An actor's knowledge and intent are more important in evaluating moral transgressions than conventional transgressions

Carly Giffin

Tania Lombrozo

Department of Psychology
University of California, Berkeley

Keywords: decision making, transgressions, mental states, moral evaluation, punishment

Department of Psychology, 3210 Tolman Hall, # 1650
Berkeley, CA 94720.
Contact: carly.giffin@berkeley.edu

Abstract

An actor's mental states – whether she acted knowingly and with bad intentions – typically play an important role in evaluating the extent to which an action is wrong and in determining appropriate levels of punishment. In four experiments, we find that this role for knowledge and intent is significantly weaker when evaluating transgressions of conventional rules as opposed to moral rules. We also find that this attenuated role for knowledge and intent is partly due to the fact that conventional rules are judged to be more arbitrary than moral rules; whereas moral transgressions are associated with actions that are intrinsically wrong (e.g., hitting another person), conventional transgressions are associated with actions that are only contingently wrong (e.g., wearing pajamas to school, which is only wrong if it violates a dress code that could have been otherwise). Finally, we find that it is the perpetrator's *belief* about the arbitrary or non-arbitrary basis of the rule – not the reality – that drives this differential effect of knowledge and intent across types of transgressions.

1. Introduction

Both laypeople's intuitions and the law accord a prominent role to a perpetrator's knowledge and intentions when it comes to assessing the severity of a transgression and how it should be punished (Young & Tsoi, 2011; Mikhail, 2009; Cushman, 2008). For example, serving someone a cup of coffee sprinkled with poison is deemed quite a bit worse when it was done *intentionally* – with full knowledge that the white powder added to the coffee was poison – than when it resulted from the false belief that the white powder was sugar (Young et al., 2007). To take a legal example, determinations of whether a defendant should be charged with murder versus manslaughter depend, in large part, on whether the killing was intentional.

Although knowledge and intent often play a central role in moral judgment, recent findings reveal that these mental states are not equally influential in evaluating transgressions of all types (Barrett et al., 2016; Chakroff et al., 2015; Hawley-Dolan & Young, 2013; Russell & Giner-Sorolla, 2011; Young & Saxe, 2011). This variation is also reflected in the law: for crimes classified as strict liability, such as speeding or statutory rape, the American legal system does not require the presence of *mens rea* – "a guilty mind" – for a defendant's conviction. Moreover, laypeople's intuitive moral judgments mirror this aspect of the law: Giffin and Lombrozo (2016) found that ignorance was less mitigating for strict liability crimes compared to "mens rea" crimes, such as theft or battery. For instance, people treated ignorance that one was speeding (even if the speedometer was broken) as less mitigating than ignorance that one committed theft (due, for example, to a false belief about an object's ownership).

Here we explore the prediction that knowledge and intent are also differentially important when it comes to evaluating transgressions of moral rules (such as hitting another person) versus conventional rules (such as violating a dress code). Specifically, we test the novel prediction that

these mental states have a larger impact in moral cases relative to conventional cases when evaluating the "wrongness" of a transgressor's actions and how a transgressor should be punished. In the sections that follow, we briefly review prior work that motivates this prediction, and we provide an overview of the four experiments that follow.

1.1 From Strict Liability to the Moral / Conventional Distinction

One basis for our prediction that knowledge and intent may be less important in evaluating transgressions of conventional rules relative to moral rules comes from recent work in the psychology of law. Giffin and Lombrozo (2016) found an attenuated role for knowledge and intent in evaluating strict liability crimes (relative to other crimes), and explained this finding by appeal to a legal distinction between transgressions that are *malum in se*, or wrong in themselves, versus those that are *malum prohibitum*, or wrong because prohibited (*US v. Morissette*, 1952). Battery[1], to take one example, is arguably wrong in itself: it is wrong to hit another person whether or not there are rules against harmful physical contact. Driving at 50-miles per hour, in contrast, is not *inherently* wrong, but becomes wrong if the speed limit is 35-miles per hour. The knowledge and intent that underlie the commission of crimes that are considered *malum in se* versus *malum prohibitum* could vary accordingly. In the former case, the intention to violate a rule (e.g., prohibiting battery) is inextricably bound with an additional intention that is itself reprehensible: an intention to harm. In the latter case, the intention to violate a rule (e.g., prohibiting speeding) is only *contingently* bound to a bad intention: were the rule against speeding not in place or the speed limit raised to 50-miles per hour, intentionally driving 50-

[1] Battery is defined in California penal code 242 as the "(1) unlawful and willful; (2) application of force or violence; (3) upon the person of another." Hitting or applying force to another in self-defense, defense of another, defense of property, or as part of your official duties would not constitute battery.

miles per hour down a well-lit, empty street would be perfectly fine. We might therefore expect that, in general, the intentions that accompany the knowing transgression of a rule that is *malum in se* will be more reprehensible than those that accompany the knowing transgression of a rule that is *malum prohibitum.*

Giffin and Lombrozo (2016) found empirical support for this proposal in two forms. First, they found that when a statute prohibiting a particular act was repealed, judgments about how wrong the corresponding transgression was went down significantly more for strict liability crimes than for mens rea crimes. This suggests that the strict liability offenses were regarded to a greater extent as wrong *because they were prohibited* (i.e., as *malum prohibitum).* Second, Giffin and Lombrozo found that, on average, participants judged strict liability laws more arbitrary than those governing mens rea crimes, suggesting that participants regarded the former as somewhat regulatory, and the latter as linked more intrinsically to non-arbitrary matters of harm – that is, to acts that are *malum in se,* or wrong in themselves.

The distinction between *malum prohibitum* and *malum in se* finds a counterpart in research on moral psychology, where scholars have differentiated between transgressions of moral and conventional rules. In classic experiments, Turiel and colleagues presented children with stories in which an actor violated a rule. The children were asked to judge how bad the actor's behavior was, both with the rule in place and in a situation in which the rule did not apply (Turiel, 2008; Weston & Turiel, 1980). They found that children as young as six judged transgressions of moral rules (such as hitting another child) wrong, even when no explicit rule was in place, suggesting that they found the act to be wrong in itself. However, transgressions of conventional rules (such as a dress code) were only judged wrong when the rule was in place, suggesting the act was wrong merely because it was prohibited.

The importance of the moral/conventional distinction is further supported by more recent developmental work, which finds that children are sensitive to the fact that people can *choose* to opt out of conventional rules, in a way that is not appropriate when the rule is moral (Josephs & Rakoczy, 2016). Moreover, the distinction is reflected in a variety of behaviors that emerge before the age of 6: 5-year-old children tattle and have stronger emotional reactions to moral transgressions than to conventional transgressions (Hardecker, Schmidt, Roden, & Tomasello, 2016), and 3- and 4-year-olds differentiate between moral and conventional rules insofar as they place more emphasis on freedom of action in the moral domain (Josephs, Kushnir, Grafenhain, & Rakoczy, 2016). Specifically, Josephs, Kushnir, Grafenhain, and Rakoczy (2016) found that children were more likely to protest when an agent violated a moral rule if the agent did so when alternative actions were available (as opposed to being constrained in their choice of actions), but that the availability of alternatives had a smaller impact on protests directed towards agents who violated conventional rules.

While scholars differ in how they conceptualize the moral/conventional distinction (e.g., Nichols, 2008; Turiel 2008b; Weston & Turiel, 1980), one important element seems to be the *arbitrariness* of a prohibition: like the rules against strict liability evaluated by participants in Giffin and Lombrozo (2016), conventional rules are somewhat arbitrary in the sense that they could have been different (e.g., the *color* specified by a dress code, the specific side of the plate on which a fork is placed), even when there are good reasons for having some sort of regulation in place. In Josephs, Kushnir, Grafenhain, and Rakoczy (2016), for example, the conventional rules specified which color marble should be placed in which box – a rule that could easily have been different.

The results from Giffin and Lombrozo (2016), combined with the developmental work on the moral/conventional distinction, generate a previously unexplored prediction: that knowledge and intent should have a greater impact on how people evaluate the "wrongness" of transgressions involving moral rules versus conventional rules, with corresponding effects for the levels of punishment deemed appropriate in each case. Put differently, the gap between how wrong it is to hit someone knowingly versus accidentally should be greater than the gap between how wrong it is to break the dress code knowingly versus accidentally. Across four experiments, we find support for this prediction and test competing hypotheses about what drives differential effects of knowledge and intent across different types of transgression.

1.2 Overview of Experiments

In Experiment 1, we find support for our key prediction. While transgressions of both moral and conventional rules are judged more harshly when an actor transgressed knowingly as opposed to unknowingly, the effect of knowledge is greater for transgressions of moral rules than of conventional rules. In Experiment 2, we test two candidate explanations for this effect: that it is driven by greater expected or actual harm in moral cases, or that the critical difference instead lies in the fact that moral transgressions are *malum in se*, whereas transgressions of convention are often *malum prohibitem*, as they involve the transgression of a somewhat arbitrary convention. We find support for the latter possibility, and in Experiment 3 go on to test the role of arbitrariness experimentally. Finally, in Experiment 4, we investigate whether it is the reality (whether a rule is *in fact* arbitrary) or the transgressor's beliefs (whether she *believes* that the rule is arbitrary) that drives judgments.

2. Experiment 1

In Experiment 1, we test the prediction that a perpetrator's knowledge and intent moderate how wrong a transgression is perceived to be and how much punishment is deemed appropriate, but that this role is greater in evaluating transgressions of moral rules (hereafter referred to as "moral transgressions") than of conventional rules (hereafter referred to as "conventional transgressions"). To do so, we compare judgments of "wrongness" (i.e., how wrong an act is judged to be) and punishment across vignettes involving transgressions of a stipulated moral or conventional rule, where the transgression is committed knowingly or unknowingly (i.e., due to an accident or false belief concerning something other than the rule itself). In other words, we test the prediction that the evaluation of moral transgressions is more "knowledge dependent" than the evaluation of conventional transgressions. For the sake of continuity with previous research, we also attempt to replicate the well-established finding that judgments concerning conventional transgressions tend to be more "rule dependent" (i.e., contingent on the presence of a rule) than those concerning moral transgressions.

2.1 Methods

2.1.1 Participants

Two-hundred-and-forty adults (105 female, 134 male, 1 other/prefer not to specify, mean age = 32, *SD* = 15) participated in the study through Amazon Mechanical Turk for monetary compensation. An additional 28 participants were tested, but were excluded for failing catch questions (27) or to ensure even numbers in all conditions (1). Participation was restricted to workers with IP addresses in the United States and an approval rating of 95% or higher on previous tasks.

2.1.2 Materials and procedure

The experimental stimuli consisted of 12 distinct vignettes, six of which concerned conventional transgressions and six of which concerned moral transgressions. Six of the vignettes (Teacher's Title, Greeting, Baseball, Dollar, Pushing, and Embezzler) were based on vignettes originally presented to children by Davidson, Turiel, and Black (1983). The Embezzler vignette was additionally modified to take place in a school setting (see Supplementary Material for full stimuli).

Each participant read only one of the 12 vignettes, leading to 12 conditions. Participants first read the *unknowing* version of their assigned vignette and answered wrongness and punishment questions. For instance, the *unknowing* Baseball vignette read:

> Jack is a boy who likes to play games, especially baseball…There's a rule that all the students on the Blue Jays' team wear blue shirts with the school logo on the back to baseball practice. The school takes this rule and respect for the school and its logo very seriously.
>
> One day, Jack was getting ready for a baseball practice. He was in a hurry to get to the bus on time, so he dressed quickly and left. Jack didn't realize he had grabbed the wrong blue shirt. So Jack went to the practice wearing a blue shirt that did not have the school logo on the back, in violation of the long-standing school policy.
>
> Because Jack was wearing the wrong shirt, he was allowed to practice that day, but was sent to the principal's office after practice.

Participants were then asked, in random order:

> *Wrongness*: "How wrong was [Actor's actions]?" Participants indicated their answer on a scale from 0 (not at all wrong) to 6 (very wrong).

> *Punishment*. "Students who break a rule at [Actor's] school are given school service hours during which they clean classrooms, organize supplies, and pick up trash on the grounds. How many hours of school improvement service should [Actor] get?" Participants indicated their answer on a scale from 0 to 6 hours.

Participants were then asked to imagine that the actor had instead violated the rule knowingly, and again rated wrongness and punishment.[2] Below is the *knowledge change* prompt from the Baseball story:

> *Knowledge Change.* "Suppose that Jack had actually realized, while he was dressing, that the shirt he was about to put on for practice violated the rule – that is, that it didn't have the logo on the back. And suppose that he decided to wear it anyway. In this case, where Jack knowingly violated the rule, how would you respond to the following questions? (Your responses may be the same as those you just provided, or they may differ.)"

In no case (for either moral or conventional transgressions) did the vignette specify that the actor violated the rule with the intention to cause harm or disruption; in most cases the motivation resulted from a desire for change (e.g., being tired of following the rule).

---

[2] In a pilot version of Experiment 1 with 160 participants, the *knowing* and *unknowing* versions of each vignette were presented to different participants in a between-subjects design. This study revealed the same effects of knowledge on judgments of wrongness as those presented here. For punishment, however, the pilot study asked about hours of detention, which seemed to generate a floor effect: participants were disinclined to assign detention hours for any of our transgressions. This motivated the change to "service hours" as the measure of punishment in Experiment 1. See Supplementary Material 2 for a full report and Supplementary Material 3 for full stimuli from that experiment.

Next, participants were told to imagine that the rule was not in place, and answered the evaluative questions a final time. These ratings were solicited to replicate prior findings that conventional transgressions are more rule dependent than moral transgressions, and thus as verification that our vignettes successfully presented transgressions that were "moral" versus "conventional" in the relevant sense. The wording of these questions (again presented in random order) was identical to that above, but preceded by a rule change.[3] Below is the *rule change* prompt from the Baseball story:

> *Rule Change.* "Finally, suppose that Jack's school had no rule prohibiting wearing a shirt without the school logo to practice, and Jack knowingly wore a shirt without the school logo to practice. In this case, with no rule about how to dress for practice in place, how would you respond to the following questions? (Your responses may be the same as those you've provided, or they may differ.)"

Finally, on a separate screen, participants were presented with a true/false comprehension question relating to the vignette they had just read, and answered one additional catch question, modeled after Oppenheimer, Meyvis, and Davidenko (2009). These questions were used to assess whether participants had read the vignette and the instructions carefully; those who

[3] This "no rule" rating was always solicited last. We feared that if we initially told participants that no rule was in place, and then later told them a rule had been instituted, they might make unwarranted assumptions about the reasons for this change. For instance, participants might assume that the actions had actually caused harm or disruption, making the rule necessary. These kind of assumptions could artificially inflate wrongness and punishment ratings. Moreover, because the primary aim of the experiment was to assess the knowledge dependence of moral and conventional violations, it seemed wise to solicit those judgments first.

answered either question incorrectly were excluded from further analyses. To conclude, participants answered demographic questions about their age and gender.

2.2 Results

Participants responded to the wrongness and punishment questions three times: for the initial *unknowing* transgression (with a rule in place), the subsequent *knowing* transgression (with a rule in place), and the final *no rule* transgression (with no rule in place). We first present the results from the first two sets of judgments as a measure of "knowledge dependence," and we then consider the final two sets to evaluate "rule dependence."

2.2.1 Knowledge dependence

To test the prediction that judgments regarding conventional transgressions are less "knowledge dependent" than those regarding moral transgressions, we performed mixed ANOVAs with knowledge status as a within-subjects variable (2: *knowing*, *unknowing*), transgression domain as a between-subjects variable (2, conventional, moral), and either wrongness or punishment ratings as the dependent variable (see Figure 1). We expected to find a main effect of knowledge status, with higher ratings for knowing transgressions than for unknowing transgressions, qualified by an interaction between knowledge status and transgression domain, with a larger effect of knowledge status for moral transgressions than for conventional transgressions.

As predicted, this analysis revealed a main effects of knowledge status for both wrongness, $F(1,238) = 465.33$, $p < .001$, $\eta_p^2 = .66$, and punishment, $F(1,238) = 382.16$, $p < .001$, $\eta_p^2 = .62$; in both cases, ratings were higher for knowing than for unknowing transgressions (see Figure 1). However, this main effect was qualified by the predicted interaction between knowledge status and transgression domain, for both wrongness, $F(1,238) = 12.84$, $p < .001$, η_p^2

= .05, and punishment, $F(1, 238) = 33.06$, $p < .001$, $\eta_p^2 = .12$. Independent-samples t-tests comparing the average difference between participants' ratings for the *knowing* and *unknowing* vignettes confirmed that the knowledge effect was greater for moral than conventional transgressions for both wrongness ($M_M = 3.01$, $SD_M = 1.74$; $M_C = 2.11$; $SD_C = 1.64$), $t(238)= 4.12$, $p< .001$, $d = .53$, and punishment ($M_M = 2.68$, $SD_M = 1.81$; $M_C = 1.45$; $SD_C = 1.34$), $t(221)= 5.97$, $p< .001$, $d = .80$ (corrected for violating Levene's test).

Finally, there was also a main effect of transgression domain for both wrongness, $F(1,238) = 36.51$, $p < .001$, $\eta_p^2 = .13$, and punishment, $F(1,238) = 73.47$, $p < .001$, $\eta_p^2 = .24$, with higher ratings for moral transgressions than for conventional transgressions.

Figure 1: Ratings for wrongness and punishment for all vignette versions as a function of transgression type. Error bars correspond to one SEM in each direction.

2.2.2 Rule Dependence

Contingency on the presence of a rule ("rule dependence") was measured by subtracting participants' wrongness and punishment ratings after the rule change from their corresponding scores for the *knowing* vignettes. Independent t-tests were performed on these difference scores, and we predicted a greater difference for conventional relative to moral transgressions.

As predicted, these analysis found that rule dependence was significantly greater for conventional than moral transgressions for wrongness (1.91 versus 1.43), $t(238)= 6.93$, $p < .001$,

d = .90, and nearly so for punishment (2.76 versus .942), $t(221) = 1.89$, $p < .06$, d = .25 (corrected for violating Levene's test).

2.3 Discussion

Experiment 1 found that relative to judgments concerning conventional transgressions, those concerning moral transgressions were more sensitive to whether the actor transgressed knowingly versus unknowingly. This was the case for judgments of wrongness and also for our measure of punishment (service hours). To our knowledge, this is the first demonstration that a transgressor's mental states differentially influence judgments of wrongness and punishment when it comes to moral versus conventional transgressions. We also found that relative to judgments concerning moral transgressions, those concerning conventional transgressions were more contingent on the presence of a rule. This is a familiar finding from the literature on the moral/conventional distinction, but helps confirm that our vignettes varied appropriately along this dimension.

While Experiment 1 supports our key prediction that knowledge and intent have a greater impact in evaluating moral versus conventional transgressions, the findings do not reveal why. An important question thus remains: what is it about the domain of the moral versus the conventional that drives the differential role of knowledge and intent? One possibility is that the moral transgressions in Experiment 1 were simply more severe than the conventional transgressions: if they involved greater harm, the associated intentions may have been regarded as more "wrong" and deserving of punishment. Indeed, moral transgressions do tend to be more severe (Smetana, 1995), and Experiment 1 found a main effect of domain, with harsher judgments for moral vignettes than for conventional vignettes.

On the other hand, past work has found that severity is insufficient to fully account for differences in judgment across the moral and conventional domains (Tisak & Turiel, 1988), though scholars have offered competing accounts of what the additional difference might be (Nichols, 2008; Turiel 2008b; Weston & Turiel, 1980). One account argues that conventions are fundamentally *arbitrary* (Turiel, 2008a), whereas moral norms stem from non-arbitrary considerations of harm (Turiel, 2008b). For instance, it's important that we have *some* convention about which side of the road to drive on, but the specific rule (to drive on the right versus the left) is arbitrary, varying from country to country. In contrast, moral rules are less likely to be regarded as arbitrary matters of preference (Goodwin & Darley, 2012; see also Sarkissian, et al., 2011).

This suggests a second possibility: that the relevant difference between the moral and conventional domains stems from the fact that moral rules are non-arbitrary and intrinsically linked to harm. As a result, knowingly violating a moral rule is intrinsically linked with an intention to harm, whereas knowingly violating a conventional rule is only contingently associated with bad intentions: there's nothing wrong with the intention to wear a shirt without a school logo; it only becomes wrong in the context of a rule prohibiting this behavior. On this view, it is not the *degree* of harm caused that differentiates the moral from the conventional, but rather *how* the harm is linked to the transgressive act. Reflecting this difference, Table 1 represents how attributions of knowledge and intent might shift at each step of Experiment 1: when the transgression is committed unknowingly, when it occurs knowingly, and when it occurs with no rule in place.

	Moral Transgression	Conventional Transgression
Unknowing Transgression	Knowledge of rule	Knowledge of rule
Knowing Transgression	Knowledge of rule Intention to break rule Intention to cause harm	Knowledge of rule Intention to break rule
No Rule	Intention to cause harm	

Table 1. Representation of the knowledge and intent one might attribute to the perpetrator at each step of Experiment 1, as a function of domain.

As the table reveals, the difference in mental states across the knowing and unknowing transgressions is greater for moral transgressions than for conventional transgressions: the former involves an additional "intention to cause harm" in the knowing case. We propose that this explains the greater knowledge dependence of moral transgressions relative to conventional transgressions. However, it's the conventional cases that involve a greater shift from the knowing transgression to the scenario without a rule: for the moral transgressions, the intention to cause harm is preserved across both cases, but for conventional cases no knowledge or intent remains. We propose that this explains the greater rule dependence of conventional transgressions relative to moral transgressions.

Of course, the violation of a conventional rule could *also* be accompanied by an intention to harm. For instance, someone could drive on the wrong side of the road with the intention to harm other drivers, or violate a dress code with the intention of hurting someone's feelings. Our contention is not that such intentions do not occur, but rather that that are not entailed by the intentional rule-breaking itself. In contrast, it's difficult to imagine a moral transgression that is *not* accompanied by an intention to harm, even if such an intention is unspecified (as was the

case for our vignettes). Except in very rarefied cases, an intention to hit or steal involves an intention to harm.

If this analysis is correct, then a key difference between moral and conventional transgressions stems from the fact that conventional rules are somewhat arbitrary – this is what renders the intention to violate a conventional rule only contingently associated with additional bad intentions. Experiment 2 tests this possibility while controlling for the overall severity of the transgressions.

3. Experiment 2

In Experiment 2, we sought to replicate the key finding from Experiment 1: that whether a transgression is committed knowingly versus unknowingly has a greater impact on the evaluation of moral transgressions relative to conventional transgressions. However, we also sought to identify what it is about moral and conventional transgressions that drives this differential effect.

To test the possibility that the moral vignettes simply involved greater harm than the conventional vignettes, we more closely matched our *knowing* moral and conventional transgressions along a variety of dimensions, including the amount of harm that a third person would have expected from the transgression, the amount of harm the actor foresaw, and the amount of harm that actually occurred. To accomplish this we pretested our stimuli, and additionally verified that these dimensions were matched in a post-test. If differences along these dimensions are responsible for the greater knowledge dependence of moral transgressions relative to conventional transgressions, then we should expect the domain differences observed in Experiment 1 to disappear in Experiment 2.

A second possibility is that knowledge dependence is driven by more than just actual harm, and instead depends on the nature of the perpetrator's intentions. Because conventional rules are somewhat arbitrary and not intrinsically linked to harm, the intentions associated with knowingly breaking a conventional rule are only bad contingently. These ideas generate two pairs of predictions. First, conventional rules should be regarded, on average, as more arbitrary than moral rules, in the sense that they could reasonably have been specified differently. Moreover, perceived arbitrariness should be a negative predictor of the magnitude of the knowledge effect. Second, the intentions associated with knowing moral transgressions should be regarded, on average, as worse than the intentions associated with conventional transgressions, and perceived badness should be a positive predictor of the magnitude of the knowledge effect. We test these predictions in Experiment 2.

3.1 Methods

3.1.1 Participants

Two-hundred-and-eighty adults (155 female, 124 male, 1 other/prefer not to specify, mean age = 35, *SD* = 11) participated in the study through Amazon Mechanical Turk as in Experiment 1. An additional 42 participants were tested, but were excluded for failing catch questions (36) or to ensure even numbers in all conditions (6).

3.1.2 Materials & Procedure

The experimental stimuli consisted of eight vignettes, five of which were used in the previous experiment (Dress Code, Lunch Table, Tardy, Throwing, and Pushing), and three of which (Hall Monitor, Candy, Burner) were created for this experiment after pre-testing items to better equate harm across transgression domains (see Supplementary Material 4 for full stimuli).

Each participant read only one of the eight vignettes, leading to eight distinct conditions. As in Experiment 1, participants first read the *unknowing* version of their assigned vignette and answered the wrongness and punishment questions used in Experiment 1. Participants were then asked to provide the same ratings assuming the actor had violated the rule knowingly, using the same knowledge change prompt as in Experiment 1. The order of the wrongness and punishment questions was randomized in each case.

After these ratings, participants rated the event, assuming the actor did know the relevant facts, along a series of dimensions related to harm and intent. The questions were presented in randomized order, and all were rated on a scale of 1 (none at all) to 7 (a great deal). The questions are provided below:

Expected Harm. "How much harm could [Actor's] actions have been expected to cause?"

Foreseen Harm. "How much harm do you think [Actor] believed his/her actions would cause?"

Actual Harm. "How much harm did [Actor's] actions actually cause?"

Intentions. "How wrong was [Actor's] intention when she behaved this way?"

Next, participants provided a rating of how arbitrary they found each of the rules used in Experiment 1 as well as the new rules added for Experiment 2, even though they had only read one rule. Participants used a 1 (not at all arbitrary) to 7 (completely arbitrary) rating scale. Below are the instructions that preceded the arbitrariness judgments:

> "Below you will see a list of school rules. We'd like you to give your intuitions about how arbitrary each rule seems to you. That is, do you believe that there's a good reason for the rule to draw the line where it does in terms of which actions are allowed versus not allowed? Or does it seem like the rule is somewhat arbitrary in the sense that it

could reasonably have been formulated a little differently, with a slightly different set of actions ruled in versus ruled out?

Please do not consult any outside resources, like other people or websites. We are interested in your own intuitions. Even if you think all the rules are somewhat arbitrary or not at all arbitrary, please take note of which seem more or less arbitrary and respond accordingly."

One concern in equating our vignettes in terms of various facets of harm is that in doing so, we may have inadvertently blurred or erased the boundary between moral and conventional transgressions. In addition to the question about arbitrariness, we therefore introduced an explicit judgment concerning the transgression domain. Participants made judgments about all rules used in Experiments 1 and 2 using a rating scale from 1 (clearly about morality) to 7 (clearly about conventions/norms). Below is the paragraph that preceded these judgments:

Morality of Action. "Actions can be wrong in multiple ways. Some are morally wrong, such as vandalizing a car. Others are wrong because they violate a group convention or norm, such as driving on the wrong side of the road.

Please rate the rules below on the following scale, from (1) clearly a rule about something moral to (7) clearly a rule about a group convention or norm. Please do not consult any outside resources, like other people or websites. We are interested in your own intuitions."

The relatively minor offense of vandalizing a car was used to help participants focus on the intended dimension (moral versus conventional) rather than on severity.

Finally, participants answered the same catch and demographic questions used in Experiment 1.

3.2 Results

3.2.1 Harm Ratings

First, we verified that we succeeded in equating the moral and conventional vignettes along various dimensions of harm. We ran independent samples t-tests comparing the three harm ratings as a function of domain (2: moral, conventional). These tests revealed that for Expected Harm and Believed Harm, there were no significant differences across domains, *p*s > .174. For ratings of Actual Harm there was a significant difference, $t(229) = 4.08$, $p < .001$, $d = .54$ (corrected for violating Levene's test), but ratings were higher for the *conventional* vignettes than for the moral vignettes ($M_C = 2.84$, $SD_C = 1.97$; $M_M = 2.04$, $SD_M = 1.19$). These findings confirm that we succeeded in making the moral transgressions no worse than the conventional transgressions along these dimensions of harm.

3.2.2 Domain Ratings

In light of the ways in which our vignettes were matched in terms of harm, it was important to test whether the moral and conventional vignettes were still clearly differentiated in terms of their perceived domain. Averaging judgments for the four moral transgressions versus the four conventional transgressions, a paired-samples t-test revealed that our moral rules were found to be based significantly more on moral precepts than were our conventional rules (M_M =2.63, $SD_M = 1.06$; $M_C = 5.98$, $SD_C = 1.01$), $t(279) = 31.71$, $p < .001$, $d = 3.80$. The same difference was observed if we considered a single rating from each participant corresponding to the moral basis of the rule in the vignette that they evaluated, and performed an independent samples t-test with domain as a between-subjects factor (M_M =2.61, $SD_M = 1.83$; $M_C = 5.96$, $SD_C = 1.36$), $t(257) = 17.32$, $p < .001$, $d = 2.16$.

3.2.3 Knowledge Dependence

Having verified that harm was no greater for the moral vignettes than the conventional vignettes (despite preserving a difference in perceived domain), we went on to test the prediction that judgments about conventional transgressions should still be less knowledge dependent than corresponding judgments about moral transgressions. Mirroring Experiment 1, we performed mixed ANOVAs with knowledge status as a within-subjects variable (2: *knowing*, *unknowing*), transgression domain as a between-subjects variable (2: conventional, moral), and either wrongness or punishment ratings as the dependent variable (see Figure 2). As in Experiment 1, we expected to find a main effect of knowledge status qualified by an interaction between knowledge state and transgression domain, with a larger effect of knowledge for moral than conventional transgressions.

Mirroring Experiment 1, the analysis revealed a significant main effect of knowledge status for both wrongness, $F(1,278) = 366.20$, $p < .001$, $\eta_p^2 = .568$, and punishment, $F(1,278) = 374.73$, $p < .001$, $\eta_p^2 = .574$, with higher ratings for the *knowing* vignettes. As predicted, these main effects were qualified by significant interactions between knowledge status and transgression domain for both wrongness, $F(1,278) = 7.42$, $p < .007$, $\eta_p^2 = .03$, and punishment, $F(1,278) = 11.45$, $p < .001$, $\eta_p^2 = .04$. Independent samples t-tests on the average difference score between the *knowing* and *unknowing* ratings revealed that the knowledge effect was significantly greater for moral than conventional transgressions for both wrongness (M_M =2.44, SD_M = 1.95; M_C = 1.83, SD_C = 1.77), $t(278) = 2.73$, $p < .007$, $d = .41$, and punishment (M_M =1.80, SD_M = 1.35; M_C = 1.26, SD_C = 1.30), $t(278) = 3.38$, $p < .001$, $d = .41$.

The main effect of transgression domain was not significant for wrongness, $F(1,278) = 2.72$, $p < .431$, $\eta_p^2 = .002$, but was significant for punishment, $F(1,278) = 8.90$, $p < .003$, $\eta_p^2 =$

.031, with punishment being higher for conventional (M_{UK} = 1.53, SD_{UK} = 1.64; M_K = 2.79, SD_K = 1.87) than moral (M_{UK} = .76, SD_{UK} = .89; M_K = 2.56, SD_K = 1.59) transgressions.

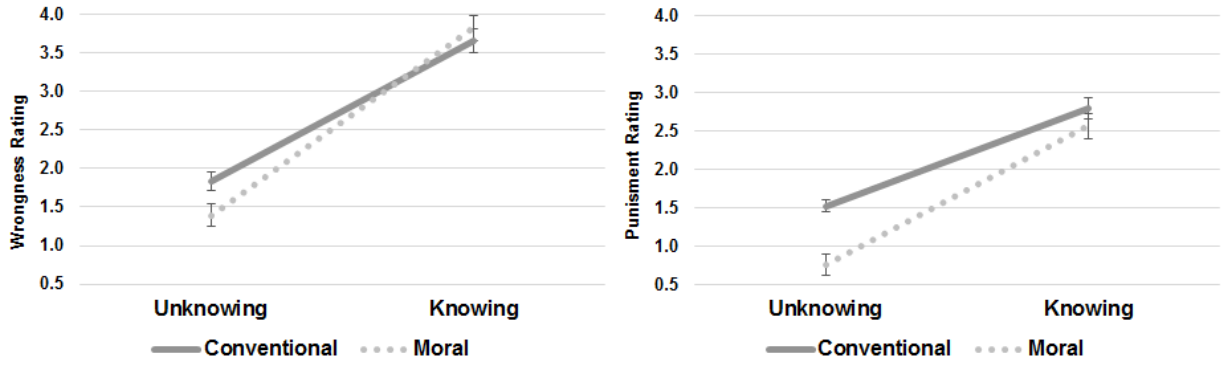

Figure 2: Ratings for wrongness and punishment for both knowledge states as a function of transgression type. Error bars correspond to one SEM in each direction.

3.2.4 Arbitrariness Ratings

To assess whether there were domain differences in the perceived *arbitrariness* of different types of transgressions, we computed a pair of averages for each participant: the average arbitrariness ratings for the four rules corresponding to our moral vignettes versus those corresponding to our conventional vignettes. A paired samples t-test revealed that moral rules were rated significantly less arbitrary than conventional rules (M_M =2.65, SD_M = .019; M_C = 3.85, SD_C = .076), $t(279) = 11.87$, $p < .001$, $d = 1.42$. This same pattern was observed if we considered a single rating from each participant corresponding to the arbitrariness of the rule in the vignette that was evaluated, and performed an independent samples t-test with domain as a between-subjects factor (M_M =2.62, SD_M = 1.95; M_C = 3.81, SD_C = 1.97), $t(278) = 5.06$, $p < .001$, $d = .61$.

3.2.5 Intention Ratings

We also predicted that the perpetrator's intention would, on average, be perceived as more wrong in the moral vignettes than in the conventional vignettes. However, an independent

samples t-test comparing intention ratings as a function of domain did not reveal a significant difference, $t(278) = .561$, $p < .575$, $d = .07$. It may be that in equating foreseen harm we effectively erased domain differences in intent, or that by asking specifically about the wrongness of the perpetrator's *intentions*, we missed out on other relevant mental states.

3.2.6 Predicting Knowledge Effects

As in Experiment 1, we found that judgments concerning moral transgressions were more sensitive to knowledge than corresponding judgments about conventional transgressions. We also found, as predicted, that participants rated conventional rules as significantly more arbitrary than moral rules. However, participants did not judge the intention of the actor to be significantly worse in the moral vignettes, as we had predicted. Regardless, we still predicted that the arbitrariness of the rule and the wrongness of the intention would be significant predictors of the knowledge effect – with arbitrariness as a negative predictor and wrongness of intention as a positive predictor. Thus, we ran simultaneous linear regressions for both wrongness and punishment with arbitrariness and intention ratings as predictors of the difference in ratings between the *knowing* and *unknowing* vignettes.

	Unstandardized Coefficients		**Standardized Coefficients**		
Model	B	Standard Error	Beta	t	Significance
Constant	1.862	.292		6.374	.000
Arbitrariness	-.146	.053	-.158	-2.724	.007
Intention	.199	.051	.225	3.885	.000

Table 2: Regression results for the wrongness knowledge effect.

As predicted, the arbitrariness of the rule was a significant, negative predictor while the wrongness of the actor's intention was a significant, positive predictor for both the wrongness

knowledge effect, $R^2 = .078$, $F(2, 277) = 12.77$, $p < .001$, and the punishment knowledge effect, $R^2 = .062$, $F(2, 277) = 10.19$, $p < .001$ (see Tables 2 and 3).

	Unstandardized Coefficients		**Standardized Coefficients**		
Model	B	Standard Error	Beta	t	Significance
Constant	1.358	.211		6.444	.000
Arbitrariness	-.094	.039	-.142	-2.435	.016
Intention	.129	.037	.203	3.471	.001

Table 3: Regression results for the punishment knowledge effect.

3.3 Discussion

Experiment 2 replicated both the punishment and wrongness knowledge effects found in Experiment 1. We found that judgments concerning moral transgressions were more sensitive to knowledge than corresponding judgments concerning conventional transgressions, despite the fact that none of the harm measures were significantly higher for moral transgressions.

Further, Experiment 2 offers support for our proposal that the knowledge effect is tracking the arbitrariness of the rule that is transgressed, as well as the intentions of the actor. We found that participants rated conventional rules to be more arbitrary, and that arbitrariness was a negative predictor of the magnitude of knowledge effects.[4] We also found that the wrongness of the intention was a positive predictor of both knowledge effects.

[4] Notably, the relationship between arbitrariness and knowledge effects held not only across domains, but also within domains: further analyses revealed negative correlations between both knowledge effect scores and how arbitrary a participant rated the rule in the vignette they read, whether analyses were restricted to only moral or conventional vignettes. For moral violations, the correlations between arbitrariness score and the wrongness knowledge effect, $r = -.160$, $p < .059$, and punishment knowledge effect, $r = -.149$, $p < .079$, were both negative and

One of our predictions was not confirmed: the actor's intentions were not perceived as significantly more wrong for knowing moral transgressions than for knowing conventional transgressions. One possible explanation is that our intention question was simply too narrow to capture the full range of mental states that participants considered. For instance, participants might have believed that while intentions were not significantly worse in the moral vignettes, the actor should have taken more care.

In sum, Experiment 2 succeeded in replicating the domain differences observed in Experiment 1 while ruling out the possibility that these differences were driven by greater anticipated, foreseen, or actual harm in the moral vignettes. The findings also supported three of our four novel predictions: we found that the conventional rules were judged to be more arbitrary than the moral rules, and that arbitrariness was associated with a smaller knowledge effect. We also found that the wrongness of the intention associated with a knowing transgression predicted the magnitude of the knowledge effect, but we did not find that intentions were perceived to be more wrong in the moral vignettes than the conventional vignettes.

4. Experiment 3

In Experiments 1 and 2 we found support for the prediction that relative to conventional transgressions, the evaluation of moral transgressions is more sensitive to the perpetrator's knowledge and intent. In Experiment 2, we found that this difference remains when moral and conventional transgressions are matched across a variety of dimensions of harm. Experiment 2

---

marginally significant. For conventional violations the correlations between arbitrariness and the wrongness knowledge effect, $r = -.133$, $p < .116$, and the punishment knowledge effect, $r = -.084$, $p < .324$, were still negative, though farther from significance.

also suggested that a relevant difference between moral and conventional transgressions is the extent to which the violated rule is perceived to be *arbitrary* in the sense that it could reasonably have been otherwise. In Experiment 3 we pursue this association experimentally by manipulating whether otherwise-matched actions violate rules that are arbitrary or non-arbitrary.

All rules in Experiment 3 involved the prevention of harm, and thus fall under the moral domain. In order to effectively manipulate whether such rules were perceived to be more or less arbitrary, we developed vignettes in an alien world for which we could stipulate associated harms without challenging participants' preconceptions about what was or was not arbitrary. Each rule specified some threshold – for example, that (alien) students were not permitted to watch gory movies below the age of 15, because younger students were more likely than older students to suffer negative emotional consequences from doing so. Thus, the presence of *some* rule was not arbitrary, but the exact threshold of 15 years *could* be. Half the participants were told that the threshold was based on a discrete developmental change that occurs at exactly age 15, and was therefore not arbitrary. The remaining participants were told that there was a more gradually changing continuum, so the exact choice of 15 was arbitrary in that 14.5 or 15.5 could reasonably have been chosen instead. We could thus test the prediction that judgments would be more sensitive to mental states for non-arbitrary rules relative to arbitrary rules using actions that were otherwise identical.

4.1 Methods

4.1.1 Participants

Two-hundred-and-forty adults (103 female, 135 male, 2 other/prefer not to specify, mean age = 33, *SD* = 11) participated in the study through Amazon Mechanical Turk as in Experiments

1-2. An additional 90 participants were tested, but were excluded for failing catch questions (46) or to ensure even numbers in all conditions (44).

4.1.2 Materials & Procedure

The experimental stimuli consisted of six distinct vignettes (Movies, Humming, Vitamins, Dumping, Laser Gun, and Speeding). The stimuli from Experiment 3 involved an alien school on an alien planet. The shift to an alien world, while regrettable in some respects, was deemed necessary as a way to exert experimental control over participants' beliefs about the relationship between different actions and possible harms.

Each vignette had two variants, one in which the rule had been set arbitrarily, and one in which the rule had been set non-arbitrarily, with arbitrariness as a between-subjects factor. For example, in the Humming vignette, all participants learned about a rule concerning where they could practice loud and potentially disruptive alien humming: they were not allowed to do so within 15 feet of an occupied classroom. The relevant excerpt from the arbitrary and non-arbitrary versions of the Humming vignette are presented below (see Supplementary Material 5 for full stimuli):

> *Arbitrary*. "…The school administration chose the distance of 15 feet somewhat arbitrarily. There isn't any reason to suspect that the noise would be much more disruptive at 14.5 feet than at 15.5 feet, but they had to choose some distance cutoff, and decided on 15 feet…"
>
> *Non-Arbitrary*. "…The school administration chose the distance of 15 feet after careful research on the unique auditory capabilities of their species. Aliens of their species can hear the noises clearly if they are made within exactly 15 feet. Beyond that distance, however, the noises fall below the critical threshold for the alien's auditory system, so they are barely audible and no longer distracting…"

Participants first read the *unknowing* version of their assigned vignette and answered the questions about wrongness and punishment from Experiments 1-2. As in Experiments 1-2, participants were then asked to imagine that the actor had knowingly violated the rule and to answer the questions once again. The question order was randomized in each case.

Finally, as in Experiment 1, participants were asked to imagine that the rule was not in force and to answer the wrongness and punishment questions a third and final time. In this experiment, we took care to inform the participant that the rule change was not due to a change in the harm caused by the action. We took this extra step because those participants who read an arbitrary version of a vignette might be more likely to believe that the rule had been changed because the action was not actually causing harm at the arbitrarily selected threshold. That is, as the rule was set arbitrarily, a rule change might be taken as an indication that the rule was ill conceived and unnecessary from the start. Below is a sample from the Humming vignette:

> *Rule Change.* "Finally, suppose that due to a clerical error, the rule against humming within 15 feet of a classroom had never been put in the school charter and therefore could not be enforced. That is, suppose that Bernice's school EFFECTIVELY had no rule prohibiting humming within 15 feet of an occupied classroom, and Bernice practiced her humming at her locker. In this case, with no rule about humming in place, how would you respond to the following questions? (Your responses may be the same as those you've provided, or they may differ.)"

After participants read all three versions of their assigned vignette (*unknowing*, *knowing*, and *no rule*), they answered catch questions and demographic questions similar to those in Experiments 1-2.

4.2 Results

4.2.1 Knowledge Dependence

To test our prediction that knowledge and intent play a greater role in evaluating transgressions of non-arbitrary rules than of arbitrary rules, we performed mixed ANOVAs with rule type (2: arbitrary, non-arbitrary) and vignette (6: Movies, Humming, Vitamins, Dumping, Laser Gun, and Speeding) as between subjects' factors, knowledge state (2: *unknowing*, *knowing*) as a within-subjects measure, and either wrongness or punishment (service hours) as the dependent variable (see Figure 3). We predicted a greater knowledge effect for the non-arbitrary transgressions, mirroring the greater knowledge effect seen for moral transgressions. This would manifest as an interaction between rule type and knowledge state.

As predicted, we found a significant interaction between knowledge state and rule type for both wrongness, $F(1,228) = 6.96$, $p < .009$, $\eta_p^2 = .03$, and punishment, $F(1,228) = 13.31$, $p < .001$, $\eta_p^2 = .06$: the increase in ratings from the *unknowing* condition to the *knowing* condition was greater for non-arbitrary transgressions than for arbitrary transgressions, both for judgments of wrongness ($M_{NA} = 2.37$, $SD_{NA} = 1.70$; $M_A = 1.81$, $SD_A = 1.60$), $t(238) = 2.62$, $p < .009$, $d = .34$, and for punishment ($M_{NA} = 2.28$, $SD_{NA} = 1.55$; $M_A = 1.58$, $SD_A = 1.38$), $t(238) = 3.65$, $p < .000$, $d = .47$.

The analysis also revealed a main effect of knowledge state, with higher ratings in the *knowing* condition than the *unknowing* condition for both wrongness, $F(1,228) = 388.96$, $p < .001$, $\eta_p^2 = .63$, and punishment, $F(1,228) = 414.22$, $p < .001$, $\eta_p^2 = .65$. We did not find a main effect of rule type, for either wrongness, $t(238) = 0.55$, $p > .580$, $d = .07$, or punishment, $t(238) = 0.27$, $p > .787$, $d = .04$. There were no main effects nor interactions involving vignette.

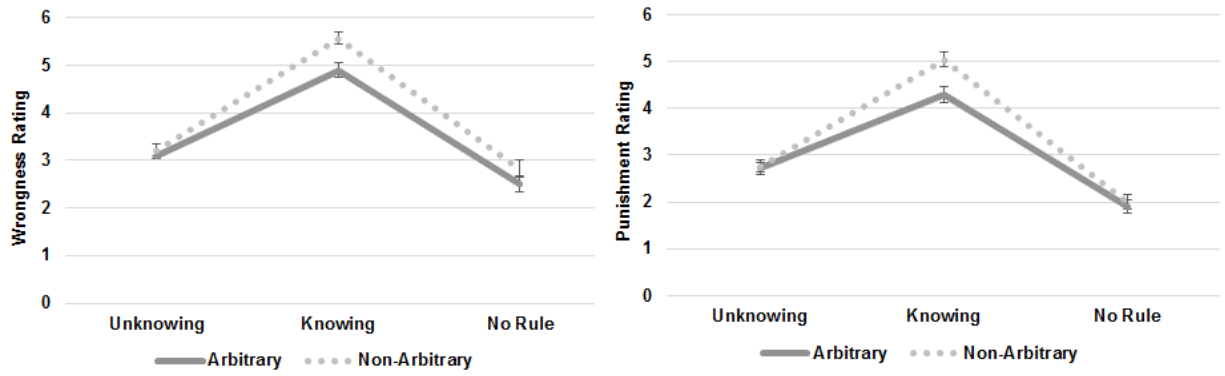

Figure 3: Ratings for wrongness and punishment for all mental states as a function of transgression type. Error bars correspond to one SEM in each direction.

4.2.2 Rule Dependence

Finally**,** might transgressions of *arbitrary* rules mimic conventional transgressions in their contingency on a rule? To answer this question, we created a measure to reflect the effect of the rule, following the same procedure as in Experiment 1. We performed a series of 2 (rule type: arbitrary, non-arbitrary) x 6 (vignette: Movies, Humming, Vitamins, Dumping, Laser Gun, and Speeding) ANOVAs to assess the rule effect. Contrary to our expectations, we found no significant difference between arbitrary and non-arbitrary transgressions for wrongness ($p > .19$), but we did see a significant difference for school service hours $F(1,228) = 6.31$, $p < .013$, $\eta_p^2 = .027$.

4.3 Discussion

Experiment 3 confirmed our prediction that knowledge dependence would be greater for non-arbitrary transgressions relative to arbitrary transgressions. We hypothesized that part of what drives the knowledge effect for moral rules is that they are typically non-arbitrary. That is, because moral rules prohibit actions that are intrinsically tied to harm, the rules could not reasonably have been otherwise (i.e., they are non-arbitrary), and a knowing transgression constitutes a knowing commission of harm. Experiment 3 tested this by setting all violations in

the moral domain, but manipulating the arbitrariness of the rules. The key prediction was that a knowing violation of a non-arbitrary rule would be deemed more wrong and more deserving of punishment than a knowing violation of an arbitrary rule, resulting in an interaction between arbitrariness and knowledge. This is precisely what we found.

Participants' ratings for both wrongness and punishment (school service hours) were less sensitive to knowledge and intent when the threshold specified by the rule was arbitrary. Thus, we were able to replicate the pattern of knowledge dependence seen in moral and conventional transgressions entirely within the moral domain by manipulating whether the rule was set arbitrarily. This finding supports our findings from Experiment 2 that the importance of these mental states is not solely a function of whether a norm involves harm, as all the transgressions considered in Experiment 3 were moral in nature and tied to a real harm, but is also influenced by the arbitrariness of the rule itself – that is, whether the rule could reasonably have been specified otherwise. These results leave open, however, whether the critical issue is *reality* (whether the rule is in fact set arbitrarily or not) or the actor's *belief* about the rule; we return to this issue in Experiment 4.

Experiment 3 failed to find significantly greater effects of a rule change for arbitrary relative to non-arbitrary rules for wrongness, although this predicted effect was observed for punishment. We suspect that this failure to replicate the pattern observed on Experiment 1 comes from the smaller effect sizes associated with rule type in the present experiment, and the corresponding loss in statistical power.

5. Experiment 4

While Experiments 2 and 3 found support for the idea that mental states are more influential when a rule is non-arbitrary, it's not clear what was driving this effect: that rules were *believed* by the actors to have arbitrary versus non-arbitrary bases, or that the *reality* was such that some transgressions crossed an arbitrarily set threshold, while others crossed a non-arbitrary threshold. In other words, was the effect driven by the knowledge and intentions themselves, or by their differential consequences in the world?

Based on our predictions and findings in Experiments 1-2, we would expect a critical factor to be the knowledge and intent of the actor. In Experiments 1-2, we suggested that moral transgressions involve a greater role for knowledge and intent because some desires and intentions are themselves inextricably bound to the harm that the action will cause, thereby influencing the perceived magnitude of wrongdoing and the appropriate level of punishment. While Experiment 2 found that intention ratings significantly predicted the magnitude of the knowledge effect, the intentions of moral transgressors were not rated significantly worse.

Thus, Experiment 4 aims to take a closer look at the role of intentions and beliefs in driving knowledge effects. If it's a perpetrator's mental states that drive such effects, then the effects found in Experiment 3 should be eliminated by stipulating that all transgressors *believed* the rules to be set arbitrarily, while varying the reality: for half of participants the rule was arbitrary (i.e., it could reasonably have been specified differently); for the other half it was not (i.e., their transgression involved crossing a non-arbitrary threshold). If instead judgments for wrongness and punishment are affected by the differential *consequences* of violating a rule that was or was not set arbitrarily (i.e., intrinsically tied to *actual* harm or not), then specifying that a transgressor always believes a rule to be arbitrary should be insufficient to eliminate the effect.

5.1 Methods

5.1.1 Participants

Two-hundred-and-forty adults (117 female, 122 male, 1 other/prefer not to specify, mean age = 32, *SD* = 12) participated in the study through Amazon Mechanical Turk as in Experiments 1-3. An additional 185 participants were excluded for failing catch questions.

5.1.2 Materials & Procedure

The experimental stimuli consisted of three of the stories used in Experiment 3 (Movies, Vitamins, and Speeding). Each vignette had two versions, one in which the actor believed the rule had been set arbitrarily, and in fact it had been, and one in which the actor believed the rule had been set arbitrarily, but in fact the rule had been set after careful, but secret, research. The three stories chosen were those that could be credibly adapted to meet these constraints. The relevant excerpts of the arbitrary and non-arbitrary version of the Movies vignette are presented below. The relevant rule was that students could not lend particular movies to any student under age 15 due to their gory content (for full stimulus materials, see Supplementary Material 6):

> *Arbitrary*: "The school administrators chose the age of 15 based on the recommendation from the studio. The studio chose the age of 15 somewhat arbitrarily. There isn't any reason why seeing the gory content would harm a 14.5-year old alien much more than a 15.5-year old alien, but they had to choose some age cutoff, and decided on 15.
>
> Neither the school administrators nor the students have any idea how the studio chose the age and believe the age was selected arbitrarily. In fact, Bernice once asked a school administrator why the age was 15, and was told that it was an arbitrary cut-off determined by the movie studio."

Non-Arbitrary: "The school administrators chose the age of 15 based on the recommendation from the studio. The studio chose the age of 15 after careful research on alien development. Aliens of their species undergo an important developmental change such that younger aliens are susceptible to the negative effects of gory and violent content, while older aliens are not. The developmental change is linked to an increase in hormone levels that occurs, like clockwork, at age 15. The studio feared that releasing this information would damage their sales and reputation so they have carefully kept it from the public. Only the top executives and a few scientists even know the study was conducted. Neither the school administrators nor the students have any idea how the studio chose the age and believe the age was selected arbitrarily. In fact, Bernice once asked a school administrator why the age was 15, and was told that it was an arbitrary cut-off determined by the movie studio."

Participants were randomly assigned to read one of the six vignettes. As in Experiments 1-3, they read an *unknowing* version followed by wrongness and punishment questions in a randomized order, and then answered the questions again after seeing a knowledge change prompt similar to those used in Experiments 1-3. Next, participants were shown the rule change manipulation from Experiments 1 and 2. We used the rule change from Experiments 1 and 2, in case the absence of a significant rule change effect for wrongness in Experiment 3 stemmed from the way the question was asked. Finally, participants answered catch questions as in previous experiments.

5.2 Results

5.2.1 Knowledge Dependence

If the differential effects of knowledge and intent found in Experiments 1-3 were driven by differences across conditions in the mental states attributed to transgressors, then we should not find differential knowledge effects here: in all cases, the transgressor had the belief that a given norm was arbitrary. To test this prediction, we performed mixed ANOVAs with rule type (2: arbitrary, non-arbitrary) as a between subjects' factor, knowledge state (2: *unknowing*, *knowing*) as a within subjects' factor, and either wrongness or punishment (service hours) as the dependent variable. Consistent with our prediction, we did *not* find an interaction between rule type and knowledge state for either wrongness, $F(1,238) = 0.770$, $p = .381$, $\eta_p^2 = .003$, or punishment, $F(1,238) = 1.24$, $p = .267$, $\eta_p^2 = .005$ (see Figure 4). There was, unsurprisingly, a main effect of knowledge state for both wrongness, $F(1,238) = 440.62$, $p < .000$, $\eta_p^2 = .649$, and punishment, $F(1,238) = 360.62$, $p < .000$, $\eta_p^2 = .602$.

To compare the results of Experiments 3 and 4 directly, we ran an additional analysis in which we treated experiment (Experiment 3 vs. Experiment 4) as a between-subjects variable. To make the two experiments as comparable as possible, we included only the three vignette pairs corresponding to those used in Experiment 4. We then performed 2 (experiment: 3, 4) x 2 (basis for rule: arbitrary, non-arbitrary) ANOVAs with the knowledge effect difference score as the dependent variable (i.e., rating for *knowing* transgression minus rating for *unknowing* transgression). This analysis revealed a significant interaction for punishment (service hours), $F(1, 356) = 5.40$, $p < .021$, $\eta_p^2 = .015$, and a marginally significant interaction for wrongness, $F(1, 356) = 3.11$, $p <= .079$, $\eta_p^2 = .009$. In both cases, the knowledge effect was greater in Experiment 3 than in Experiment 4, which supports the idea that it's an agent's beliefs about the

harm associated with rule breaking, and not the actual harm caused, that partially or wholly determines differential effects of mental states across types of rule breaking.

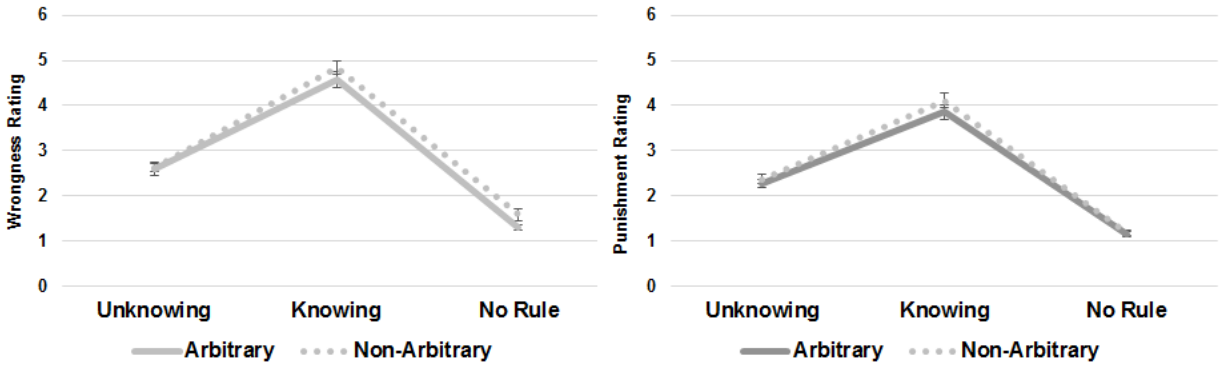

Figure 4: Ratings for wrongness and punishment (service hours) for all vignette variants as a function of transgression type. Error bars correspond to one SEM in each direction.

5.2.2 Rule Dependence

Difference scores were created as in previous Experiments to test for rule dependence. We did not find any significant difference between the arbitrary and non-arbitrary conditions (all *p*'s > .35).

5.3 Discussion

Experiment 4 supports our prediction that it is an actor's *belief* that a rule is arbitrary that participants use to evaluate wrongness and ascribe punishment, not the actual consequences of transgressing an arbitrary or non-arbitrary rule: the differential knowledge effects across conditions completely disappeared when the actor's knowledge was held constant across different realities. This finding supports our broader contention that what drives the knowledge effect in moral and conventional (or non-arbitrary and arbitrary) transgressions is an actor's knowledge and intent concerning his or her actions.

Experiment 4 also supports Experiment 2 in suggesting that the presence, absence, or degree of *actual* or *potential* harm are not what really drive our knowledge effects. If harm were

driving the differential knowledge effect, we would have expected to see a greater knowledge effect when a transgression violated a rule that was in fact non-arbitrary relative to cases in which the rule was in fact arbitrary, regardless of the perpetrator's beliefs. This, along with the careful matching of harm in Experiment 2, provides compelling support that it's a transgressor's mental states that are predominantly responsible for the magnitude of the knowledge effect.

6. General Discussion

In Experiments 1 and 2 we found evidence that the evaluation of moral transgressions is more knowledge dependent than the evaluation of conventional transgressions – that is, "knowing what you're doing" is more damning, and ignorance is more mitigating, when the rule that's violated is moral as opposed to conventional. We also found that the evaluation of conventional transgressions is more rule dependent than that of moral transgressions, replicating findings from past literature. In Experiments 2-4, we found evidence that part of what drives differential effects of knowledge and intent is the typically non-arbitrary nature of moral rules, and especially a transgressor's *beliefs* about the basis for a violated rule.

The finding that relative to moral transgressions, the perceived severity of conventional violations is contingent on the presence of a rule is not new; however, to our knowledge, this is the first demonstration that judgments concerning the wrongness of moral transgressions are more sensitive to knowledge and intent than those concerning conventional transgressions, with corresponding effects for punishment. That said, our findings nicely complement recent work with children by Josephs, Kushnir, Grafenhain, and Rakoczy (2016), who found that another aspect of an actor's actions – whether they were chosen freely or under constraint – had a greater impact on children's protests concerning violations of moral versus conventional rules. Our

study differs from this recent work in manipulating actors' mental states directly (rather than manipulating constraints on their choices), and in considering explicit judgments of wrongness and punishment (rather than children's "protests"). Our study is thus the first to systematically vary whether a transgressor "knew what she was doing" in violating a rule, and to do so across moral and conventional cases.

Despite the novelty of our result, the relationship between knowledge dependence and the conventionality of a rule is consistent with prior work on the evaluation of strict liability crimes. In Giffin and Lombrozo (2016), we argued that knowledge is less important in laypeople's judgments concerning strict liability crimes because such crimes tend to involve the violation of a rule with somewhat arbitrary – and therefore arguably "conventional" – elements. For example, speeding is a strict liability crime, and it involves the violation of a somewhat arbitrary speed limit. Driving 40 miles per hour is not inherently wrong, but it is wrong when it occurs in a 35-mile zone, and the designation of 35 miles (as opposed to 34 or 36.5) as the limit is somewhat arbitrary. It is not clear that the harm or consequence is much different from 35 to 40 or 40 to 45, even though there's a good reason for specifying *some* speed limit. In other words, the rule is arbitrary in the sense that it could reasonably have been specified differently. This feature of "arbitrariness" could help explain why strict liability crimes and conventional violations behave similarly.

Experiment 2 found support for an association between arbitrariness and knowledge dependence, and Experiment 3 set out to manipulate the role of arbitrariness directly. We hypothesized that if part of what makes conventional violations less sensitive to a transgressor's knowledge and intent is their more arbitrary specification, then rules that are expressly characterized as somewhat arbitrary – even when they involve harm – should also show less

sensitivity to these mental states. This is precisely what we found: judgments concerning the violation of arbitrarily set rules depended less strongly on the actor's knowledge and intent, whether participants were evaluating the severity of wrongdoing or the number of service hours that should be required as punishment. Finally, in Experiment 4 we found evidence that what matters is the transgressor's *belief* that a rule has been set arbitrarily; not necessarily the actual consequences of breaking a rule that involves an arbitrary threshold.

Our findings raise a number of important questions for future research. First, how do our findings relate to prior work (Barrett et al., 2016; Chakroff et al., 2015; Hawley-Dolan & Young, 2013; Russell & Giner-Sorolla, 2011; Young & Saxe, 2011), which documents a weaker effect of knowledge and intent in evaluating purity violations, such as incest, relative to harm violations, such as battery? Like conventional transgressions, purity violations differ from moral transgressions in the primary locus of harm. In a typical harm violation, such as poisoning, the victim is another person (Gray & Wegner, 2009, 2012; Gray, Young, & Waytz, 2012). This is also the case for the moral transgressions considered in our experiments. However, in the case of purity violations, such as incest, the victim is often the self (Young & Tsoi, 2013; Young & Saxe, 2011). Conventional transgressions could mirror purity violations not in the focus on self per se, but in lacking another person as an identifiable victim. Supreme Court Justice Morissette argued that legal wrongs that are *malum prohibitum* are offenses against *the authority of the state*: even in the absence of certain harm to others, disregarding convention is a potential harm to society or social order. Individuals may be harmed downstream, but they are not typically identifiable as individuals at the time of the transgression. It could be that the central importance of knowledge and intent in moral transgressions stems from the role of an identifiable victim other than the self. In such cases, it may be especially important to track the mental states of the

perpetrator to evaluate moral character and prevent future harms. Potentially consistent with this line of thought, Josephs, Kushnir, Grafenhain, and Rakoczy (2016) explain their results by appeal to the idea that violations of moral norms involve a focus on the *actor*, whereas violations of conventional norms shift the focus from the actor to the *consequences* of the violation.

Second, in what sense must rules be arbitrary for the role of mental states to be attenuated? If our account is right, then the critical factor is whether a rule is linked to intrinsic harm, and thus constrained such that it couldn't reasonably have been otherwise. In our experimental manipulation of arbitrariness (Experiments 3-4), we created rules that were arbitrary because they set a threshold that could reasonably have been otherwise. It was useful to adopt this approach to arbitrariness because all vignettes were set in the moral domain and involved some level of harm. It's worth noting, though, that this differs from the way in which conventional rules are often arbitrary: rather than identifying a critical threshold, many specify conditions for group coordination (e.g., driving on the right rather than the left). We expect that comparable effects would be found in such cases, but this is at present an untested prediction. A further question concerns the relationship between conventionality and arbitrariness and whether all forms of conventionality involve some inherently arbitrary elements.

Third, might there be important boundary conditions on the effects we report here? It's instructive to consider conventional transgressions accompanied by truly bad intentions. For example, consider a person who intentionally drives on the wrong side of the road, not only with the intention of violating traffic rules, but also with the intention to generate a harmful collision. Although the anticipated harm stems from a violation of convention, the addition of the intention to cause harm seems to shift the offense from the domain of the conventional to the domain of the moral: the primary offense isn't in violating rules of traffic, but in violating rules against

aggravated battery or perhaps vehicular homicide. We take it as evidence for our distinction between conventional and moral transgressions that it is so difficult to construct instances of purely conventional transgressions that involve genuinely bad intentions. It appears that once an action is accompanied by genuinely bad intentions, it shifts domains from a conventional to a moral transgression.

Fourth, how might culture moderate the effects we report? Prior work has found that the relative importance of intent in moral evaluations can differ across cultures (Barrett et al., 2016). Moreover, there is documented cultural and individual variation in the perceived boundary between moral and conventional wrongdoing (Hauser, Cushman, & Young, 2008). Our vignettes are set in a modern, westernized context, and our sample is from the United States. It is possible that cultures that exhibit an attenuated role for mental states in moral evaluation will show even less sensitivity when it comes to conventional violations. Investigating other cultures will be crucial in painting a more accurate and complete picture of human moral psychology.

Finally, how might our results from Experiments 3-4 translate to more "earthly" situations? We chose to situate Experiments 3-4 in an alien world because we feared that participants' beliefs about how arbitrary – or not – a rule is on earth would be relatively hard to manipulate. However, to the extent we could convincingly state that our actors reasonably believed a real moral rule was set arbitrarily, we would expect to see a reduction in the influence of mental states on judgments. Precisely which mental states matter and why, however, is another question that merits further research.

In sum, our findings are consistent with prior work demonstrating the importance of mental states in moral judgment. However, our findings go beyond prior work in pointing to a differential role for knowledge and intent when it comes to evaluating moral versus conventional

transgressions. We also suggest a basis for this difference: to the extent a rule is believed to reflect non-arbitrary considerations, the intention to violate the rule will be bound up with intentions to cause other consequences, such as harm. This makes a knowing transgression more severe, and ignorance more mitigating, for moral transgressions than for conventional transgressions.

References

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113* (11), 4688–4693. doi: 10.1073/pnas.1522070113

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2015). When minds matter for moral judgment: intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, 1-9. doi:10.1093/scan/nsv131

Cushman, F. (2008). Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment. *Cognition*, *108*, 353-380. doi: 10.1016/j.cognition.2008.03.006

Davidson, P., Turiel, E., & Black, A. (1983). The Effect of Stimulus Familiarity on the Use of Criteria and Justifications in Children's Social Reasoning. *British Journal of Developmental Psychology, 1*, 49-65. doi: 10.1111/j.2044-835X.1983.tb00543.x

Giffin, C. & Lombrozo, T. (2016). Wrong or Merely Prohibited: Special Treatment of Strict Liability in Intuitive Moral Judgment. *Journal of Law & Human Behavior, 40* (6), 707-720. doi: 10.1037/lhb0000212

Goodwin, G. P., & Darley, J. (2012). Why are some moral beliefs seen as more objective than others? *Journal of Experimental Social Psychology, 48*, 250-256. doi: 10.1016/j.jesp.2011.08.006

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*, 505–520. doi:10.1037/a0013748

Gray, K., &Wegner, D. M. (2012). Morality takes two: Dyadic morality and mind perception. In M. Mikulincer, & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil*. (pp. 109–127).Washington, DC: American Psychological Association. doi:10.1037/13091-006

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*, 206–215. doi:10.1080/1047840X.2012.686247

Hardecker, S., Schmidt, M., Roden, M., & Tomasello, M. (2016). Young children's behavioral and emotional responses to different social norm violations. *Journal of Experimental Child Psychology*, *150*, 364-379. doi:10.1016/j.jecp.2016.06.012

Hauser, M. D., Young, L., & Cushman, F. A. (2008). Reviving Rawls' linguistic analogy. In W. Sinnott-Armstrong (Ed.), *Moral Psychology and Biology*. New York: Oxford University Press.

Hawley-Dolan, A., & Young, L. (2013). Whose Mind Matters More—The Agent or the Artist? An Investigation of Ethical and Aesthetic Evaluations. PLoS ONE 8(9): e70759. doi:10.1371/journal.pone.0070759.

Josephs, M. & Rakoczy, H. (2016). Young children think you can opt out of social-conventional but not moral practices. *Cognitive Development, 39*, 197-204. doi:/10.1016/j.cogdev.2016.07.002

Josephs, M., Kushnir, T., Grafenhain, M. & Rakoczy, H. (2016). Children protest moral and conventional violations more when they believe actions are freely chosen. *Journal of Experimental Child Psychology, 141*, 247-255. doi: /10.1016/j.jecp.2015.08.002

Mikhail, J. (2009). Is the Prohibition of Homicide Universal? Evidence from Comparative Criminal Law. *Brooklyn Law Review, 75*, 497-516.

Nichols, S. 2008. Sentimentalism Naturalized. In W. Sinnott-Armstrong (ed.) *The Psychology and Biology of Morality*. Cambridge, MA: MIT Press.

Nucci, L. P, & Herman, S. (1982). Behavioral disordered children's conceptions of moral, conventional, and personal issues. *Journal of Abnormal Child Psychology, 10*, 411-426. doi: 10.1007/BF00912330

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872. doi:10.1016/j.jesp.2009.03.009

Russell, P., & Giner-Sorolla, R. (2011). Moral Anger, but Not moral Disgust, Responds to Intentionality. *Emotion, 11*(2), 233-240. doi: 10.1037/a0022598

Sarkissian, H. Parks, J., Tien, D., Wright, J.C., & Knobe. J. (2011) Folk moral relativism. *Mind & Language, 26* (4), 482-505. doi: 10.1111/j.1468-0017.2011.01428.x

Smetana, J. G. (1995). Morality in context: Abstractions, ambiguities, and applications. In R. Vasta (ed.), *Annals of child development* (Vol 10, pp. 83-130). London: Jessica Kingsley.

Smetana, J. & Braeges, J. (1990). The Development of Toddlers ' Moral and Conventional Judgments. *Merrill-Palmer Quarterly, 36* (3), 329-346.

Tisak, M. & Turiel, E. (1988). Variation in Seriousness of Transgressions and Children's Moral and Conventional Concepts. *Developmental Psychology, 24*(3), 325-357. doi: 10.1037/0012-1649.24.3.352

Turiel, E. (2008a). The Development of Children's Orientations toward Moral, Social, and Personal Orders: More than a Sequence in Development. *Human Development, 51*(1), 21-39. doi: 10.1159/000113154

Turiel, E. (2008). Thought About Actions in Social Domains: Morality, Social Conventions, and Social Interactions. (2008). *Cognitive Development, 23*, 136-154. doi: 10.1016/j.cogdev.2007.04.001

Turiel, E. (1994). The development of social-conventional and moral concepts. In *Fundamental Research in Moral Development*. Puka, B. (Ed.). New York, NY: Routledge.

United States v. Morissette, 342 U.S. 246 (1952).

Weston, D. & Turiel, E. (1980). Act-Rule Relations: Children's Concepts of Social Rules. *Developmental Psychology, 16(5),* 417-424. doi: 10.1037/0012-1649.16.5.417

Young, L., Cushman, F., Hauser, M., Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *PNAS,104* (20), 8235-8240. doi:10.1073/pnas.0701408104

Young, L., & Saxe, R. (2011). When Ignorance is No Excuse: Different Roles for Intent Across Moral Domains. *Cognition, 3,* 202–214. doi: 10.1016/j.cognition.2011.04.005

Young, L. & Tsoi, L. (2013). When Mental States Matter, When They Don't, and What That Means for Morality. *Social and Personality Psychology Compass*, 7(8), 585-604. doi: 10.1111/spc3.12044

Acknowledgements

We gratefully acknowledge Elliot Turiel for helpful discussion and for sharing stimulus materials. This work also benefited from conversations with Auden Dahl, Mahesh Srinivasan, Shaun Nichols, and Walter Sinnot-Armstrong and his lab group. This work was partially supported by NSF Career award DRL-1056712 and a James S. McDonnell Foundation Scholar Award in Understanding Human Cognition.