (e.g. "if harm was caused intentionally, the act was immoral"), as is specified on a single-process model, or whether instead moral judgments arise via a competitive interaction between moral evaluations that draw from causal and intentional representations independently (e.g. "if a harm was caused, the act was immoral" versus "if the harm was intended, the act was immoral"), as is specified on a two-process model.

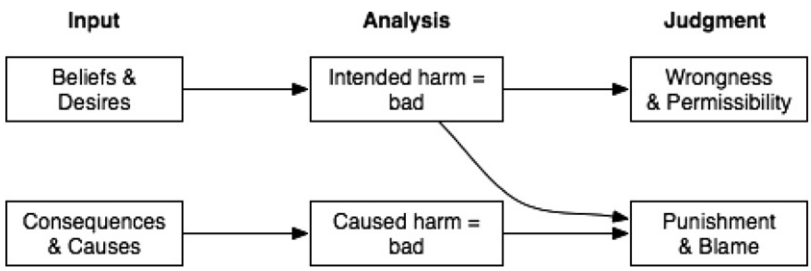

Fig. 3. A two process model of moral judgment.

One source of evidence for competition between distinct processes comes from functional neuroimaging (Young, Cushman, Hauser, & Saxe, 2007). Young and colleagues found neurological signatures of cognitive conflict selectively for cases of accidental harm: a circumstance where the putative causal process would assign blame on the basis of causal responsibility for harm, but where the putative intentional process would deny blame on the basis that the harm was unintended. On a two-process model cognitive conflict would be expected, as subjects mediate between conflicting moral judgments. On a single-process model, no opportunity for conflict arises: causal and intentional information are integrated as sufficient or necessary conditions and a single moral judgment follows. Experiments 3 and 4 provide further, novel evidence for competition between distinct processes of moral judgment.

5. Experiment 3

The results of Experiment 2 demonstrate that causal and intentional factors both play an important role in judgments of deserved punishment and blame. One possibility is that these factors are integrated into a single process of moral judgment. As noted above, however, there are reasons to suspect that causal and intentional factors contribute to distinct process of moral judgment and operate competitively in the assignment of punishment and blame. Experiment 3 is designed to test for such a competitive interaction. It is motivated by an unexpected finding in Experiments 1 and 2: the failure to find a significant interaction between the consequence factor and the belief and desire factors. Such an interaction would have been predicted on the standard attribution theory model, which posits causal responsibility as a necessary condition for blame and punishment. If no harmful event occurs, no causal responsibility can be attributed, and the agent should be fully exonerated without regard to his mental state. In contrast to this prediction, mental state information exerted an equally strong influence in cases of consequence and no consequence.

A potential explanation for this pattern of results is that (1) the causal process of moral judgment is only fully engaged in the presence of a harmful consequence and that (2) when it fails to be activated, the mental-state process competitively dominates resultant judgments of deserved punishment and blame. Specifically, when subjects were presented with cases of no harmful consequence in Experiments 1 and 2, they may have been prompted to rely relatively more heavily on an analysis of mental states because the option of causal analysis was foreclosed by the absence of any harmful consequence. When no consequence occurs, there is no causal analysis to be performed.

If this account is correct, subjects judgments of an agent's deserved punishment and blame should distinguish between cases when the agent causes no harm and no harm occurs, versus cases when the agent causes no harm but harm *does* occur by some independent mechanism. In cases of the latter type – when a harm occurs but the agent is not the cause – the causal process of moral judgment would be fully engaged, leading to the exoneration of the causally non-responsible agent. This, in turn, would yield the predicted interaction between causal and intentional factors

in judgments of punishment and blame: no punishment and blame when the agent is not causally responsible, and punishment and blame dependent upon mental state when the agent is causally responsible.

Experiment 3 was conducted identically to Experiments 1 and 2 except that no-consequence cases were substituted with no-cause cases: where no harm occurred at all in Experiments 1 and 2, the harm occurred but was caused by some independent mechanism in Experiment 3. Based on the above analysis, this manipulation should yield decreased judgments of punishment and blame in Experiment 3, as compared to Experiments 1 and 2. Moreover, a significant mental-state by cause interaction should be observed in Experiment 3. However, these differential patterns should emerge only for judgments of punishment and blame, which draw on the causal process of moral judgment. As a control condition, judgments of wrongness were also tested. Hypothesized to rely solely on an analysis of mental state, judgments of wrongness should not differ between Experiments 1 and 3.

It should be noted that the pattern of responses predicted by the two process model are quite counterintuitive. Why should somebody who attempts to cause severe bodily harm be judged less harshly when the harm is attributable to an independent source, as compared to when the harm does not occur? Why should this be predicted only for judgments of punishment and blame, but not for judgments of wrongness? The counterintuitive predictions of the two process model, along with the opportunity to demonstrate differential reliance on the causal process versus the mental-state process within a single class of moral judgment (blame and punishment), make Experiment 3 a stronger test of the proposed two-process model.

5.1. Methods

Methods for Experiment 2 were identical to Experiments 1 and 2 except that subjects were tested with scenarios where a harmful consequence always occurred, but the agent either was or was not the cause of the harmful consequence. As in Experiments 1 and 2, the agent's belief and desire were also manipulated for a fully crossed $2 \times 2 \times 2$ design. An example of the old "consequence" factor used in Experiments 1 and 2 is provided alongside the new "cause" factor used in Experiment 3:

Consequence (Experiments 1 and 2):

Jenny welds the metal, and her partner's hand is burned.
or:
Jenny welds the metal, but her partner happens to let go and is not burned at all.

Cause (Experiment 3):

Jenny welds the metal, and her partner's hand is burned.
or:
Jenny welds the metal, but her partner happens to let go and is not burned by Jenny. However, Jenny's partner picks up a different piece of hot metal and is burned.

5.2. Results

For judgments of wrongness, the pattern of results in Experiment 3 paralleled the pattern of results from Experiment 1: the belief factor accounted for a large proportion of variability while the cause factor accounted for a much smaller proportion of variability (Fig. 4a). In Experiment 3 the contributions of belief, desire and consequences were 64%, 18% and 6%, respectively, compared to 62%, 21% and 3% in Experiment 1. All three factors contributed significantly to the model (Belief: $F(1,7) = 90.78$, $p < .001$, partial $\eta^2 = .93$; Desire: $F(1,7) = 90.69$, $p < .001$, partial $\eta^2 = .93$; Consequence/Causation: $F(1,7) = 29.96$, $p = .001$, partial $\eta^2 = .81$). The remaining 12% of variability was accounted for by interactions and error. There was a significant Belief-by-Desire interaction, but it accounted for less than 1% of the variability. None of the other interactions reached statistical significance at $p < .05$.

The wrongness conditions in Experiments 1 and 3 were compared directly by combining data, treating consequence and causation as a single factor, and including condition as an additional within-context factor. There was a significant main effect for condition ($F(1,7) = .76.20$, $p < .001$, partial $\eta^2 = .92$), reflecting slightly lower wrongness ratings in Experiment 3. However, no significant interactions were obtained between condition and belief ($F(1,7) = .257$, $p = .638$, partial $\eta^2 = .04$),

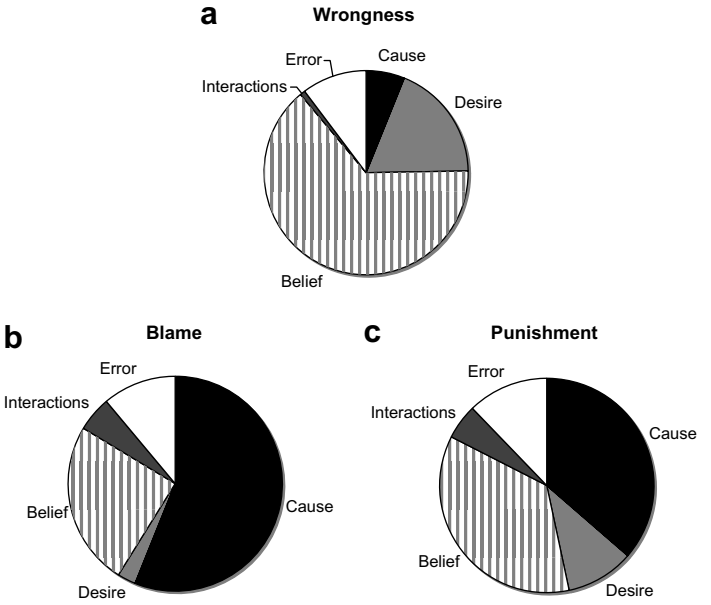

Fig. 4. Results of Experiment 3. (a) Proportion of within-context variability explained by each factor for the wrongness condition. (b) Proportion of within-context variability explained by each factor for the blame condition. (c) Proportion of within-context variability explained by each factor for the punishment condition.

desire ($F(1,7) = .717$, $p = .425$, partial $\eta^2 = .09$), or consequence/causation ($F(1,7) = 1.535$, $p = .255$, partial $\eta^2 = .18$). When probed on wrongness, subjects treated the consequence and causation factors nearly identically. The lack of significant differences between the consequence and causation conditions for the wrongness question in Experiments 1 and 3 was confirmed by conducting a second repeated-measures ANOVA that treated individual responses to single trails, rather than mean responses to individual scenarios, as the unit of analysis.

For judgments of blame, the pattern of results in Experiment 3 contrasted with the pattern of results from Experiment 1: 56% of the variability was explained by the cause factor in Experiment 3 compared to only 22% by the consequence factor in Experiment 1 (Fig. 4b). The main effect of belief accounted for 25% of the variability, as compared to 50% for Experiment 1, while the main effect of desire accounted for only 3% of the variability, as compared to 13% for Experiment 1. All three factors contributed significantly to the model (Belief: $F(1,7) = 44.09$, $p < .001$, partial $\eta^2 = .86$; Desire: $F(1,7) = 5.07$, $p = .008$, partial $\eta^2 = .65$; Consequence/Cause: $F(1,7) = 132.25$, $p < .001$, partial $\eta^2 = .95$). The remaining 17% of variability was accounted for by interactions and error. In contrast to the relatively small interactions in all conditions of Experiments 1 and 2, there was a significant belief-by-cause interaction that accounted 5% of the variability in the blame condition of Experiment 3. Subjects were most sensitive to beliefs when the agent did cause the harm, as compared to when the agent did not cause the harm. None of the other interactions reached statistical significance at $p < .05$.

The blame conditions in Experiments 1 and 3 were compared directly by combining data, treating consequence and causation as a single factor, and including condition as an additional within-context factor. There was a significant main effect for condition ($F(1,7) = 53.08$, $p < .001$, partial $\eta^2 = .88$), reflecting markedly lower blame ratings in Experiment 3. Additionally, significant interactions were obtained between condition and belief ($F(1,7) = 5.747$, $p = .048$, partial $\eta^2 = .45$), desire ($F(1,7) = 5.63$, $p = .049$, partial $\eta^2 = .45$), and especially consequence/causation ($F(1,7) = 37.08$, $p < .001$, partial $\eta^2 = .84$). When probed on blame, subjects treated the consequence and causation factors quite differently. In particular, subjects appeared to be far more willing to exonerate the agent from blame when the another agent caused a harmful consequence, as compared to when no harmful consequence occurred (Fig. 5a). The significant differences between the blame conditions of Experiments 1 and 3 were confirmed by conducting a second repeated-measures ANOVA that treated individual responses to single trails, rather than mean responses to individual scenarios, as the unit of analysis.

The difference between Experiments 1 and 3 on the punishment probe was broadly similar to the pattern revealed for the blame probe: a much larger proportion of variability was explained by the cause factor in Experiment 3 than the consequence factor in Experiment 1 (Fig. 4c). The main effect of cause on punishment in Experiment 3 accounted for 36% of the variability, as compared to 20% for the consequence factor in Experiment 1. The main effect of belief accounted for 36% of the variability as compared to 38% for Experiment 1, and the main effect of desire accounted for 10% of the variability, as compared to 30% for Experiment 1. All three

factors contributed significantly to the model (Belief: $F(1,7) = 41.90$, $p < .001$, partial $\eta^2 = .86$; Desire: $F(1,7) = 59.97$, $p < .001$, partial $\eta^2 = .90$; Consequence/Causation: $F(1,7) = 54.63$, $p < .001$, partial $\eta^2 = .95$). The remaining 18% of variability was accounted for by interactions and error. As with the blame condition, for the

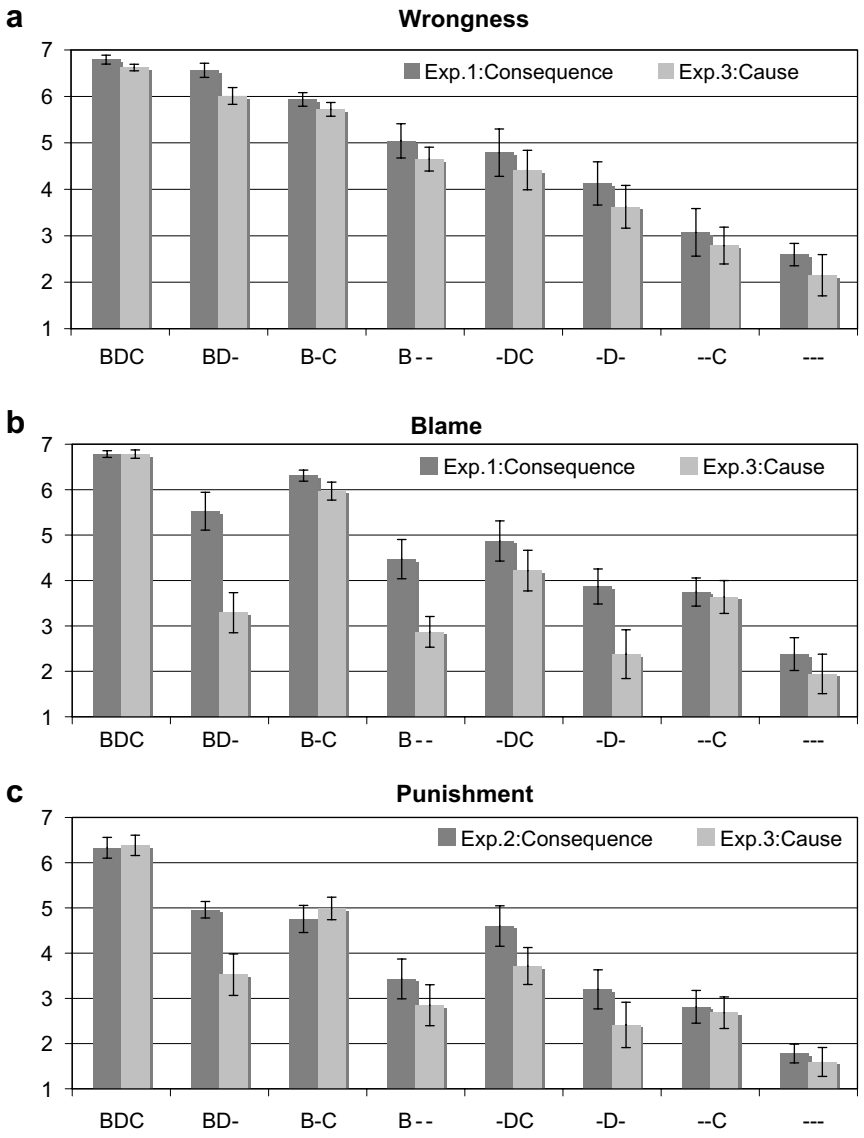

Fig. 5. Direct comparisons between consequence condition of Experiments 1 and 2 versus cause condition of Experiment 3 for (a) wrongness, (b) blame and (c) punishment. Mean moral judgments for each scenario context are grouped by combination of belief, desire and cause/consequence factors.

punishment condition there was a significant belief-by-cause interaction that accounted 5% of the variability, such that subjects were most sensitive to beliefs when the agent did cause the harm, as compared to when the agent did not cause the harm. None of the other interactions reached statistical significance at $p < .05$.

The punishment conditions in Experiments 1 and 3 were compared directly by combining data, treating consequence and causation as a single factor, and including condition as an additional within-context factor. There was a significant main effect for condition ($F(1,7) = 31.07$, $p < .001$, partial $\eta^2 = .82$), reflecting substantially lower punishment ratings in Experiment 3. Additionally, significant interactions were obtained between condition and consequence/causation ($F(1,7) = 9.89$, $p = .016$, partial $\eta^2 = .59$), as well as desire ($F(1,7) = 15.72$, $p = .005$, partial $\eta^2 = .69$), but not between condition and belief ($F(1,7) = 0.08$, $p = .791$, partial $\eta^2 = .01$). Paralleling the results with the blame probe, when probed on punishment subjects treated the consequence and causation factors quite differently. Subjects were more willing to exonerate the agent from blame when the another agent caused a harmful consequence, as compared to when no harmful consequence occurred (Fig. 5b). The significant differences between the punishment conditions of Experiments 2 and 3 were confirmed by conducting a second repeated-measures ANOVA that treated individual responses to single trails, rather than mean responses to individual scenarios, as the unit of analysis.

5.3. Discussion

Experiment 3 yielded two key results. First, subjects were less likely to assign blame and punishment to agents who attempted harm when the harm was brought about by some alternative means, as compared to when the harm failed to occur entirely (as in Experiments 1 and 2). Second, there was the significant interaction between belief and causation for both the blame and punishment conditions of Experiment 3: subjects were more sensitive to beliefs when the agent did cause the harm, compared to when the agent did not cause the harm. This interaction was observed only for the blame and punishment conditions, and not for the wrongness condition; furthermore, it was observed only in Experiment 3, and not in Experiments 1 and 2. It was, however, very robust, generalizing across eight different scenario contexts and two different dependent measures.

Both of these results can be understood by appeal to a causal process of moral judgment that assigns punishment and blame to agents causally responsible for harmful consequences. Such a process would lead to the exoneration of the non-responsible agent in Experiment 3, yielding the observed interaction between mental state and causal factors. But such a process would stumble against the no-consequence cases of Experiments 1 and 2. Absent any consequence, causal responsibility cannot be assigned, and the mental state process of moral judgment competitively dominates in these cases. This leads to harsher moral judgments in cases of malicious beliefs and desires, and eliminates the interaction of mental state and consequential factors.

In contrast to judgments of blame and punishment, the results of all three experiments suggest that when answering questions about wrongness and permissibility, subjects are only minimally influenced by causal/consequential information. Instead, they directly query the agent's mental state at the time of his or her behavior. This represents an alternative process of moral judgment: if an agent believes that his action will cause harm – and, to a lesser extent, if the agent desires to cause harm – then his action is wrong, and impermissible. Importantly, this process of moral judgment can be applied whether or not a harmful consequence actually occurs, and no matter who causes it.

The finding that wrongness judgments were unchanged between Experiments 1 and 3 constrains the available alternative explanations of the observed effects in the punishment and blame conditions of Experiment 3. The mitigation of blame and punishment judgments when an alternative causal means produced harm cannot be attributed to the 'distracting' influence of another blameworthy agent in the scenario, or the absorption of a fixed quantity of available condemnation by the blameworthy agent. Both of these explanations would apply as much to judgments of wrongness as judgments of blame and punishment. A two process model is better suited to account for the discrepancy between the wrongness and blame conditions: judgments of the wrongness of attempted harms are not mitigated by the presence of a harm caused by alternative means simply because judgments of wrongness do not depend upon the analysis of causal responsibility.

6. Experiment 4

Experiment 3 revealed that people assign less punishment to attempted crimes when harm coincidentally happens to befall the intended victim by some independent mechanism, as compared to when no harm befalls the victim at all. One explanation for this effect, which might be termed "blame blocking", holds that the evaluation of causal blame for a harmful event can block the consideration of an attempted harm-doer's malicious intent. However, the methodology employed in Expeirments 1–3 has several limitations. For instance, subjects responded to several items in a row, introducing task demands and interference effects between items, and scales were anchored in abstract terms (e.g. "very much punishment") that could have been interpreted differently between experiments.

Experiment 4 is designed to further test the blame blocking hypothesis, replicating the results of Experiment 3 while addressing these concerns. First, judgments of each case were collected in a between-subjects design, avoiding potential task demands or interference effects that might have influenced previous results. Second, responses were made on a scale anchored at each point with specific prison sentences, enhancing the comparability of response scales between conditions. Third, the harm caused in the "Harm" case – an allergic response to hazelnuts – was presented as an instance of bad luck unconnected to any specific perpetrator. This mitigates the risk that the punishment assigned to the athlete is diminished because punishment itself is a limited resource, and is assigned to some other malicious perpetrator.

In Experiment 4, subjects are presented with one of two scenarios in a between-subjects design. In the "No Harm" case, they read about an athlete who attempts to poison his rival by sprinkling poppy seeds on a salad which the rival has been served at a banquet, believing his rival to be allergic to the seeds. However, the rival is not allergic to poppy seeds and is unharmed. The "Harm" case is constructed identically, except that the salad happens to contain hazelnuts, the rival happens to be allergic to hazelnuts, and the rival therefore dies. The death of the rival by hazelnuts is therefore completely causally independent of the athlete's attempt to harm the rival by poppy seeds. Subjects used a nine-point scale specifying the degree of jail time deserved by the rival (e.g. "None", "7 years", "Life in prison"). This scale was modeled on research conducted by Robinson and Darley on folk intuitions concerning criminal liability (Robinson & Darley, 1995). In keeping with the results of Experiment 3 and the blame blocking model, subjects were predicted to assign longer prison sentences in the "No Harm" case relative to the "Harm" case.

6.1. Methods

Subjects voluntarily logged on to the Moral Sense Test website. They were presented with one of two scenarios in a between-subjects design:

"No Harm" Case:

Smith and Brown are two runners scheduled to compete in a championship race. Smith holds the current world record and is widely expected to win the race. Brown plans to eliminate Smith from the race.

Brown is absolutely sure that Smith is allergic to poppy seeds, and that eating poppy seeds will kill him. Brown decides to sprinkle some poppy seeds on Smith's food if Smith gets up to go to the bathroom. But it turns out that Brown is incorrect: Smith is not allergic to poppy seeds at all.

At the athlete's banquet a few days before the race, Smith takes a few bites of his salad and then gets up to go to the bathroom. Brown sprinkles poppy seeds on Smith's food. Smith comes back and finishes the salad. The poppy seeds don't harm Smith at all.

"Harm" Case:

Smith and Brown are two runners scheduled to compete in a championship race. Smith holds the current world record and is widely expected to win the race. Brown plans to eliminate Smith from the race.

Brown is absolutely sure that Smith is allergic to poppy seeds, and that eating poppy seeds will kill him. Brown decides to sprinkle some poppy seeds on Smith's food if Smith gets up to go to the bathroom. But it turns out that Brown is incorrect: Smith is not allergic to poppy seeds at all. Instead, Brown is fatally allergic to hazelnuts.

At the athlete's banquet a few days before the race, there are hazelnuts in the salad that Smith is served. Smith takes a few bites and then gets up to go to the bath-

room. Brown sprinkles poppy seeds on Smith's food. Smith comes back and finishes the salad. The poppy seeds don't harm Smith at all, but because of the hazelnuts, Smith dies.

Subjects were instructed to imagine that they were on a jury evaluating the case against Brown. They were asked then, "How much prison time does Brown deserve?" Responses were recorded on a 9-point scale anchored as follows: none, 6 months, 1 year, 2 years, 4 years, 8 years, 16 years, 32 years, Life. Subjects were excluded from analysis if they responded in fewer than 20 seconds, judged in pilot research to be the minimum possible comprehension and response time, or if they indicated that they had participated in previous versions of the Moral Sense Test.

6.2. Results

Because subject's responses were not normally distributed, and because the response scale included categorical values (e.g. no punishment; life in prison), all results were analyzed using nonparametric tests. A total of 200 usable responses to each scenario were analyzed. As predicted, subjects assigned more punishment in the "No Harm" case than the "Harm" case, but this effect was only marginally

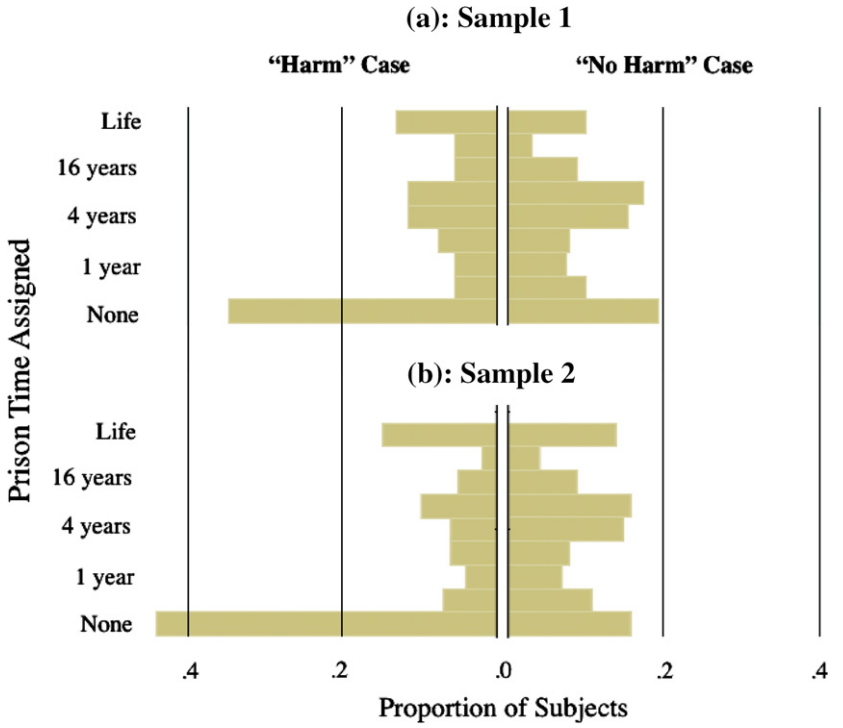

Fig. 6. Histograms representing the proportion of subjects assigning each punishment level in the "Harm" and "No Harm" conditions (depicted back-to-back). Histograms are presented separately for (a) the first sample ($N = 200$ per condition) and (b) the second sample of subjects ($N = 100$ per condition).

significant (Mann–Whitney Ranked Sums: $Z(400) = 1.6$, $p = .11$). However, inspection of the distribution of responses revealed a striking effect: subjects were nearly twice as likely to assign no punishment at all in the "Harm" case (34.5%) than in the "No Harm" case (19.5%; see Fig. 6a). This effect was highly significant (χ^2 (1, $N = 400$) 11.4, $p = .001$). By contrast, among those individuals who *did* assign punishment, there was no reduction in the "Harm" case – in fact, there was a marginally significant trend in the opposite direction (Mann–Whitney Ranked Sums: Z $(292) = 1.6$, $p = .10$).

A second sample of subjects was tested to confirm the observed effect that subjects in the "Harm" condition are more likely to assign no punishment (vs. any punishment at all), compared to subjects in the "No Harm" condition. Power analysis revealed that the observed effect should be replicable with a probability exceeding 90% by testing 100 subjects per condition. A second sample of this size was tested, and the observed effect was replicated (Fig. 6b): subjects were more than twice as likely to assign no punishment in the "Harm" case (45%) than in the "No Harm" case (16%; χ^2 (1, $N = 200$) 19.8, $p < .001$). As in the first sample, no significant difference in the amount of punishment assigned in each case was observed selectively among those subjects who assigned any punishment at all (Mann–Whitney Ranked Sums: Z $(139) = 0.9$, $p = .36$).

6.3. Discussion

The results of Experiment 4 confirmed the results of Experiment 3: the punishment assigned to attempted crimes is mitigated when harm befalls the intended victim by some alternative means, compared to when no harm befalls the victim at all. This result provides further support for the "blame blocking" model offered above. Moreover, in Experiment 4, this effect occurs in a very specific manner. Subjects were twice as likely to assign *no punishment at all* in the harm case relative to the no-harm case, but among those subjects who assign *any* punishment, the level of punishment assigned in each condition was identical. This highly specific effect was replicated in two independent samples of subjects; however, it depends in part upon the interpretation of a null result, and replicating it with new scenario texts remains an important direction for future research.

The specific pattern of results observed in Experiment 4 dovetails elegantly with the two-process model offered above. On this account, the occurrence of a harmful consequence in the "Harm" case triggers an assessment of causal responsibility that points away from the attempted harm-doer. This process of locating causal blame blocks any assessment of the attempted harm-doer's malicious mental state, and subjects consequently assign no punishment at all to the attempted harm-doer. However, among those subjects who do persist in assessing the harm-doer's mental state, punishment was observed to be equally strong in the harm and no-harm cases. These subjects are hypothesized to be relying on a mental-state analysis alone, and of course the mental state of the attempted harm-doer is identical in both the harm and no-harm cases.

At a broader level, the "blame blocking" phenomenon reported in Experiments 3 and 4 suggests that the occurrence of a harmful consequence triggers an analysis of

causal responsibility that plays a key role in the assignment of punishment. When causal responsibility points away from an agent, this can effectively block assessment of the agent's malicious intent. This blocking alone would be compatible with a single-process model of moral judgment that proceeds in two steps: first the assessment of causal responsibility, then the assessment of a culpable mental state. But such a single-process model would predict a similar blocking effect when no harmful consequence occurs at all, leaving the first step uncompleted. In contrast to this prediction, subjects readily assigned punishment to attempted harm-doers when no harmful consequence occurred. Thus, in the absence of any harmful consequence to trigger the assessment of causal responsibility, an assessment of the agent's culpable mental state dominates punishment judgments. This competitive interaction between causal and mental state analyses lends support to the proposed division between two processes of moral judgment.

7. General discussion

The present study yields two general findings. First, judgments of wrongness and permissibility are overwhelmingly determined by an analysis of culpable mental states, while judgments of deserved punishment and blame show relatively enhanced sensitivity to an analysis of causal responsibility. This finding contradicts the commonsense notion that acts are punished if and only if they are wrongful. Second, judgments of deserved punishment for failed attempts to harm are mitigated when the intended harm coincidentally occurs by some independent mechanism, relative to when the harm does not occur at all. This result contradicts the commonsense notion that moral judgments of an agent depend only on whether the agent causes a harm, and not on whether the harm happens to occur independently.

The finding that various types of moral judgment depend differently on mental state and causal/consequential information has important methodological implications. Recent studies of moral judgment have relied on dozens of different questions, including appropriateness (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), blame (Pizarro et al., 2003), moral responsibility (Woolfolk, Doris, & Darley, 2006), punishment (Oswald, Orth, Aeberhard, & Schneider, 2005), permissibility (Cushman et al., 2006), and victim anger (Erber, Szuchman, & IPrager, 2001) to name just a few. The choice of a particular type of moral judgment should be made with awareness of the unique patterns of dependency on consequence, cause, belief and desire elicited by that type of judgment. Likewise, comparisons between studies must be interpreted in light of the classes of moral judgment tapped by each.

The identification of different classes of moral judgment also reconciles the divergent results and theories of two research traditions: the cognitive developmental tradition, which proposes that adult judgments of wrongness are dominated by intentional factors, and attribution theory, which proposes that judgments of punishment and blame rely on a conjunction of intention and causal responsibility. The results of the present study suggest that these theories can be understood not as competing accounts of a single phenomenon, but as complimentary accounts of distinct phenomena.

The data presented here also underscore the usefulness of dividing "intention" into belief and desire components. The results of all three experiments indicate that judgments of wrongness and permissibility are more heavily determined by belief than desire. For instance, subjects rated it relatively more wrong to weld together pieces of metal when believing that this would burn the person holding the metal even if there was no desire to burn this person, only a desire to weld the metal. Subjects rated it relatively less wrong to weld together pieces of metal desiring that this would burn the person holding the metal so long as there was no belief whatsoever that it actually would burn the person holding the metal. Of course, subjects may have inferred desires from the statement of beliefs, and visa-versa, blurring the sharp distinctions drawn in the stimuli themselves. Nevertheless, the data presented here imply that what matters most when making moral judgments is our belief that we will cause harm, rather than our desire to cause harm.

The precise roles of belief and desire in moral judgment represent an essential topic for further investigation. The conceptual distinction between belief and desire in moral judgment has been noted at least as early as Heider's theory of responsibility (Heider, 1958), but studies have typically conflated the two dimensions. A notable exception is Yuill & Perner (1988), which shows that young children predict how "cross" a victim will be with an agent earliest in cases in which foreseeability and desire are conflated, later in cases of desire alone, and finally in cases of foreseeability alone. Meanwhile, researchers have drawn important distinctions between the representations of beliefs and desires, articulating separate mechanistic, ontogenetic and evolutionary accounts of each (reviewed in Saxe, Carey, & Kanwisher, 2004; Tomasello, Carpenter, Call, Behne, & Moll, 2005). These distinctions are likely to play a critical role in the development of theories of moral judgment.

Finally, the present study provides evidence for a distinction between two processes of moral judgment, one which assesses causal responsibility for harm and another which assesses culpable mental states. A key feature of a two-process model is that assessments of causal responsibility and mental culpability act competitively to determine moral judgments. Just such a competitive interaction is demonstrated in Experiments 3 and 4 by the significant differences in judgments of deserved blame and punishment when somebody fails to cause a harm that does not occur, as compared to when somebody fails to cause a harm that does occur by some independent means. When an agent commits a failed attempt but the harm occurs by some independent means, causal responsibility is assigned to the independent means, and assessment of the culpable mental state of the agent is competitively blocked. This leads to reduced judgments of punishment and blame. But when no harm at all occurs, causal responsibility cannot be assigned. In these cases an assessment of mental culpability dominates, leading to increased judgments of punishment and blame.

The results of Experiment 4 lend further support to the proposed division between causal and mental state processes. Replicating the pattern of results from Experiments 1–3, subjects assigned less punishment to an attempted harm-doer if the intended harm coincidently occurred by some independent causal mechanism. In

Experiment 4, this shift towards less punishment occurred in a highly specific manner: subjects were twice as likely to assign *no punishment* when the harm occurred coincidentally compared to when no harm at all occurred, but among those subjects who assigned *any punishment*, there was no significant difference in the degree of punishment assigned. This all-or-nothing effect is consistent with the two process account proposed: subjects are either influenced by the occurrence of the harmful consequence to rely on a process of causal responsibility, assigning no punishment at all, or persist in evaluating the mental state of the attempted harm-doer, assigning equal punishment in both conditions. However, this specific effect was obtained using a single pair of scenarios and it's interpretation depends in part on a null result, so it remains an important topic for further investigation.

The proposed two-process account puts a new twist on decades-old research into the development of moral judgment in children. Piaget (1965/1932) and later Kohlberg (1969) characterized the young child's moral system as dominated by causal and consequential reasoning, and also as heavily focused on punishment. The results of the present study provide a new insight into the developmental linkage between consequence and punishment, showing that into in adulthood there remains a fundamental relationship between causal/consequential analysis and punishment. Traditionally, what happens next in the child's development has been described as a shift within a single psychological system for moral judgment from reliance on causes and consequences to reliance on mental states. The two process model proposes an alternative developmental characterization: the existing consequence-based psychological system for moral judgment is augmented by a new and distinct mental-state-based system. The new system maps on to the adult concepts of wrongness and permissibility. Meanwhile, the old system continues to play a critical role in contributing to judgments of blame and punishment, but in a manner additionally constrained by the output of the new, mental-state system. This constraint presumably operates at least in part because of the explicitly held theory that wrong acts should be punished, and punishable acts must be wrong.

What, precisely, does it mean to refer to two "processes" of moral judgment? This term is favored in the present study because of the linkage of distinct inputs (causal versus mental state information) with distinct analyses ("causal responsibility = bad" versus "culpable mental state = bad") and distinct outputs (blame and punishment versus wrongness and permissibility). Moreover, Experiments 3 and 4 demonstrate that that subjects can be induced to rely more heavily on one process or the other depending on the availability of the necessary inputs to each process, and Experiment 4 suggests that this manipulation has an "all-or-nothing" effect on their subsequent judgments of punishment.

In what way does the two-process account differ from standard single-process models? On any theory, separate cognitive processes must be engaged to assess causal and intentional information – the claim that causal reasoning is accomplished separately from mental state reasoning is obviously true, and alone it does not warrant a distinction between separate processes *of moral judgment*. Rather, the critical issue is whether causal and intentional factors are integrated prior to

the output of a single valenced response (as on a single-process model), or whether separate valenced responses are computed on the basis of each factor and then act competitively to determine judgments of wrongness, punishment, etc (as on a two-process model). The "blame blocking" phenomenon reported in Experiments 3 and 4 provides key evidence for the existence of competition between the two processes: the assignment of causal blame competitively blocks the assignment of mental culpability, while the silencing of the causal assessment allows the mental-state assessment to dominate.

There are several important categories of evidence that would provide further support to the two process theory: the demonstration that the processes have separate developmental patterns of emergence (as suggested by the classic studies of Piaget and Kohlberg), the demonstration that they can operate in parallel and generate cognitive conflict (as suggested by various philosophical and legal dilemmas, as well as recent neuroimaging research (Young et al., 2007)), and the demonstration of dissociable neural networks subserving each process. These are each critical areas for further investigation.

An additional topic for future research is the relationship between the model presented here and other multi-system models of moral judgment that suggest a division between affective and cognitive processes (e.g. Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001) or between automatic and controlled processes (Cushman et al., 2006; Haidt, 2001; Pizarro & Bloom, 2003). Greene (2008) has proposed that affect plays a critical role in generating retributive intuitions. Future studies should address whether this affective response is driven by the emotional salience of a harmful outcome, as opposed to a malicious intention.

7.1. Conclusions

What are the causal and intentional properties of actions that lead to moral judgment? This question has framed decades of research in moral psychology and remains at its heart (Alicke, 2000; Baron & Ritov, 2004; Cushman et al., 2006; Darley & Shultz, 1990; Hauser, 2006; Heider, 1958; Mikhail, 2000; Piaget, 1965/1932; Pizarro et al., 2003; Shaver, 1985; Weiner, 1995). The present study suggests that the answer is not so simple as has been presumed: different moral judgments make use of causal and intentional properties of actions in systematically different ways. This finding sheds light on inconsistencies among past traditions of psychological research and has important implications for the design of future investigations.

The present study also provides evidence consistent with the existence of two distinct processes of moral judgment: one triggered by harmful consequences that delivers a negative moral judgment of the causally responsible agent, and another triggered by analysis of the mental states underlying action that delivers a negative moral judgment of an agent who believes that his action will cause harm, and to a lesser extent, an agent who desires for his action to cause harm. Further characterizing these two systems will contribute not only to our understanding of the psychology of moral judgment, but also to our understanding of fundamental dilemmas of philosophy and the law.

Acknowledgements

Thanks to Susan Carey, Liane Young, Marc Hauser, Walter Sinnott-Armstrong, the editors and several anonymous reviewers for their insightful comments on this manuscript.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 103*, 37–49.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74–85.

Bratman, M. E. (1989). Intention and personal policies. In J. E. Tomberlin (Ed.), *Philosophical perspectives. Philosophy of mind and action theory* (Vol. 3). Blackwell.

Cushman, F. A., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*, 281–289.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Darley, J. M., Klosson, E. C., & Zanna, M. P. (1978). Intentions and their contexts in moral judgments of children and adults. *Child Development, 49*(1), 66–74.

Darley, J. M., & Shultz, T. R. (1990). Moral rules – their content and acquisition. *Annual Review of Psychology, 41*, 525–556.

Erber, J. T., Szuchman, L. T., & IPrager, I. G. (2001). Ain't Mishehavin': The effects of age and intentionality on judgments about misconduct. *Psychology and Aging, 16*(1), 85–95.

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*(9), 1589–1602.

Fincham, F. D., & Roberts, C. (1985). Intervening causation and the mitigation of responsibility for harm doing. *Journal of Experimental Social Psychology, 21*(2), 178–194.

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology, 20*, 113–120.

Forguson, L. (1989). *Common sense*. London: Routledge.

Greene, J. D. (2008). The secret joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology and biology*. New York: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Hall, J. (1947). *General principles of criminal law*. Indianapolis: Bobbs-Merrill Company.

Hart, H. L. A., & Honore, T. (1959). *Causation in the law*. Oxford: Clarendon Press.

Hauser, M. D. (2006). *Moral minds: How nature designed a universal sense right and wrong*. New York: Harper Collins.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language, 22*(1), 1–21.

Hebble, P. W. (1971). Development of elementary school childrens judgment of intent. *Child Development, 42*(4), 583–588.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Imamoglu, E. O. (1975). Children's awareness and usage of intention cues. *Child Development, 46*, 39–45.

Karniol, R. (1978). Childrens use of intention cues in evaluating behavior. *Psychological Bulletin, 85*(1), 76–85.

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

Kohlberg, L. (1969). Stage and sequence: The cognitive–developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151–235). New York: Academic Press.

McLaughlin, J. A. (1925). *Proximate cause. Harvard law review, 39*(2), 149–199.

Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 1A theory of justice'*. Unpublished Ph.D. Thesis. Ithaca: Cornell University.

Nagel, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

Nelson Le Gall, S. A. (1985). Motive outcome matching and outcome foreseability – effects on attribution of intentionality and moral judgments. *Developmental Psychology, 21*(2), 332–337.

Oswald, M. E., Orth, U., Aeberhard, M., & Schneider, E. (2005). Punitive reactions to completed crimes versus accidentally uncompleted crimes. *Journal of Applied Social Psychology, 35*(4), 718–731.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Piaget, J. (1965/1932). *The moral judgment of the child*. New York: Free Press.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review, 110*(1), 193–196, discussion pp. 197–198.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*, 653–660.

Robinson, P. H., & Darley, J. M. (1995). *Justice, liability and blame*. Boulder: Westview Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 1–27.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioral Science, 13*(3), 238–253.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development, 57*(1), 177–184.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675–735.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford Press.

Wellman, H. M., Cross, D., & Bartsch, K. (1986). Infant search and object permanence: A meta-analysis of the A-not-B error. *Monographs of the Society for Research in Child Development, 51*, 1–51, 62–67.

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

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*(2), 283–301.

Young, L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in childrens' judgments of actors responsibility and recipients emotional reaction. *Developmental Psychology, 24*(3), 358–365.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67*(5), 2478–2492.