Causation, Norm Violation and Culpable Control

Mark D. Alicke

Ohio University

David Rose

Carnegie Mellon University

Dori Bloom

Ohio University

Causation, Norm Violation and Culpable Control

Human brains do spectacular things: They solve complex logical puzzles, compose symphonic masterpieces, conceive technological marvels, and create enduring artworks, for starters. But before they can embark on these prodigies, human brains must achieve something that they share in common with all brains—they must evaluate and differentiate which creatures, objects, and conditions will facilitate their prospects and well-being, and which will do them harm. Evaluation is the most fundamental component of human judgment (Osgood, Suci, & Tannenbaum, 1959) and one of the most important cognitive capacities for survival.

Virtually all meaningful human actions are automatically evaluated (Bargh & Chartrand, 1999; Fazio, et al., 1988). These evaluative reactions intrude on the judgments and attributions that people make about their own and others' behavior. So, when people make focal judgments about the components of a human act, such as whether it caused a particular outcome, whether the outcome was foreseen or foreseeable, whether the action was intentional or involuntary, and whether incapacities or situational constraints excuse or mitigate it, they are influenced by their peripheral evaluative reactions to the actor, the actor's behavior, or the outcomes that ensue (Alicke, 1992, 2000, 2008). As a result, when people are asked to identify, for example, the primary cause of an event, they accord privileged status to actions that arouse positive or negative evaluations. In this way, causal attributions reflect a desire to praise or denigrate those whose actions we applaud or deride.

At least, that's our story. There are two prominent alternatives to this assumption about the primacy of evaluation, the first fairly implausible in light of the extant data, the

second much in favor. The less credible view conflates how action components *should* be evaluated (in accordance with the criteria of Anglo-American jurisprudence and rational prescriptions for justice and fairness), with how people actually evaluate others. These prescriptive models stipulate that blame and responsibility require causation, intention, foresight or foreseeability, and the absence of mitigating or extenuating circumstances (Piaget, 1932; Shaver, 1985; Weiner, 1995). The primary value of such models is to translate fundamental legal and philosophical tenets into normative models of blame. Such models do a reasonable job of predicting blame or responsibility ascriptions under ideal conditions (e.g., Fincham & Shultz, 1981; Karlovac & Darley, 1988; Shultz, Schleifer, & Altman, 1981). They falter, however, once the funk and muck of real life events are transported to the judgment task, primarily because, in our view, they fail to account for the contribution of evaluative reactions to the judgment process (Alicke, 1992, 2000; Alicke et al., 2009).

The second alternative pertains specifically to causal judgments rather than to blame and responsibility *per se*. According to this view, of the various causal influences that compete for recognition, observers will elevate the most unusual or abnormal condition to primary causal status. This view harkens back to Hart and Honore's classic treatment of Causation in the Law (1959), was resurrected and further developed in Kahneman and Miller's (1986) norm theory, and is the basis of much current thinking and research on counterfactual reasoning (Mandel, Hilton, & Catellani, 2005). Because an event can have numerous abnormal causes, further refinements are needed, and have been supplied. One view is that people grant privileged status to causal conditions that, if altered, would prevent a harmful or unfortunate outcome (Mandel & Lehman, 1996).

The second, related view, is that people favor causes that identify an intervention that would alter the event's outcome (Collingwood, 1940). Because people are more likely to imagine interventions that change negative events into positive ones than the reverse, this view is similar to the first in that it entails citing the cause that, if changed, would negate the harmful outcome.

Both of these views, therefore, assume that causal ascriptions are based on a species of counterfactual reasoning. This reasoning highlights interventions that would undo the outcome that occurred, especially when the outcome is harmful or undesirable. In most cases, the prepotent cause will be the one whose negation improves the present state of affairs. Hitchcock and Knobe (in press) have explicitly endorsed this intervention approach as an alternative to what we call the evaluation (and they call the blame) perspective. In their view, the causal candidate that deviates most from the normal state of affairs will be identified as the primary cause because it provides the most suitable target for intervention. We refer to their approach and its cognates as the norm violation view.

Before reporting the results of the studies that we conducted to distinguish between our evaluation perspective and Hitchcock and Knobe's norm violation view, we want to clarify the basis of our disagreement with Hitchcock and Knobe's position to avoid exaggerating the differences in our views, and to elaborate the specific assumptions that underlie our position.

Clarifying the Basis of the Debate

Hitchcock and Knobe's analysis of how laypeople ascribe causation emphasizes the importance of identifying abnormal conditions that, if altered, would restore an event

to its more normal state. Discerning such intervention points highlights ways to improve one's own and others' prospects. Hitchcock and Knobe identify three types of norm violations that serve this purpose. First, abnormality in a statistical sense can be informative. Changing the behavior of people who do unusual things effectively restores an event to its normal state, and this capacity of statistically abnormal actions enhances their perceived causal potency. Second, Hitchcock and Knobe clearly recognize that moral or ethical transgressions provide a basis for heightened causal ascriptions. Finally, norm violations also include deviations from proper functioning. A malfunctioning machine, for example, would hinder a company's operations. Hitchcock and Knobe assume that the distinctions among these different types of norm violations are relatively unimportant, and that what ultimately matters is whether altering a particular causal candidate (e.g., fixing the machine) restores an event to its normal state and makes a bad situation better.

Because Hitchcock and Knobe clearly recognize that moral or evaluative judgments matter, the main point of contention concerns why they matter. We agree with Hitchcock and Knobe that norm violations are almost certainly the primary determinants of causal citations for events that do not involve human agents. Similarly, we concur that norm violations rule the causal roost for benign events that lack nefarious motives, undesirable or reckless actions, or harmful outcomes. We grant, therefore, that norm violations suffice to explain heightened causal efficacy for some types of events.

But the areas in which we disagree with Knobe and Hitchcock are significant in that they involve the events to which most of their examples apply, namely, those involving undesirable or harmful behavior. The crux of the disagreement concerns the

fundamental motivation that drives the ordinary person's construal of social events. Identifying ways to improve things that go wrong has obvious instrumental value and may consciously guide much of people's behavioral analyses. Nevertheless, primitive motives for revenge and retribution have weighed heavily in human affairs at least since people began recording them and can impede rational decision strategies. We assume that blame represents a symbolic form of retribution. Blame expresses disapprobation for the actor's motives or actions and, to borrow Joel Feinberg's, terminology, "stains" the actor's character (Feinberg, 1970). Because people generally prefer to view themselves as fair and rational, they must support these blame attributions with evidence. Exaggerating an actor's causal role in an event is one way in which this can be achieved. We elaborate this assumption below in the context of outlining the culpable control model of blame.

The Culpable Control Model

The culpable control model assumes that people generally try to follow cultural prescriptions for ascribing blame. In short, they seek to ascertain whether a person negligently or intentionally caused, or could have caused, harm to another's person or property, and if so, whether situational pressures (e.g., coercion, provocation) or personal incapacities (e.g., ignorance, mental illness) were sufficient to excuse or mitigate blame. These considerations comprise three linkages (See Figure 1) that represent distinct ways of exerting control during an action sequence: a link from mind to behavior (did the behavior occur on purpose?), from behavior to consequence (how strong was the causal connection between the actor's behavior and the outcomes that occurred?), and from mind to consequence (did the consequences come about as foreseen)?

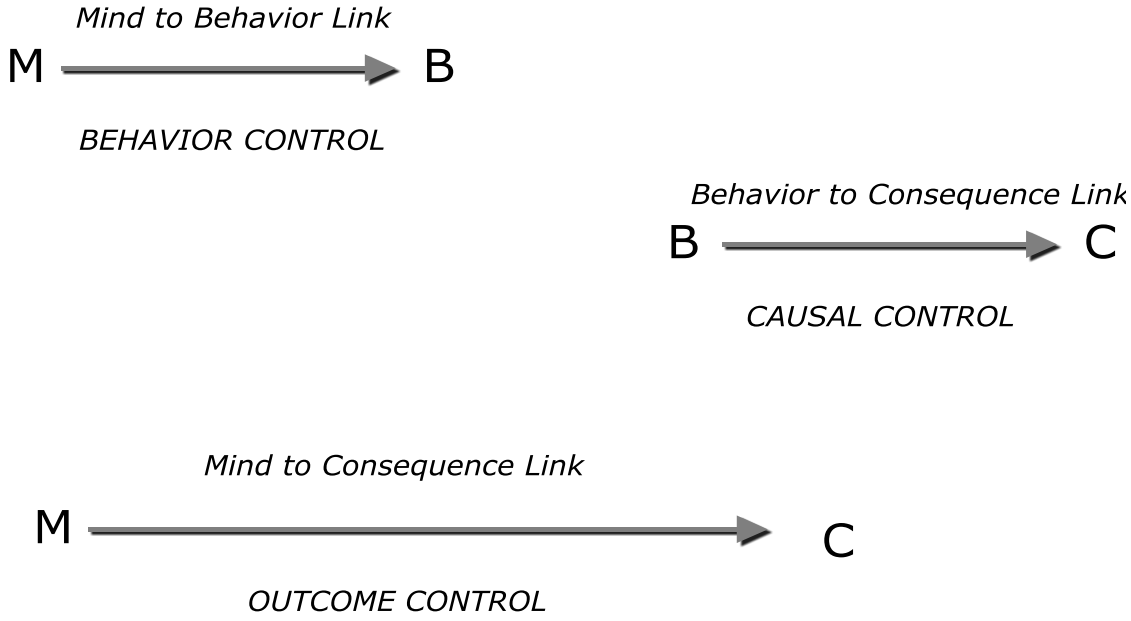

Figure 1: Structural Linkages among mental, behavioral and consequence elements

In addition to assessing actual control, observers also estimate potential control which involves judging whether the consequences that occurred *should have* been foreseen. Potential control is important in assessing negligent behavior in which harmful consequences are effected unintentionally but irresponsibly. Taken together, these linkages comprise assessments of behavior control (acting purposively), causal control (causing one or more harmful outcomes), and outcome control (causing the outcome in the desired and foreseen manner), and represent the degree of actual or potential control an individual exerted (or could have exerted) over an event.

At the same time that people consciously assess these aspects of control, they also spontaneously evaluate the actor, his or her actions, and the outcomes that occurred. We assume that spontaneous evaluations occur in response to the central elements of control

(behavior, causal and outcome), as well as to peripheral features of the event such as the actor's or victim's race or character, or the degree of harm that occurred. For example, an observer might react unfavorably to the knowledge that an actor spent a long time planning a despicable act (behavior control), or to the fact that the act was committed by someone who belongs to a disliked ethnic or racial group. Some evaluations are virtually endemic to control estimations, especially when assessing potential control. It is difficult, for example, to isolate negative reactions to what someone did or caused from determinations of what they should have done or known, especially under highly ambiguous circumstances. When spontaneous evaluations are sufficiently strong, the culpable control model assumes that the control elements (behavior, causal and outcome) that observers analyze are processed in a "blame validation" mode. Blame validation entails exaggerating a person's actual or potential control over an event to justify the desired blame judgment, or altering the threshold for how much control is required for blame.

The phrase "culpable control" reflects the fact that the desire to blame or find someone culpable intrudes on assessments of mental, behavior and outcome control. In a sense, culpability, which is supposed to be the output of the judgment, becomes part of the process of assessing the blame criteria. Much of the current debate in the literature on blame and causation is couched in terms of two simple models: blame attributions determine causal attributions (blame → cause), or the reverse (cause →blame)[1]. The culpable control model is usually characterized as endorsing the former relationship.

[1] What we call negative evaluations or spontaneous reactions are often referred to as “intuitions” or “emotions” by others (Damasio, 1994; Green, Sommerville, Nystrom, Darley & Cohen, 2001; Green & Haidt, 2002; Haidt, 2001; Heubner, Dwyer & Hauser, 2009; Pizarro & Bloom, 2003) .

However, a more complete characterization of the culpable control model would be: negative evaluative reaction→initial blame hypothesis→ blame validation processing →enhanced causal control→blame. In other words, negative evaluations or spontaneous reactions lead to the hypothesis that the source of the evaluations is blameworthy, and to an active desire to blame that source. This desire, in turn, leads observers to interpret the available evidence in a way that supports their blame hypothesis. In the present discussion, the primary avenue for supporting or validating a blame hypothesis is to increase perceptions of causal control, but more generally, it can also entail enhancing perceptions of behavior and outcome control.

The ultimate effect of perceived control and negative evaluations on blame is a compensatory one. We assume that some level of behavior, causal and outcome control is required to blame an actor for an actual or attempted offense. Even extremely prejudiced observers are unlikely to blame someone whose behavior was completely accidental, or who was causally unconnected to the harmful consequences. However, given a requisite baseline level of perceived control, strong negative evaluations increase blame ascriptions and alter judgments of the control elements that should, ideally, be assessed independently in ascribing blame. On the other hand, when the control evidence is overwhelming, there is scant opportunity for negative evaluations to skew the blame process.

Study 1: Doctors Violating Norms--Fortuitously

Our first study was based on one that Hitchcock and Knobe conducted to illustrate their norm violation position. In this scenario, an intern wants to administer a new drug to a patient with kidney problems but must obtain the signature of the pharmacist and the

attending physician. The pharmacist signs off but the physician realizes that the hospital has banned the drug due to its dangerous side effects. Nevertheless, the physician consents and the patient recovers with no adverse reactions. When asked to rate the physician's and pharmacist's causal roles in the patient's recovery, participants gave higher ratings to the physician. Since the outcome was favorable, Hitchcock and Knobe argue that the culpable control model cannot account for the findings because people do not blame others for favorable outcomes. In their view, heightened causal attributions occur because the physician's behavior is counternormative. As they state it: "our own account makes no mention of any sort of moral judgment regarding the effect. Instead, it posits a role for judgments about whether the candidate cause was itself a norm violation." Specifically, the physician's deviation from the normal state of affairs, regardless of the outcome of the event, leads people to view it as a suitable target for intervention, which in turn leads them to select him as the primary cause.

We believe that the culpable control model provides a more compelling account of causal attributions in the physician scenario. Hitchcock and Knobe would agree that the physician's behavior provides a basis for negative evaluations, although they would emphasize the counternormativeness of the physician's behavior. The culpable control model, however, assumes that this negative evaluation encourages observers to believe that the physician is blameworthy and that they seek to validate their desire to blame him. The nature of the outcome, rather than being irrelevant, as Hitchcock and Knobe maintain, provides a basis for either justifying the desire to blame the physician or for attenuating that desire. A negative outcome, such as the patient's death, would fuel the desire to blame the victim, which would be reflected in heightened causal attributions to

the physician. On the other hand, a positive outcome, such as occurred in Hitchcock and Knobe's scenario, would attenuate the physician's perceived causal role. These assumptions could not be tested in Hitchcock and Knobe's scenario because they included only a positive outcome condition and simply compared the physician's perceived causal influence to that of the pharmacist, whose causal role was minimal.

A more complete analysis of causal attributions in Hitchcock and Knobe's scenario requires conditions that vary both the normativeness of the physician's behavior and the nature of the outcome. To this end, we expanded Hitchcock and Knobe's scenario to include conditions in which the physician's behavior was normative (i.e., he followed the hospital's policy and refused to administer the drug) or counternormative (i.e., he administered the drug, as in Hitchcock and Knobe's scenario), and whether the patient experienced a positive outcome (i.e., he recovered with no side effects, as in Hitchcock and Knobe), a negative outcome (i.e., death), or no outcome information was provided. We assessed ratings of the physician's causal impact on the patient's outcome, as well as positive versus negative evaluations of the physician's decision to administer or to refrain from administering the drug.

In this context, the main difference between the culpable control and norm violation views concerns the role of outcome information in causal attributions. Hitchcock and Knobe stipulate that whereas the culpable control model is based on the goodness or badness of the event's outcomes (since people can only be blamed for bad outcomes), their own norm-violating position applies to behaviors rather than outcomes. As the previous discussion of the culpable control perspective makes clear, this does not quite accurately characterize the model. Positive and especially negative evaluative

reactions can occur in relation to the actor's intentions, motives, actions, and outcomes, as well as to a host of other features such as his or her race, gender, or personality. Nevertheless, what is most germane for present purposes is that the valence of the outcomes *do* matter in the culpable control model. Because Hitchcock and Knobe claim that only behavioral norm violations count in causal ascriptions, they would predict no effects due to the event's outcomes.

The culpable control model, however, makes specific predictions regarding the interplay between the normality of the physician's behavior and the outcomes that it produces. When the physician evokes negative evaluations by violating the hospital's policy, the death of the patient is a severe aggravating circumstance, which raises his perceived causal influence beyond where it would reside if he followed the hospital's policy and produced the same outcome. By contrast, the physician who behaves appropriately by following the hospital's policy is less likely to be penalized for the patient's death. Specifically, there should be a statistical interaction such that the physician who violates hospital policy is seen as more causal when the patient dies than when the patient lives, whereas no such difference should occur for the physician who follows hospital policy.

Results

The findings of Study 1 are illustrated in Figures 2 and 3[2]; the first figure depicts ratings of blame-praise whereas the second shows causation ratings.

---

[2] A total of 319 participants (Male=121, Female=193, Did Not Indicate=5) were selected from an introductory psychology course.

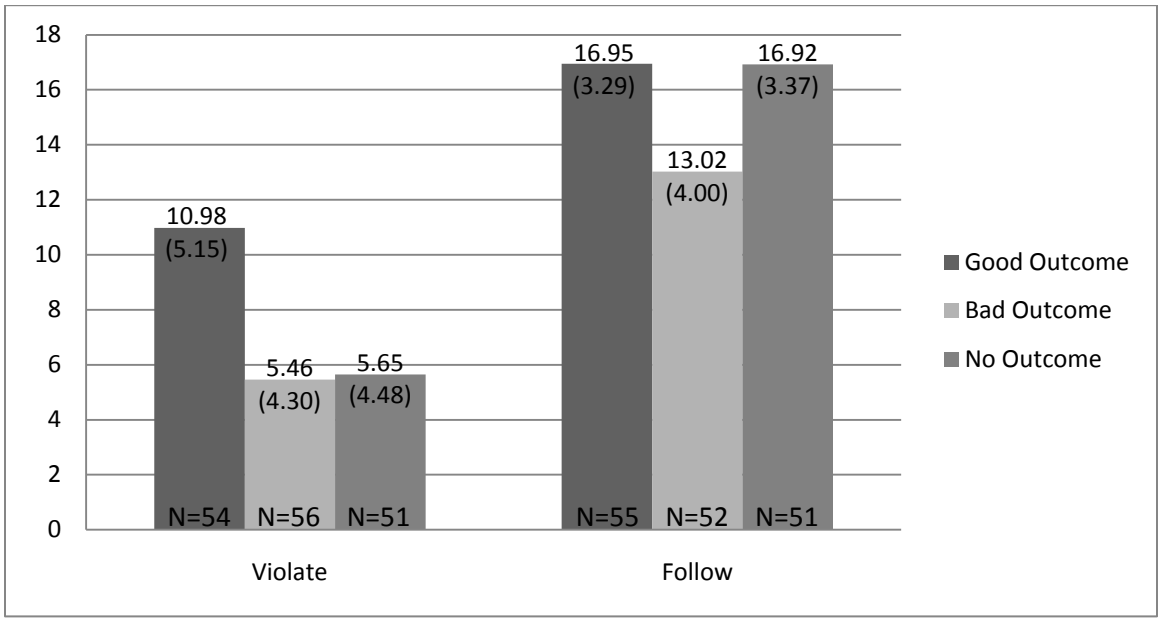

Figure 2: Blame-Praise Ratings[3]

Significant interactions between the drug decision (i.e., to violate the norm and administer the drug or to adhere to the norm and refuse the drug) and the outcome of the decision (i.e., recovery, death, or no outcome) on evaluations (i.e., blame-praise) of the physician's decision, $F(2, 313) = 11.28$, $p < .0001$, and on ratings of his causal involvement, $F(1, 213) = 13.05$, $p < .001$, supported culpable control predictions. These findings show that when the physician violated hospital policy, he was viewed more negatively when the patient died than when he lived, $F(1,313) = 48.59$, $p < .0001$, and was viewed as more causal in the former case than in the latter, $F(1,213) = 9.59$, $p < .01$. When the physician followed hospital policy, evaluations of his decision were relatively favorable regardless of the outcome, although they were significantly reduced by the fact

[3] The question concerning evaluation was a follows: How would you evaluate the attending doctor's decision to administer/not administer the drug? A 21 point scale was used with 1 anchored at "extremely blameworthy" and 21 anchored at "extremely praiseworthy". In this and the following studies, mean values are placed at the top of the graph, standard deviations inside the bar at the top, and sample sizes at the bottom of the bars.

of the patient's death, $F(1, 313) = 23.92$, $p < .0001$.

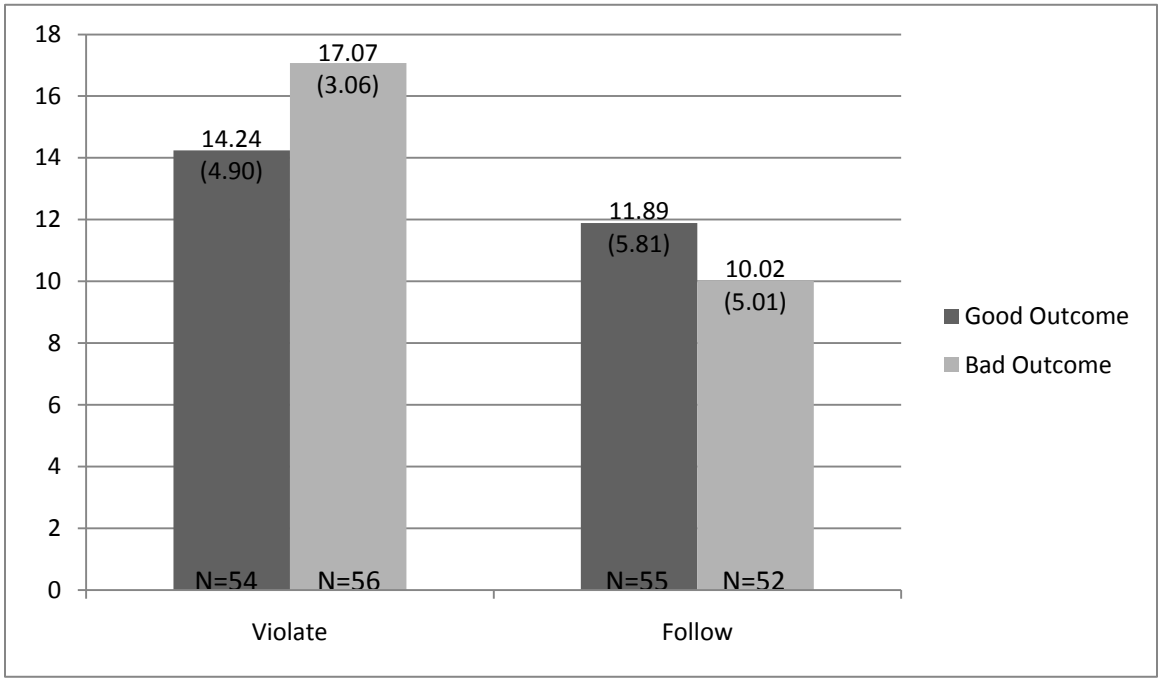

Figure 3: Causal Ratings[4]

The pattern of results displayed in Figure 2 shows the importance of positive and negative evaluations of the physician's behavior and the outcome that the patient experiences. First, the physician who violated the hospital's policy but had the fortuitous outcome of the patient recovering (as in Hitchcock and Knobe's scenario) was viewed far more negatively (i.e., as relatively more blameworthy) than the physician who respected the hospital's policy and obtained the same outcome, $F(1,313) = 56.28$, $p < .0001$. In fact, the physician who violated hospital policy and obtained a positive outcome was viewed even more negatively than the physician who followed hospital policy but obtained a negative outcome (i.e., the patient's death), $F(1,104) = 6.38$, $p = .012$.

[4] The question concerning the doctor's causal role was as follows: How much did the attending doctor's actions and decisions cause the patients recovery/death? A 21 point scale was used with 1 anchored at "not at all the cause" and 21 anchored at "very much the cause".

Clearly, violating the hospital's policy is by itself a potent source of blameworthiness in this context.

Comparison between the no-outcome conditions were also revealing. As we assumed, the physician who violated hospital policy was viewed very negatively, whereas the physician who followed hospital policy was viewed very positively, $F(1, 313) = 188.24$, $p < .0001$. However, the patient's recovery had no corresponding effect when the physician was already evaluated positively in that favorable evaluations were about equal in the recovery and no outcome conditions.

In general, therefore, the findings of this study are consistent with the culpable control model and very difficult to explain from a norm violation perspective. First, the data support the culpable control model's assumption that there are two important sources of evaluation in this scenario: the physician's decision to go along with or to violate the hospital's policy, and the patient's death or survival. The results show that the physician in Hitchcock and Knobe's original scenario would have been viewed very negatively if not for the fortuitous outcome of the patient's recovery, and he was not viewed very positively even with this happy consequence. These findings are consistent with the culpable control assumption that the physician's decision to contravene hospital policy provided a strong initial basis for blame. When the patient died as a result, this negative outcome was a severe aggravating circumstance which further elevated the physician's perceived causal role. However, when the physician was viewed positively for following the hospital's policy, he was not viewed as any more causal as a result of the patient's death in comparison to when the patient survived. Since the norm violation view

explicitly disavows the influence of outcomes on causal judgment, it cannot explain this interaction pattern.

Study 2: A Direct Test of Norms Vs. Evaluation

The most direct way to adjudicate between the norm violation and culpable control views is to implement a design that varies the goodness or badness of an actor's behavior simultaneously with whether it violates or adheres to a norm. We therefore created a story in which a group of students who lived on the same floor of a dormitory obtained a copy of the final exam for their biology class. The students either cheated or didn't cheat on the test. One student, John Granger, went along with the group (norm condition) or did not go along with the group (counternorm condition). This design, therefore, included four conditions: a) Granger follows the norm and cheats on the test (norm, bad); b) Granger follows the norm and does not cheat on the test (norm, good); c) Granger deviates from the norm and cheats on the test (counternorm, bad); d) Granger deviates from the norm and refuses to cheat on the test (counternorm, good).

The biology class comprises 80 students and is graded on a curve such that 20 people will receive a grade of A, 20 a grade of B, 20 a grade of C, and 20 students will receive a D. Granger's score was the 20th highest score in the class, which means he was the last student to receive a grade of A. The 21st student was a pre-med student who received a B, and as a result, missed the GPA cutoff she needed to get into the medical school she was hoping for by .07 GPA points. Participants were asked to indicate the extent to which they thought Granger was the cause of the student failing to meet the medical school cutoff, the degree to which he was to blame, and also to rate the goodness or badness of his actions.

Results

Our most fundamental prediction was that judgments of causation and blame would be based more on whether Granger's behavior was good or bad than whether it was normative or counternormative. The results confirmed this prediction: Granger was seen as less causal when his behavior was good (M = 3.37) than when it was bad (M = 5.20), $F(1, 178) = 17.12$, $p < .0001$, and he was also blamed less when his behavior was good (M = 2.92) than when it was bad (M = 5.19), $F(1, 178) = 27.98$, $p < .0001$. Overall, there was no main effect of whether his behavior was normative or counternormative on causal ratings, $F(1, 178) = 2.10$, $p < .15$, or on blame, $F(1, 178) = 1.59$, $p < .21$.

However, these findings were qualified by an interaction that revealed the same pattern on causal, $F(1, 178) = 4.24$, $p < .05$, blame, $F(1, 178) = 5.58$, $p < .02$ and evaluative, $F(1, 178) = 38.96$, $p < .001$ judgments (see Figures 4,5 and 6).

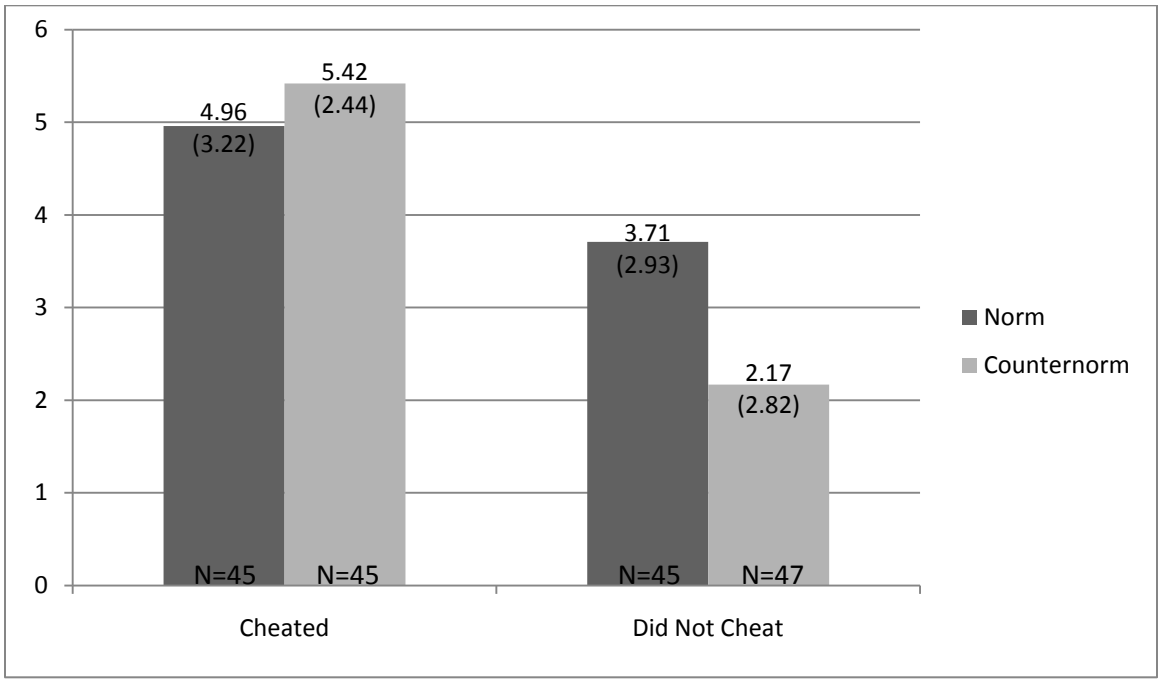

Figure 4: Blame Ratings[5]

When Granger cheated on the test, his causal impact and blameworthiness were uninfluenced by whether he was the only one who cheated or one of a group of cheaters. However, when he behaved admirably by refusing to cheat, he was seen as less causal and less blameworthy when he was the only one who took the moral high ground than when everyone else also refrained from cheating.

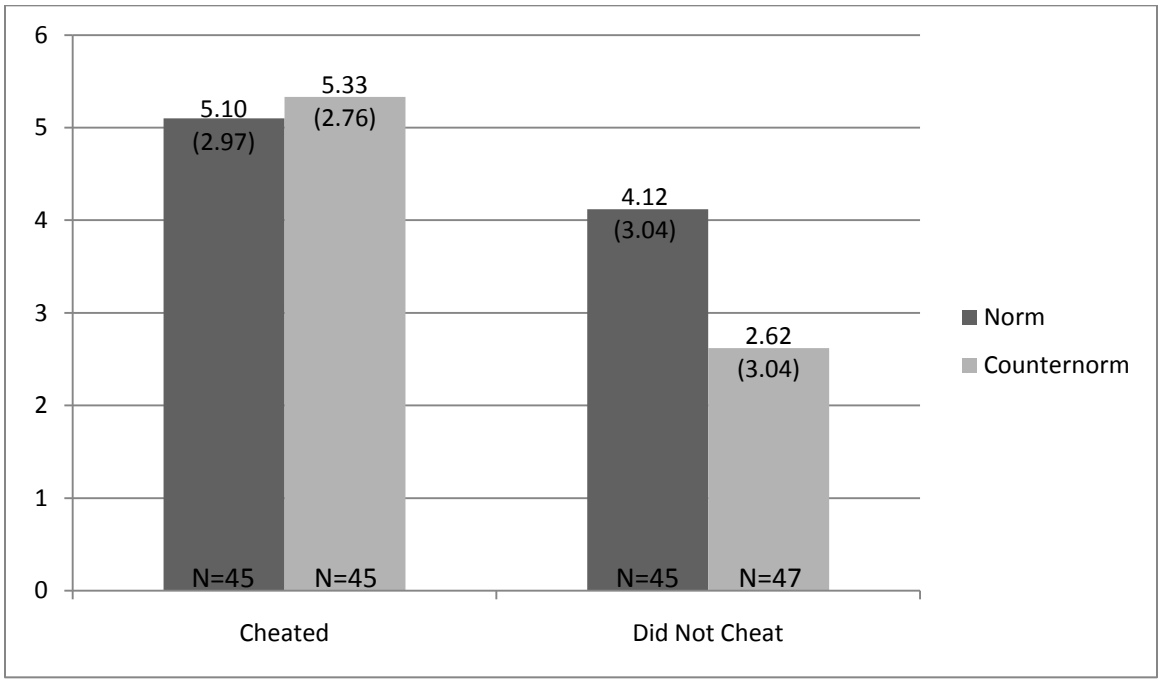

Figure 5: Causal Ratings[6]

These data show clearly that causal judgments for behavioral outcomes are not determined by norm violations alone. Whether a norm violation influences causal judgment depends on the way it is evaluated. Clearly, people can violate norms by doing

[5] Ratings were made on a 0-9 scale labeled anchored by "not at all to blame" and "very much to blame."
[6] Ratings were made on a 0-9 scale anchored by "not at all the cause" and "very much the cause."

good things or bad things. When an actor behaves badly, in this instance by cheating, his perceived causal influence and blameworthiness are maximized regardless of what everyone else did. Essentially, participants apply a deontological principle which states "Don't Cheat" regardless of what others are doing. However, things are a bit more nuanced for positive behaviors. People who go against the crowd to do the right thing are rewarded by being assigned *less* causal impact and decreased blameworthiness. Specifically, the actor who refused to cheat when everyone else on his floor cheated was seen as less causal and less blameworthy for the prospective medical student's misfortune. In contrast to the norm violation view, which predicts that actions that violate norms will be seen as *more* causal, we have shown that a person who behaves admirably by violating the norm is seen as less causal and less blameworthy as a result.

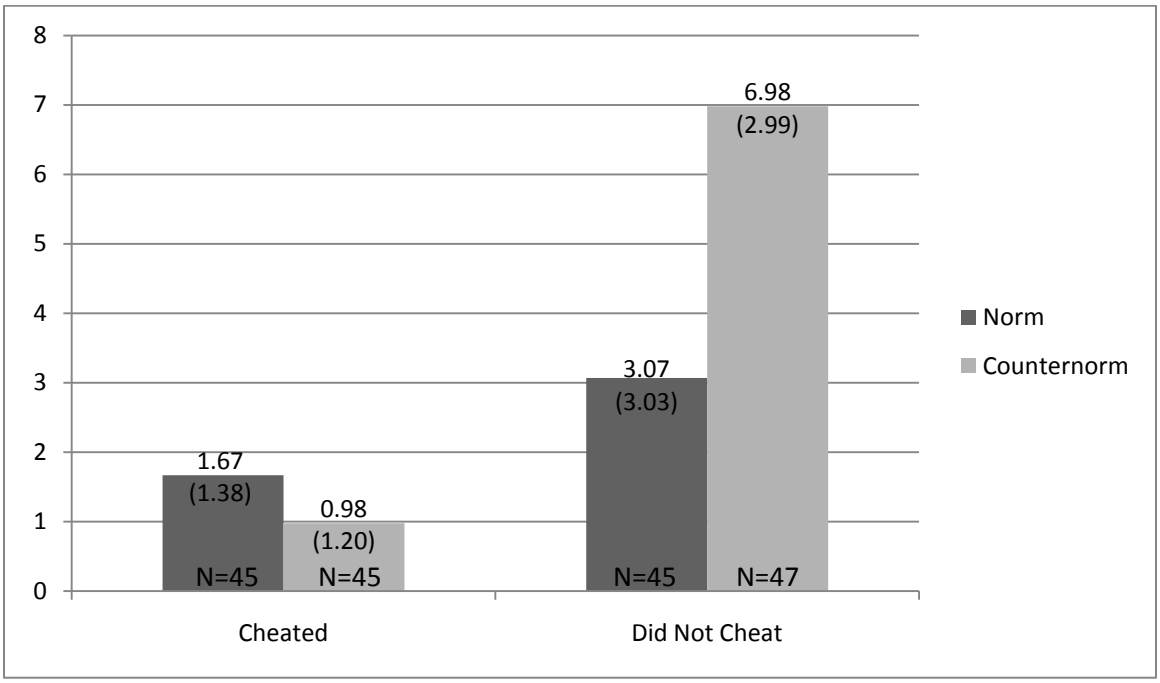

Figure 6: Evaluative Ratings[7]

It could be argued that there are conflicting norms in the situation that we created. That is, in addition to the local norms in which a group of individuals chose to cheat or not to cheat, there are general social norms that proscribe cheating. One might say that the student who violated local norms by not cheating nevertheless honored general proscriptions against cheating. Indeed, conflicting norms are the rule rather than the exception in evaluating undesirable human actions. In Hitchcock and Knobe's physician scenario, for example, the physician who violated hospital policy by administering a drug that he thought would help the patient could be said to have followed a general norm of "do what's best to help others."

Our point, which is strongly supported by the data we have presented, is that the effects of either local or general norm violations on causal judgments depend upon the evaluative tone (goodness or badness) of the behaviors they entail. The reason the student who violated local norms of cheating was seen as less causal was not that he adhered to a general norm *per se*, but that he did something good and praiseworthy, namely, exhibited integrity and independence. However, for the student who cheated either solely or as part of a group, local and general norm violations made no difference to causal attributions since both violations were equally bad and blameworthy.

Study 3: Chicken or Egg: Blame → Cause, or Cause → Blame?

In her contrast of the culpable control model versus Knobe's (2006) view, Driver (2008) suggests that the culpable control model entails that blame judgments precede and

[7] Ratings were made on a 0-9 scale anchored by "very bad" and "very good."

determine causal ascriptions (i.e., blame → cause), whereas Knobe's model stipulates the opposite relationship (i.e., cause→ blame). As we discussed earlier, the culpable control assumption is a bit more complex than this (specifically, we have suggested: negative evaluative reaction→initial blame hypothesis→ blame validation processing →enhanced causal control→blame), but we agree with Driver that demonstrating the viability of the blame → cause model would provide strong support for the culpable control position. For this reason, we designed one more study to test the blame → cause argument.

We created a new scenario in which a homeowner shot and killed an intruder who, unbeknownst to the homeowner, turned out to be an innocent and sympathetic victim or a dangerous criminal. The culpable control model predicts that, with all other things being equal, an actor who harms a likable victim will be seen as more causal than one who harms a dislikable victim. This prediction is based on the assumption that people will react more negatively to someone who harms a likable victim, and will therefore augment his perceived causal role to express their disapprobation. Because the norm violation view explicitly states that outcomes of events don't matter in causal assignment, it would predict no difference between these conditions since the antecedent event--that of the homeowner shooting a presumed intruder--is identical in both cases.

Specifically, the story (see Appendix A) was one in which the victim, Edward Poole, was either a dangerous ex-convict who broke into a home (negative characterization), or a physician who entered a home at the neighbor's request to feed her cat while she was away (positive characterization). In both cases, Poole was shot by one of the homeowners, Turnbull (who came home unexpectedly and didn't know of his wife's arrangement), who confronted Poole as he was climbing the stairs inside the house.

The most basic prediction is that Turnbull will be blamed more, and seen as more the cause of the victim's death, when the victim is characterized positively as opposed to negatively. In addition, we ran a causal search on the data to test whether blame determines causal judgments or causal judgments determine blame.

We were also interested in establishing boundary conditions for the hypothesized victim characterization effect. We assume that effects based on evaluative reactions will be cancelled when an actor's causal role is unambiguous, that is, when an actor obviously is, or obviously is not, an important causal contributor to an event. To test this, we created a causal overdetermination condition in which the external circumstances negated the actor's causal influence. Because people do not generally stray too far from the objective evidence, we assume that they will refrain from ascribing heightened causality when the data unequivocally fail to support such a judgment. However, we also contend that this restraint is tenuous. Given even a small degree of ambiguity, evaluative reactions should again exhibit substantial effects on causal assessment.

Three different versions of the circumstances surrounding Poole's death were created. In the first, an autopsy revealed that Poole suffered a brain aneurysm virtually at the moment that he was shot by Turnbull. Under these circumstances, even those who have strong reactions to an innocent victim's death will be reluctant to ascribe more causal influence to Turnbull because his behavior was unnecessary to produce Poole's death in the immediate situation. Accordingly, we expected to obtain uniformly low causal ratings in these conditions regardless of whether Poole was characterized negatively or positively.

Once these constraints on causal interpretation are loosened, however, effects of Poole's negative versus positive characterization should be observed. In the second version of the story, participants were told that the autopsy indicated that Poole was seriously ill and would have died from a brain tumor within a few weeks. In contrast to the previous condition, we assumed that the constraint in this condition--that Poole would have died in a few weeks from a brain tumor--would introduce sufficient causal ambiguity to restore positive and negative evaluation effects. Thus, higher causal ratings were predicted in the condition in which Poole was characterized positively than in which he was characterized negatively. In fact, we predicted that these effects would be approximately equal to those in the control condition in which no further information was provided about Poole's medical condition. In sum, we wanted to show that only severe constraints on causal judgment (i.e., causal overdetermination), and not moderate ones (i.e., the victim would have died in the near future), mitigate effects of positive and negative characterizations on causal assignment. Findings such as these would suggest a pervasive influence of evaluation effects on causal judgment and would show that they are eradicated only when objective information about countervailing causal forces is exceptionally compelling.

In addition to asking participants to indicate the extent to which they thought that Turnbull was the cause of Poole's death, we also asked them to indicate the extent to which they thought that he was to blame. We expected to find the same pattern of effects on blame as on causation.

Results

As the culpable control model predicts, the homeowner was both blamed more,

$F(1, 254) = 55.22$, $p < .0001$, and seen as more the cause of the victim's death, $F(1, 254) = 13.53$, $p < .0001$, when the victim was characterized positively as opposed to negatively. The more specific contrasts were also consistent with our predictions (See Figures 7 and 8).

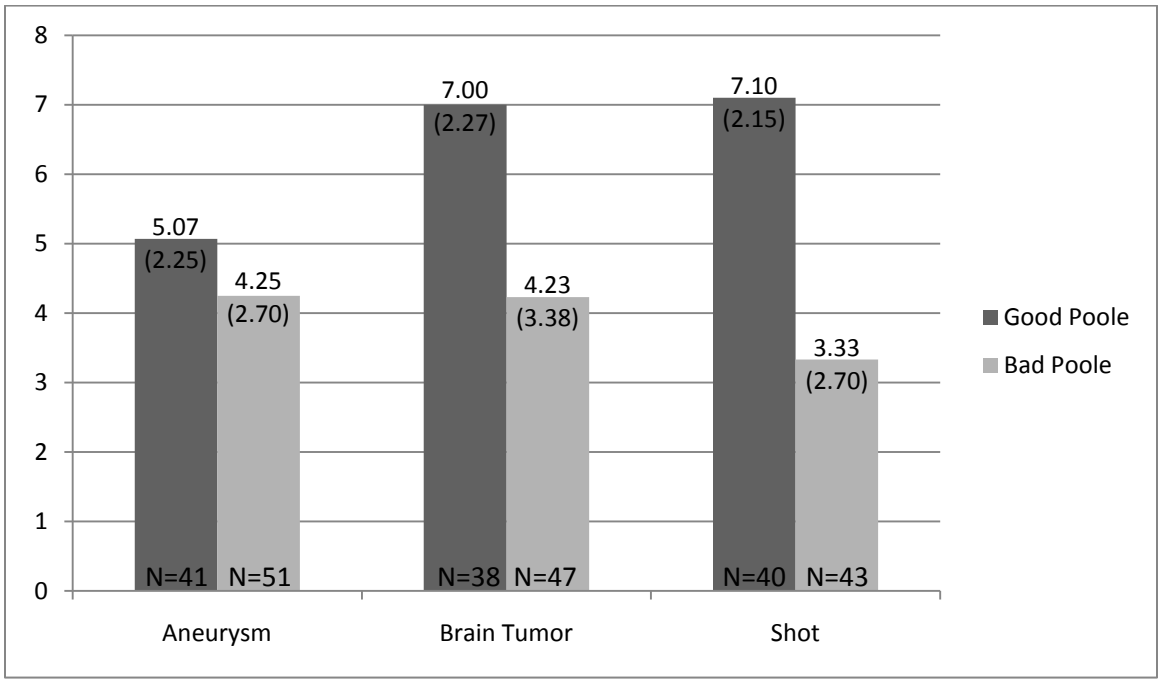

Figure 7: Blame Ratings[8]

The difference between the positive and negative victim characterization conditions was significant when the victim's death was delayed by two weeks, both on ratings of Turnbull's causal influence, $F(1,254) = 4.109$, $p < .05$, and on his blameworthiness, $F(1,254) = 4.804$, $p < .01$. These same findings were obtained on ratings of causation, $F(1,254) = 13.381$, $p < .0001$, and blameworthiness, $F(1,254) = 41.822$, $p < .0001$, when the victim died immediately after being shot by the actor without any further qualification. As we predicted, however, this same effect was not obtained when the

[8] Cause and blame ratings were made on the same scales as in the previous study.

victim would have died anyway in the immediate situation due to an aneurysm ($p > .05$).

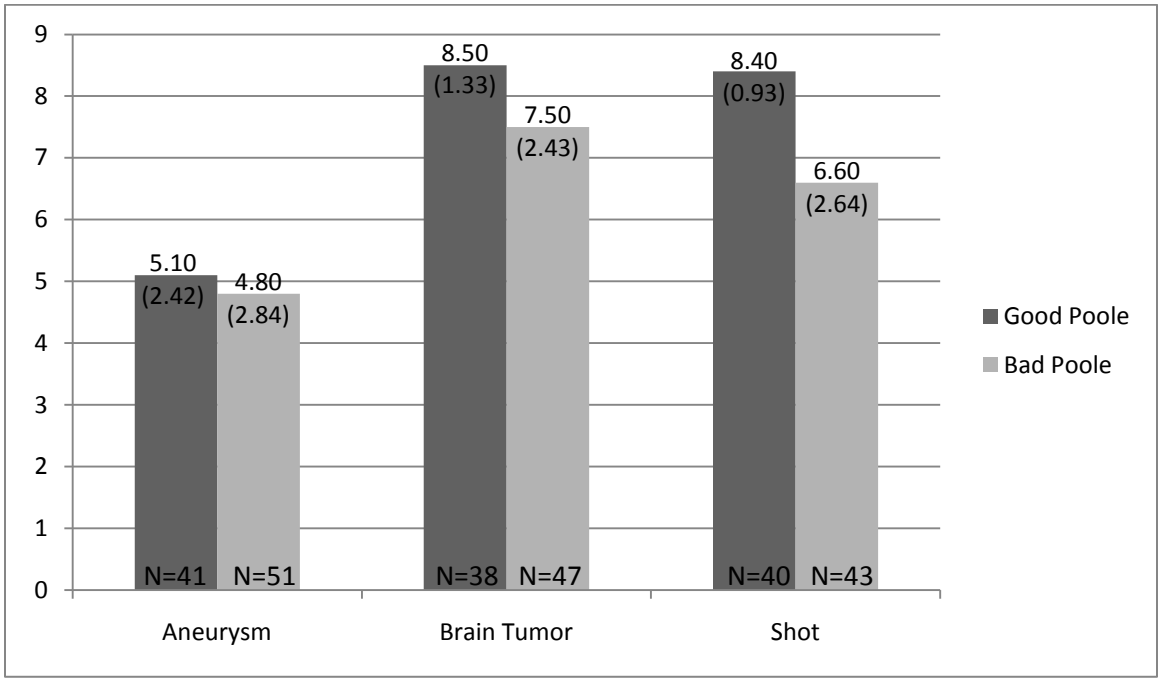

Figure 8: Causal Ratings

Finally, we used the Tetrad IV[9] software program to test whether cause→blame or blame→cause. Tetrad runs an automated "search" on the data set and detects the causal model that best explains the relationship among variables in the observed data. The model[10] that we obtained is shown in Figure 8:

[9] Tetrad IV is available at www.phil.cmu.edu/projects/tetrad

[10] The implied covariance matrix used to estimate the model is as follows:

	Blame	Cause	Outcome	Poole
Blame	8.7958			
Cause	2.9474	7.0799		
Outcome	0.0000	0.8218	0.6789	
Poole	0.6033	0.2022	0.0000	0.2491

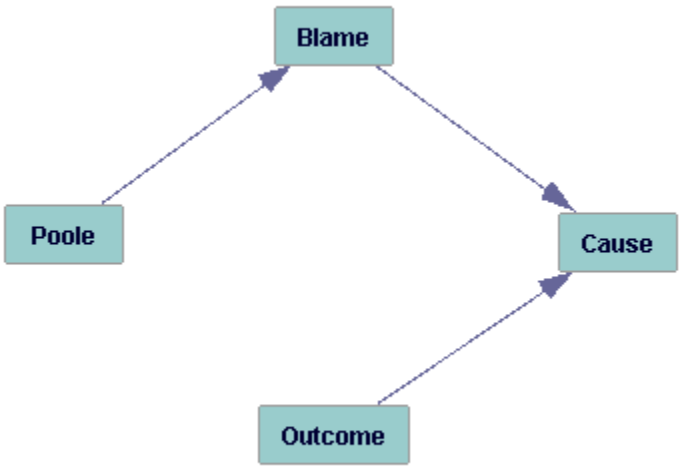

Figure 9: Blame→ Cause Model[11]

This model, which we will refer to as the Blame→Cause Model, shows that: a) Blame and Outcome (aneurysm vs. brain tumor vs. shot) independently influence causal judgments, and b) Poole's characterization (helpful neighbor vs. ex-convict) affects causal judgments only through the intermediary of blame. This model is a good fit of the data df=7, X^2=6.74, p=.46, and provides strong evidence for the Blame→Cause model in the Turnbull case.

---

[11] The equations for the variables in the above model are as follows:
Poole=N(1.4567, .2491)
Blame=2.4216 • *Poole* + N(1.5277, 7.3349)
Cause=.3351• *Blame* + 1.2015 • *Outcome* + N(2.6043, 5.0972)
Outcome = $N(1.9685, .6790)$

Culpable Control, Norm Violations and Causation: An Overview

We have argued that norm violations are insufficient to account for causal attributions for human events that involve undesirable behavior and/or harmdoing. The data presented in this paper cast serious doubt on whether any evidence has yet been adduced to show that the norm violation account is a better model of ordinary causal attributions than the culpable control model. What we have demonstrated is that blameworthiness has a pervasive influence on causal judgment and that it can account for the types of effects that Hitchcock and Knobe describe.

Whereas Hitchcock and Knobe view moral or evaluative considerations as one species of norm violation, we view norm violations as one aspect of evaluative judgment. One might say that we are barking up the same tree but from different angles. Nevertheless, the distinction is a vital one. The norm violation position depicts causal judgment as a largely rational process whereby people seek to identify actions that, if changed, would improve an undesirable or harmful state of affairs. Moral or evaluative considerations are merely one type of norm violation that can help to identify where interventions would remedy a bad situation. In other words, by emphasizing an actor's wrongdoing as a causal factor, observers disclose how to improve his or her future behavior.

We see three primary limitations to the norm violation position as Hitchcock and Knobe have stated it. First, like us, they advocate a functional view of causation as it applies to human events, especially harmful ones. However, in promoting interventions as the primary motive in causal assignment, they are likening people to engineers who appraise non-optimal situations with an eye toward improving them. Feinberg (1970), in

distinguishing among different criteria for assessing causation, has referred to this as the "engineering criterion." The engineering criterion is a perfectly apt metaphor for causal analysis when there is no strong basis for positive or negative evaluations, but as we have shown, norm violations alone cannot account for variations in causal ascriptions when a basis for positive or negative evaluation is present. Consider again the scenario in which the physician violated the hospital's norm by recommending a drug that was prohibited. We agree with Hitchcock and Knobe that if interventions could be identified to discourage doctors from violating hospital policies, patients would generally have better outcomes. But as the findings of our second study suggest, if observers had applauded rather than disapproved of the physician's actions, they would have seen him as less, rather than as more, causal for his intervention. In our view, the "person as engineer" metaphor applies best to relatively mundane human actions rather than to the events that fascinate philosophers, lawyers, psychologists, and other enquiring minds. For more interesting cases of ordinary causal judgment, we would replace the engineering criterion with what Feinberg calls the "stain criterion," thereby transforming the "person as engineer" into the "person as evaluator."

The second limitation of the norm violation approach as an account of causal judgment in human affairs is the claim that such violations refer only to behaviors and not to other elements of harmful events such as outcomes. This seems like an arbitrary stipulation: Most psychological theories of counterfactual reasoning acknowledge that norms apply both to actions and outcomes (Mandel, Hilton, & Catellani, 1985). Thus, in thinking about how harmful events could have turned out differently, observers consider

different paths that could have been taken and different consequences that could have been achieved.

While this weakness can be rectified simply by extending the norm violation view to encompass outcomes, our third issue with a pure norm violation approach is that it depicts causal judgment as a highly rational, controlled process. Affective responses have no special role, and evaluations are important only as applied to moral transgressions, which constitute one type of norm violation. Presumably, observers scrutinize an actor's selfish, greedy, or malevolent actions primarily because they contain the blueprints to remedy a harmful event. By contrast, the culpable control model assumes that evaluative reactions accompany virtually all human events in which good or bad actions or outcomes occur. These evaluative reactions can, and do, influence causal judgment (Alicke, 1992; Alicke, 2000; Alicke et al., 2009; Alicke & Zell, 2009; Mazzocco & Alicke, 2004). Some evaluative reactions are emotionally-charged, such as when an actor behaves despicably or produces horrendous outcomes, whereas others may simply entail goodness or badness assessments with little emotion. Although affectively-charged events are probably more susceptible to bias than more mundane ones, evaluative reactions generally provide the opportunity for pervasive biases in causal judgment.

For a norm-violation approach to have priority over an evaluation based model, it must provide at least some examples in which the influence of norm violations on causal judgment cannot be explained with reference to the praiseworthiness or blameworthiness of some element of the action sequence. Another of the examples that figures prominently in Hitchcock and Knobe's norm-violation position is a scenario in which

administrative assistants and faculty members routinely take pens from a receptionist's desk. The administrators are allowed to take the pens whereas the faculty members are not. As the story develops, an administrator and a faculty member each take a pen from the receptionist's desk, leaving her penless when an important message arrives. Knobe and Fraser (2008) asked participants to rate the extent to which the administrator or the faculty member was the cause of the receptionist's misfortune. Participants gave much higher causal ratings to the faculty member.

Hitchcock and Knobe interpret this as evidence that people accord primary causal status to actions that violate norms. Since moral transgressions are merely one type of norm violation, and different types of norm violations are effectively interchangeable, what matters most for them in the pen scenario is that the professor violated a norm, not that he did something blameworthy.

We disagree heartily with this interpretation. In our view, the primary reason for highlighting the faculty member's role in this unfortunate event must surely be that he is a depraved pen pilferer. Again, while we acknowledge that norm violations are important to causal assignment, norm violations that entail undesirable behavior are dignified with special causal status because they support blame attributions. Rather than demonstrating that norm violations identify interventions that improve events, Hitchcock and Knobe have shown that observers express their disapproval of an individual who violates explicitly-stated rules by saddling him with heightened causal responsibility.

We ran two small studies to assess our contention that perceptions of the professor's bad behavior underlie causal judgments in this scenario. Consistent with Knobe and Fraser's methodology, we had separate groups of participants (N = 265) rate

the extent to which they agreed or disagreed that the professor or the administrative assistant caused the secretary's problem (1-7 scale ranging from totally disagree to totally agree). We then asked participants to explain their causal ratings. A single coder who was blind to the purpose of the study coded the explanations.

The results for causal judgments of the professor and administrative assistant replicated Knobe and Fraser's findings in that the professor was seen as significantly more causal (M = 4.23, SD = 1.69) than the administrative assistant (M = 2.68, SD = 1.66), $t(263) = 7.52$, $p < .001$. Codings of the explanations revealed that participants cited the professor as the cause of the secretary's plight 66% of the time; the other 34% of the explanations were scattered among other causes including the administrative assistant, the receptionist, the other faculty members, and the prohibitory rule itself. Of those who cited the professor as the main cause, all but one indicated as their reason the blameworthiness of his behavior. Not a single participant mentioned that he violated a norm. Interestingly, virtually all other causal citations also entailed the blameworthiness of someone's actions. For example, those who cited the receptionist stated that she should have hidden her pens or stood up for herself to prevent the pen thievery. Those who saw the administrative assistant as the main cause explained this by saying that the administrative assistant should have known better and that although taking the pens was permitted, the assistant should have foreseen that this would lead to a shortage. The explanations that participants provided, therefore, support the culpable control contention that people do not merely view causation for harmful events in terms of norm-violating actions. Rather, when asked to explain their causal judgments, they emphasize that someone did something that warranted blame.

We then conducted one more variation of the penless secretary scenario (N = 71). In addition to asking whether the professor or administrator was more causal, we also asked participants to rate the badness or goodness of the professor's behavior on a 7-point scale ranging from "very bad" to "very good." Ratings of the professor's and the administrative assistant's causal influence replicated the findings of the previous study as well as those of Knobe and Fraser in that the professor was seen as more causal (M = 3.89, SD = 1.45) than the administrative assistant (M = 2.43, SD = 1.46), $t(70) = 4.27$, $p < .001$. Furthermore, the average rating of the professor's behavior was M = 3.00, which was below the scale midpoint, indicating a negative view of his actions. In fact not a single participant rated the professor's behavior above the scale midpoint. Clearly, therefore, the professor's behavior was viewed as relatively "bad" and blameworthy in this context. In the realm of offensive or harmful human behavior, blame is the engine that makes norm violations matter.

Concluding Comments

We have shown that evaluations are an important component of causation judgments for undesirable or harmful actions, and we have demonstrated that causation judgments, at least in some circumstances, are determined by perceived blameworthiness rather than the reverse. Norm violations are important determinants of perceived causal influence, but they are effective because they indicate whether an actor has done something exceptionally good or exceptionally bad. In fact, ascribing causation is, by itself, rarely the ultimate goal of the layperson's behavioral analysis. From the standpoint of the culpable control model, causation is but one of the criteria (along with intent, foresight, foreseeability, and mitigating circumstances) that determines the extent to

which actors are blamed or praised for the consequences they attempt or achieve. Judgments of these criteria, however, are strongly influenced by actors' evaluative reactions to the people involved in an action sequence, their behavior, and the consequence that occurred or could have occurred. These evaluative influences are not exceptions to an otherwise rational process--they are essential components of lay behavioral analyses because they stem from observers' most fundamental motives of discerning what objects, events and people are likely to facilitate their goals and well-being and which endanger their prospects.

References

Alicke, M.D. (2008). Blaming badly. *Journal of Cognition and Culture*, 8, Special Issue—*On Folk Conceptions of Mind, Agency, and Morality*, 179-186.

Alicke, M.D. (1992). Culpable causation. *Journal of Personality and Social Psychology* 63, 368-378.

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

Alicke, M.D., Davis, T.L., Buckingham, J.T., & Zell, E. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin, 34*, 1371-1381.

Alicke, M.D., Davis, T.L., & Pezzo, M.V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition, 12*, 281-308.

Alicke, M.D., & Zell, E. (2009). Social attractiveness and blame. *Journal of Applied Social Psychology*, *39*, 2089-2105.

Bargh, J.A. and Chartrand, T.L. (1999). The unbearable automaticity of being. *American Psychology* 54, 462-479.

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

Collingwood, R.G. (1940). *An essay on metaphysics*. Oxford, Clarendon Press.

Damasio, A. (1994). *Descartes' Error: Emotion, reason, and the human brain*. New York: Putnam

Driver, Julia (2008). "Attributions of Causation and Moral Responsibility." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 423–439. Cambridge: MIT Press.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C. and Kardes, F.R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology* 50, 229-238.

Feinberg, J. (1970). *Doing and deserving: Essays in the theory of responsibility*. Princeton, NJ: Princeton University Press.

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

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI study of emotional engagement in moral judgment. *Science, 293(14),* 2105-2108.

Green, J.D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science, 6(12),* 517-523.

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

Hart, H.L.A., & Honore, T. (1959). *Causation in the law*. London: Oxford University press.

Heubner, B. Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Science, 13(1),* 1-6.

Hitchcock, C., & Knobe, J. (in press). *Journal of Philosophy*.

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

Karlovac, M., & Darley, J.M. (1988). Attribution of responsibility for accidents: A negligence law analogy. *Social Cognition, 6,* 287-318.

Knobe, Joshua (2006). *Folk Psychology, Folk Morality*. Dissertation.

Knobe, J., & and Fraser, B. (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality,* pp. 441–447. Cambridge: MIT Press.Mandel,

D.R., & Lehman, D.R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, *71,* 450-463.

Mazzocco, P., J., & Alicke, M.D. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology, 26,* 131-146.

Osgood, E.E., Suci, G.J. and Tannenbaum, P.H. (1957). *The measurement of meaning*. University of Illinois Press, Urbana, IL.

Piaget, J. (1932). *The moral judgment of the child*. London: Routlege & Kegan Paul.

Pizarro, D., & Bloom, P. (2003). The intelligence of moral intuitions: Comment on Haidt. *Psychological Review, 110(1),* 193-196.

Shaver, K. (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 Behavioural Science, 13,* 238-253.

Appendix A

Positive Characterization; Death Imminent (Aneurysm)

1. Edward Poole was a physician who had recently moved into a new neighborhood.

2. Poole had become friends with the Turnbulls who lived on his street.

3. Mrs. Turnbull asked Poole if he could feed her cat who stayed in an upstairs bedroom while she and her husband were out of town at separate conferences.

4. On Nov. 2 of 2004, Poole used the key that Mrs. Turnbull had given him and headed upstairs to the room in which the Turnbulls kept the cat.

5. Neither Poole nor Mrs. Turnbull realized that Mr. Turnbull had returned home after finding out that his conference had been cancelled at the last minute.

6. Poole tripped over one of the children's toys and made a loud nose as he was going up the stairway.

7. Turnbull heard the noise and took a licensed gun from his drawer.

8. Turnbull shot at Poole just as he was about to enter the room where the cat stayed.

9. The bullet hit Poole in the chest and back and killed him almost instantly.

10. The autopsy conducted on Poole showed that he had suffered a brain aneurysm almost at the same time that he was shot by Turnbull. Thus, Poole would have died even if Turnbull had not shot him.

Positive Characterization; Death Delayed (Brain Tumor)

1. Edward Poole was a physician who had recently moved into a new neighborhood.

2. Poole had become friends with the Turnbulls who lived on his street.

3. Mrs. Turnbull asked Poole if he could feed her cat who stayed in an upstairs bedroom while she and her husband were out of town at separate conferences.

4. On Nov. 2 of 2004, Poole used the key that Mrs. Turnbull had given him and headed upstairs to the room in which the Turnbulls kept the cat.

5. Neither Poole nor Mrs. Turnbull realized that Mr. Turnbull had returned home after finding out that his conference had been cancelled at the last minute.

6. Poole tripped over one of the children's toys and made a loud nose as he was going up the stairway.

7. Turnbull heard the noise and took a licensed gun from his drawer.

8. Turnbull shot at Poole just as he was about to enter the room where the cat stayed.

9. The bullet hit Poole in the chest and back and killed him almost instantly.

10. The autopsy conducted on Poole showed that he had an advanced, inoperable brain tumor that would have killed him within two weeks. Thus, Poole would have soon died even if Turnbull had not shot him.

Positive Characterization; Shot Dead

1. Edward Poole was a physician who had recently moved into a new neighborhood.

2. Poole had become friends with the Turnbulls who lived on his street.

3. Mrs. Turnbull asked Poole if he could feed her cat who stayed in an upstairs bedroom while she and her husband were out of town at separate conferences.

4. On Nov. 2 of 2004, Poole used the key that Mrs. Turnbull had given him and headed upstairs to the room in which the Turnbulls kept the cat.

5. Neither Poole nor Mrs. Turnbull realized that Mr. Turnbull had returned home after finding out that his conference had been cancelled at the last minute.

6. Poole tripped over one of the children's toys and made a loud nose as he was going up the stairway.

7. Turnbull heard the noise and took a licensed gun from his drawer.

8. Turnbull shot at Poole just as he was about to enter the room where the cat stayed.

9. The bullet hit Poole in the chest and back and killed him almost instantly.

Negative Victim Characterization; Death Imminent (Aneurysm)

1. Edward Poole was released from prison after serving an 18 year sentence for the rape of an 11 year-old girl.

2. Poole was living in a neighborhood with an old friend who had also recently been released after serving a 6 year sentence for armed robbery.

3. On Nov. 2 of 2004, Poole broke a window in the house of John Turnbull with a baseball bat and headed upstairs toward the room of his youngest daughter.

4. Turnbull heard the noise and took a licensed gun from his drawer.

5. Turnbull shot at Poole just as he was about to enter the girl's room.

6. The bullet hit Poole in the back and chest and killed him almost instantly.

7. The autopsy conducted on Poole showed that he had suffered a brain aneurysm almost at the same time that he was shot by Turnbull. Thus, Poole would have died even if Turnbull had not shot him.

Negative Victim Characterization; Death Delayed (Brain Tumor)

1. Edward Poole was released from prison after serving an 18 year sentence for the rape of an 11 year-old girl.

2. Poole was living in a neighborhood with an old friend who had also recently been released after serving a 6 year sentence for armed robbery.

3. On Nov. 2 of 2004, Poole broke a window in the house of John Turnbull with a baseball bat and headed upstairs toward the room of his youngest daughter.

4. Turnbull heard the noise and took a licensed gun from his drawer.

5. Turnbull shot at Poole just as he was about to enter the girl's room.

6. The bullet hit Poole in the back and chest and killed him almost instantly.

7. The autopsy conducted on Poole showed that he had an advanced, inoperable brain tumor that would have killed him within two weeks. Thus, Poole would have soon died even if Turnbull had not shot him.

Negative Characterization; Shot Dead

1. Edward Poole was released from prison after serving an 18 year sentence for the rape of an 11 year-old girl.

2. Poole was living in a neighborhood with an old friend who had also recently been released after serving a 6 year sentence for armed robbery.

3. On Nov. 2 of 2004, Poole broke a window in the house of John Turnbull with a baseball bat and headed upstairs toward the room of his youngest daughter.

4. Turnbull heard the noise and took a licensed gun from his drawer.

5. Turnbull shot at Poole just as he was about to enter the girl's room.

6. The bullet hit Poole in the back and chest and killed him almost instantly.