The Meanings of Moral Wrongs

Nicolas Restrepo

## Introduction

Cultural differences in moral worldviews are well documented in both academic journals and newspaper headlines. The renewed interest in the sociology of morality has been accompanied by an impulse to document these fault-lines (Hitlin and Vaisey 2013). Researchers have shown that practices that are mundane amongst certain groups cause outrage amongst others. These cultural differences, however, are aggregates: they are the result of the everyday judgments of individuals who are adept at intuiting what is right and what is wrong. In tracing the variation of moral cultures, we make the assumption that the members of different social groups see immorality in different places. Nonetheless, the process whereby actors come to perceive a practice as immoral remains opaque. In this paper, I shed light on how individuals attribute immorality to events, thus providing more robust cognitive foundations for the sociological study of morality.

There is widespread debate about how actors identify and categorize moral transgressions. Currently, one of the most plausible explanation is that we attribute moral wrongdoing through template matching (Schein and Gray 2018). In other words, when we perceive an event, we compare it to our idea of what the typical moral transgression looks like. If the perceived event closely matches our cognitive template of an immoral act, then we are likely to view it as a transgression. Even though this theory has gained traction, the evidence to support it remains limited. Here, I provide an attempt to operationalize and directly test it. I take advantage of the fact that the bulk of research on moral cognition has used hypothetical vignettes. Using sociological approaches to measuring social meaning, I can produce “translations” of these vignettes which allow me to examine the connotative meanings of these fictional scenarios. This helps me accomplish two goals: first, I identify shared semantic features amongst the scenarios in order to come up with a principled way of outlining what a prototypical immoral act looks like; second, I calculate the extent to which the vignettes match this template, aiming to shed light on how distance from the exemplar is related to the attribution of immorality. By looking at the underlying semantic attributes of moral scenarios, then, it is possible to provide an empirical test of one of the leading theories of moral decision-making.

My argument will have the following structure. First, I will outline the theoretical debates that inform and motivate this analysis. Second, I will explain how we can use sociological approaches to social meaning in order to translate vignettes into a format amenable for the analysis of their underlying semantic features. Third, I will describe the methods and results of the first study, which shows that the translations of the scenarios are interpreted consistently by respondents. Fourth, I discuss the design and findings of study 2, where I use reaction time data to examine whether the attribution of immorality occurs through template matching. Lastly, I will close by discussing the implications and limitations of the study, as well as the potential avenues for research that it opens up.

## Theoretical Background

### Towards a unified vision of morality

Although research on how individuals come to perceive actions as immoral has been disjointed, some common ground has been reached. Evidence from different fields suggests that actors do not reach moral judgments through complex logical deliberations, weighing their commitments to universal principles against the immediate implications of their actions (Greene 2014, 2015). Rather, they rely on “gut feelings” to instinctively decide what is right or wrong. We intuitively know what is impermissible, even if we are not capable of articulating why (Haidt and Joseph 2004). Questions remain, though, about the underlying cognitive processes that make individuals such swift arbiters of immorality, and about how these vary across social contexts. In answering these questions, scholars have generally put forward two types of explanations: one furthers that there are qualitatively different types of moral transgressions, and that actors use distinct cognitive strategies to understand them; another suggests that moral wrongs are all judged based on one underlying logic.

Explanations that contend that morality consists of differentiated domains have emerged as a response to an empirical puzzle: the fact that there are events that individuals find immoral, but which do not involve harm (Haidt, Roller, and Dias, 1993). Dietary and sexual taboos are often cited as examples of these seemingly harmless wrongs. The most prominent theory in this line of work is Moral Foundations Theory (henceforth MFT) (Graham et al. 2013). MFT scholars argue that the diversity of moral judgements is explained by the fact that morality is underpinned by five distinct foundations: authority, fairness, loyalty, purity, and harm. We attribute immorality differently depending on the moral foundation that appears to be breached. For instance, seeing a soccer player bend the rules to their favor is substantively different that watching them deliberately try to injury a rival. Both are immoral acts but while the former evokes normative sentiments about fairness, the latter is concerning because of the harm that might be inflicted on the other player. Variations in moral worldviews arise as individuals and groups emphasize certain foundations – or combinations of foundations – over others (Haidt and Joseph 2007). MFT, then, provides a pluralist vision of moral judgement, capable of explaining diversity in moral prohibitions.

A more unified conceptualization of morality – the one that MFT precisely argues against - has regained traction throughout the last decade. This push has been motivated by evidence that shows that the five moral foundations are all highly correlated (Gray and Keeney 2015b). This suggests that transgressions are cut from the same cloth and cannot be parceled out into different groups. Gray and Schein (2018) contend that the unifying dimension of moral decision-making is harm. Their work shows that when individuals are asked to envision immoral acts, they tend to bring up quintessentially harmful practices, and that when asked to explain what makes certain actions immoral, they tend to resort to ideas of harmfulness (Schein and Gray 2015, 2018). Individuals, thus, judge actions as immoral insofar as they can perceive them as harmful. Perceptions of harm even mediate our understanding of practices where no entities appear to be directly affected: incestual marriages, for instance, are often described as harming the health – the gene pool – of the group (Schein and Gray, 2018). Their argument is that, regardless of the superficial differences between moral violations, the central mechanism through which individuals come to see them as immoral is harm. Thus, rather than being based on differentiated logics, moral judgement is underpinned primarily by perceptions of harm.

Now, whether morality is conceived as differentiated or continuous has important implications for our understanding of how individuals come to perceive actions as immoral. The former position operates under the implicit assumption that moral cognition is mostly about classification. The attribution of immorality consists of mapping practices unto the appropriate moral code. Nonetheless, this vision of moral cognition as underpinned by discrete classification is not tenable given current evidence. As current research compels us to adopt a more continuous conceptualization of moral decision-making, we need to find cognitive models that are congruent with such a view. While alternative explanations have emerged, we still lack the empirical grounding to adjudicate their validity. In this study, I address this gap by providing an empirical test for one of the most plausible contemporary theories of moral cognition: the idea that individuals attribute immorality through template matching.

### Template matching and moral cognition

Currently, there is a working theory of moral cognition that is congruent with a more unified vision of morality and it relies on the notion of template matching (Gray, Young, and Waytz 2012). Template matching is a cognitive process whereby individuals categorize information by contrasting it with salient mental exemplars built from experience and socialization (Schein and Gray 2018; Harnad 2017). Thus, category membership is not ascribed through inclusionary/exclusionary criteria but rather through proximity to a mental template. For example, when individuals are asked to picture a bird, they are more likely to envision something that looks like a robin than an ostrich (Rosch 1999). What the proponents of this position ague, then, is that when we are asked to picture a moral transgression, we have a particular mental image of what that looks like (Schein and Gray 2018). When confronted with an event, we compare it with this cognitive template in order to assess its harmfulness and immorality. Moral cognition, then, does not entail sorting an event into discrete categories. Instead, it involves placing a violation on a continuum, more or less distant from our mental template of a typical moral wrong.

Proponents of this theory provide a description of this mental template, but many details remain opaque. Immoral actions, they argue, have a dyadic structure: they involve an intentional agent directing a damaging behavior towards a vulnerable patient (Schein and Gray 2018). An event that would resemble such a template would be a calculating criminal murdering a child. When we witness an event, we compare it to this template and assess its immorality accordingly. Though plausible, this definition remains vague. It is unclear how we can define what constitutes a vulnerable victim or a damaging behavior without incurring in tautologies. Therefore, we do not know how to gauge whether an event fits into this template, let alone how to calculate how practices might deviate from it. Thus, even though there is a plausible working definition of the structure of prototypical moral wrongs, we still lack a clear strategy to operationalize it.

Furthermore, the evidence supporting this view of moral cognition is scant. We lack direct empirical tests of whether the attribution of immorality indeed occurs through template matching. As a consequence, it is unclear how we are supposed to conceptualize this process. For instance, it is possible to conceive the relationship between immorality and distance from the exemplar as linear: the further away an event is from our cognitive template, the harder it is for us to perceive it as immoral. However, the relationship could also be parabolic. As events become more distant from our mental template, the attribution of immorality becomes more difficult but only up to one point. After that, practices start to look obviously not immoral. Thus, events that are very far away from our idea of what a typical moral transgression is – such as hugging one’s mom – are quickly flagged as innocuous. While the idea of moral cognition occurring through template matching has gained prominence, we lack even basic understanding about how this process works. My argument is that we can start to build a principled empirical understanding of this process by analyzing the underlying semantic features of moral transgressions.

### The social meanings of moral trangressions

The central challenge in the effort to test this theory is outlining a principled definition of the prototypical moral violation. An important step towards that goal is to realize that mental templates are eminently social (Hunzaker and Valentino 2019; Hunzaker 2016). The cultural landscapes in which we are embedded shape how we envision certain categories, and the exemplars around which they are organized. For instance, recent research suggests that our ideas of what the typical African American (Monk 2019) or benefit claimant looks like (Hunzaker and Valentino 2019; Martin and Desmond 2010) vary depending on the social positions we occupy. The key implication here is that we can treat the effort to outline a template of moral transgressions in the same way as scholars have outlined prototypes in other social categories. We can examine the underlying cultural meanings associated with immoral acts in order to look for regularities, piecing together an outline of a prototypical moral transgression. There are two main challenges in this approach: first, we are dealing not with a single entity but with a dyadic structure, consisting of an agent directing a behavior towards an recipient; second, we need to find rigorous measures of cultural meaning that we can map onto these different components. Given fortuitous theoretical overlaps, it is possible to use the dictionaries of affective meanings collected by Affect Control Theory (henceforth ACT) scholars to address these challenges.

ACT is premised upon the idea that individuals act in order to maintain the cultural meanings associated with their identities (Heise 2007). Due to the central role that meaning plays in this theory, ACT scholars have developed rigorous strategies to approach its measurement and analysis. For the purposes of this paper, it is worth highlighting three ideas that capture ACT’s approach to meaning: first, that identities and behaviors have distinct cultural meanings; second, that these meanings are widely shared amongst members of the same group; and third, that these semantic structures can be reduced to a set of measurable dimensions (Heise 2007; Hunzaker 2016). In order to operationalize the latter, this theory incorporates Osgood et al.’s (1975) work, arguing that the meaning of any concept can be effectively distilled to three dimensions: evaluation, potency, and activity. In other words, we need three pieces of information to broadly capture the cultural meaning of a concept: we need to know whether it is good or bad (evaluation), whether it is weak or strong (potency), and whether it is lively or quiet (activity) (Osgood 1969). Extensive cross-cultural work has shown that these three dimensions broadly capture connotative meanings, both across cultures and throughout time (Osgood et al. 1975).

Using these dimensions, it is possible to collect rigorous measures of the meanings that groups attach to concepts. These aggregate meanings are collected by asking respondents to rate concepts on three, nine-point scales corresponding to each of the dimensions (Heise 2014). These measures are then aggregated, and the average is computed, yielding the evaluation, potency, and activity (EPA) values for each concept. Importantly, these EPA profiles are accurate and informative. Babies, for instance, are very good, very powerful, and somewhat lively, while murderers are depicted as very bad, very powerful, and slightly active. This framework, then, helps render cultural meanings measurable and formalizable. As part of their research agenda, ACT researchers have collected large dictionaries of affective meanings. I take advantage of these dictionaries to map reliable measures of cultural meaning onto the different components of moral transgressions.

Research on moral cognition has primarily used fictional scenarios to examine how actors assess immorality. In these studies, respondents are presented with a series of hypothetical events and they must decide how immoral or harmful they are. These vignettes tend to share a dyadic grammatical structure much like the one discussed above: they often involve an actor directing a behavior towards a recipient. Incidentally, this structure (actor-behavior-object) is the main unit of analysis of ACT. This means that the dictionaries of affective meanings contain many of the terms we would need to translate widely used fictional scenarios into a format that allows for the analysis of their connotative meanings. Let me provide an example. A scenario that has been shown (Clifford et al. 2015) to be interpreted reliably is:

You see a teacher hitting a student’s hand with a ruler for falling asleep in class.

Ascribing appropriate EPA profiles to each component of the sentence, we can produce the following translation:

Teacher [E = 2.5; P = 2.31; A = 0.32] hits [E = -2.66; P = 1.30; A = 2.12] lazy student [E = -0.42; P = -0.76; A = -1.56]

This format provides a wealth of additional information about these vignettes. Much like an x-ray, the translation allows to go below the level of denotative meanings and to examine the underlying semantic building blocks of the event. This information, in turn, is useful in the effort to understand template matching in the context of moral cognition. For instance, it is possible to examine what semantic features play an important role in the attribution of immorality. Furthermore, we can look for identifiable attributes of typically harmful actions – such as murder – and prototypically vulnerable victims – such as children. Thus, by going down to the level of connotative meanings, we gain valuable pieces of information that will prove essential for testing whether moral cognition occurs through template matching.

In the following two studies, I use this semantic approach to assess the prototypicality of moral transgressions and to analyze how deviations from those regularities shape moral decision-making. In Study 1, I analyze which semantic attributes – e.g. the agent’s evaluation or the object’s activity – are most informative for understanding the attribution of immorality. This, in turn, helps me reach a principled formalization of what the template of an immoral act looks like. In Study 2, I examine if an event’s distance from said template is predictive of how long it takes individuals to categorize it as immoral or harmful. Using reaction time data, then, I provide a test of whether the attribution of immorality occurs through template matching.

## Study 1

### Data and Methods

The data for this study was collected through the Prolific platform. Prolific is a crowdsourcing website that allows researcher to connect with participants interested in completing different types of tasks. This platform has advantages over its counterparts: it focuses on data quality and has a replenishing pool of participants (Peer et al. 2017). The latter is important because it ensures that respondents are not so familiar with surveys and experiments that completing such tasks becomes a purely mechanistic endeavor. In total, I reached out to 205 participants and, after removing cases with missed attention checks, the total number of usable responses was 194.

For this study, I translated 25 widely used vignettes into the format that allows me to explore their connotative meanings. In order to compile this list, I drew on the work of Gray and Keeney (2015a) and Clifford et. al. (2015) I chose scenarios that had been previously used and that I could translate using ACT’s dictionaries of meaning. Importantly, I also decided to include five scenarios for each of the five moral foundations. Although the idea that moral wrongs belong to distinct categories has been recently criticized, it remains useful to examine whether systematic differences exist. The full list of translations is provided in the supplemental materials. Each participant was presented with the 25 translated scenarios and asked to rate how immoral, harmful, and unexpected they were. They provided their ratings using five-point Likert scales that went from, for example, “not harmful at all” to “extremely harmful”. The decision to include these three scales was based on previous work, which had examined these dimensions (Gray and Keeney 2015a; Schein and Gray 2018).

This study has two main goals: first, to confirm that the translations are interpreted appropriately and that they produce results that are consistent with previous findings; second, to examine which semantic features have the biggest impact in perceptions of immorality. In order to accomplish the former, I compare the average immorality rating of each scenario with its harmfulness and unexpectedness ratings. Previous work suggests that both relationships should be positive. To address the second goal, I fit cross-classified, multi-level models predicting immorality ratings on the event’s semantic components – like actor’s evaluation or object’s potency. These models include varying intercepts for individuals and scenarios in order to account for systematic differences between events and respondents. The results will give me a principled basis to reach a formal definition of a prototypical moral transgression.

### Results

Figure 1 shows the relationship between the average immorality of each scenario and its average harmfulness and average unexpectedness. The results further reaffirm the close relationship between immorality and harm. Moreover, the positive relationship between immorality and unexpectedness echoes the idea that severe transgressions tend to be perceived as more unexpected because they often entail the breaching of social norms. For the purposes of this study, the most important point is that the translated scenarios are interpreted consistently. There are encouraging patterns that are consistent with findings in the literature: the “harm” transgressions tend to be rated as the most severe, while the “purity” violations rank amongst the most unexpected. This suggests that there was not a significant loss of information when the vignettes were translated into the new format.

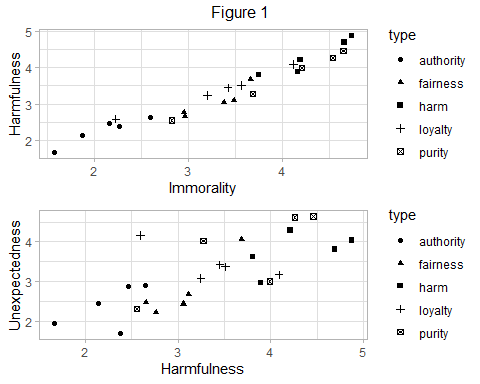
In order to select the most informative model, I use both AIC and BIC, and the results are available on the supplemental materials. Across the board, the most informative model is a cross-classified model that includes all semantic features but that foregoes additional covariates, such as the gender or political ideology of the respondent. Figure 2 shows the coefficient plots for two models. In the second model, I exclude the so-called “purity” scenarios from the analysis, following previous literature that has argued that these transgressions are distinctive, either because they a fundamentally different type of violation or simply because they are unconventional. Both models suggest that there are three semantic features that are particularly informative when considering the immorality of an event: the behavior’s evaluation, the behavior’s potency, and the object’s potency. While only the coefficient for behavior potency reaches conventional levels of statistical significance in the first model, all three coefficients become significant in the model that excludes the purity transgressions. This reaffirms the idea that these violations might indeed have idiosyncratic semantic structures, though – as it will be discussed below – it does not necessarily imply that they belong to a distinct category of moral wrongs.

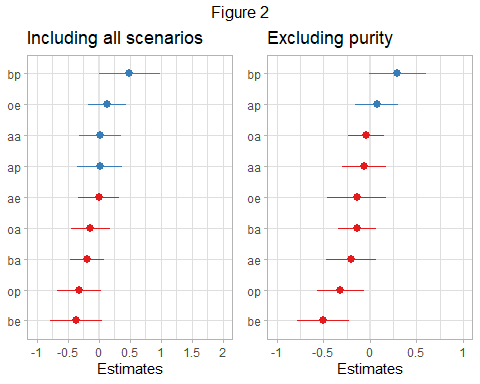
At this point, nonetheless, conventional statistical significance is not considerably important because the analysis remains mainly exploratory. I am interested in the features that are most informative for predicting an event’s immorality. In this respect, both models agree: one should look at the behavior’s evaluation, the behavior’s potency, and the object’s potency. Reassuringly, these results are consistent with Gray and Schein’s dyadic theory of morality. Negative and potent behaviors are perceived as more immoral. In turn, as the object’s potency increases, the immorality of an acts tends to decrease. This offers a more concrete definition of Gray and Schein’s (2018) notion of a “vulnerable patient”. As the recipient of an action appears less potent, said action is considered more immoral.

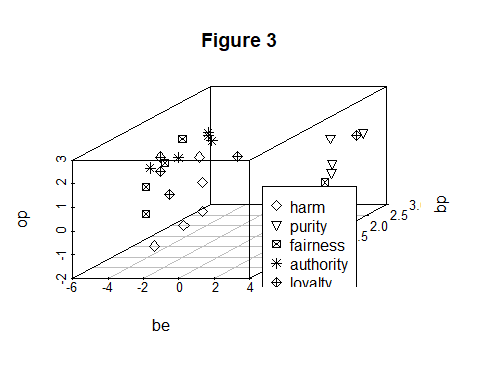
Furthermore, establishing the importance of these three features allows me to place events in a three-dimensional space in order to compare them. Figure 3 shows this space. This is key for two main reason. First, it helps me reach a principled definition of a prototypical moral transgression: it should involve a potent and negative event directed at a very weak object. Using the terms that I have tested, the most appropriate event I can formulate is:

A person kills a child.

The second reason why building such a three-dimensional space is important is because it allows me to calculate distance between the scenarios and the template defined above. This is the key piece of information that helps me test whether template matching mediates the attribution of immorality.







## Study 2

### Data and Methods

The data for this study was also collected through the Prolific platform. I reached out to a total of 200 participants and, after controlling for missed attention checks, there were a total 184 usable responses. In this study, I asked participants to categorize the same 25 scenarios as either immoral or not immoral, harmful or harmless. For each question, the scenario appeared in the middle of the screen and the dichotomous categories were shown at the bottom, on opposite sides of the page. Using their keyboards, the participants chose the category to which they think the event belongs. The outcome variable of interest is the time it takes respondents to categorize the violations. Reaction time here is meant to stand as a proxy for cognitive effort, with higher times reflecting that the categorization task is more difficult (Gray and Keeney 2015a; Moore 2017).

Each participant had to classify all scenarios on two dimensions and, therefore, I have a total of 9200 recorded times. To ensure that I could capture the initial reactions as reliably as possible, I designed the questionnaire so that participants had a ten-second limit to classify each scenario (with a one-second pause between questions). Out of all the responses, only 44 were above 10 seconds and 22 were below 1 second. Given that these data most likely reflect misclicks or technical difficulties, I used median imputation to replace these values. I ran the same analyses after deleting these observations with no appreciable differences in the results.

The main goal of this analysis is to examine the relationship between distance from the prototypical moral transgression – a person kills a child – and the time it takes individuals to categorize the scenarios. I define distance from the template as the Euclidean distance between that point and a particular event in the three-dimensional space presented in Figure 3. Thus, if the prototypical transgression is , then distance for scenario will be defined as:

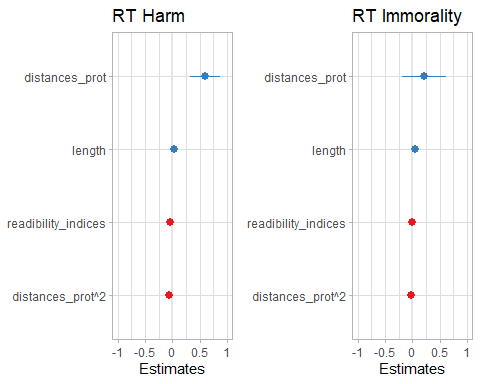
Finding that this distance is informative of reaction time would constitute novel evidence supporting the idea that moral cognition is underpinned by template matching. In order to test this proposition, I fit cross-classified, multi-level models where the main independent variable is distance from the prototype and the outcome variable is reaction time. The models include varying intercepts for both respondents and scenarios. They also include controls for length and readability. Reaction time data can be fickle and the length, and complexity of scenarios certainly play a role in how long it takes respondents to categorize them. The models, then, include a term for the number of characters in each event and another coefficient with their respective Flesch-Kincaid readability score – an established measure of the level of education necessary to understand a text.

### Results

The most informative model, in this case, is the one that depicts a quadratic relationship between distance from the prototype and reaction time. Although the Information Criteria values are close, this specification is narrowly better. This relationship appears to follow similar patters for the classification of immorality and harmfulness, but the coefficients only reach conventional levels of significance in the latter. This is certainly an important limitation but given the close interconnection between immorality and harm, the results are still insightful for understanding template matching in the context of moral cognition. Figure 4 depicts the quadratic relationship: events that are very close and very far from the prototype take little time to classify. In other words, it is easy to see why prototypical events are harmful, while situations that are very far from that point – such as kissing one’s spouse – are quickly perceived as harmless. The results indicate that distance from a mental template might be informative for understanding moral cognition and that this relationship should be considered non-linearly.

The model suggests that the scenarios that are most difficult to categorize are those that lie at a middling distance from the prototypical immoral act. This relationship between ambiguity and reaction time is also found in other parts of the data. For example, Figure 5 shows the predictions of a cross-classified, multi-level model that regresses reaction time of harmfulness on the proportion of participants who classified that event as harmful. I provide a full coefficient table in the supplemental materials, but the most important point is that the relationship is parabolic: the events that had the most variance in how they were categorized took the longest to be classified. This finding furthers validates the decision to model the relationship between distance from the prototype and reaction as quadratic.

A concern with these results is that most of the scenarios that are very distant from the prototype are the so-called purity scenarios. Study 1 shows that these scenarios are consistently ranked amongst the most harmful. Thus, it is possible that these transgressions are driving the downward turn of the quadratic effect. Nonetheless, I ran the same models taking these scenarios out of the sample and the quadratic relationship still held; these results are available in the supplemental materials. More interestingly, however, it is possible to fit a logistic regression predicting whether an event was classified as harmful using distance from the prototype as the main dependent variable. Figure 6 shows the coefficients of this multi-level, cross-classified model. Increases in distance from the prototype lead to considerably lower odds of an event being classified as harmful. Again, let me re-emphasize that the so-called purity scenarios are both quite distant from the prototype and yet are very harmful. Thus, the relationship evidenced in Figure 4 could be said to occur not because of these scenarios but rather in spite of them. As it will be discussed below, this fact carries important implications.



## Discussion

At a general level, this study provides a novel technique for analyzing moral transgressions. As mentioned above, the majority of work on moral cognition has used hypothetical scenarios to elicit moral judgments from participants (Clifford 2017). In this paper, I show that we gain valuable additional information by going below the level of denotative meanings and analyzing the underlying semantic structures of these vignettes. What makes it possible to access these connotative meanings is the existence of large repositories of empirically validated semantic profiles collected by ACT scholars. We can use these dictionaries to translate fictional vignettes into formats that allow us to explore their semantic structures. Doing so yields insightful findings. For instance, Study 1 shows that, echoing Gray and Schein’s (2018) arguments, transgressions have shared regularities: they involve a negative and potent behavior directed towards a weak object. Recall that, according to these authors, the template of a moral wrong involves an intentional actor directing a damaging behavior towards a vulnerable patient. In this study, I am able to bring rigor to this definition: a damaging behavior could be considered an action which is low on evaluation but high on potency; a vulnerable patient could be described as an actor who has a low potency score.

By establishing these regularities, I am able to outline a formal definition of what the template of an immoral act looks like and to calculate deviations from said template. Finding important semantic dimensions for the attribution of immorality helps me place the scenarios on a Cartesian plane defined by those dimensions. By “maximizing” the values on these dimensions, I can represent the prototypical moral wrong as a point in this space and, therefore, I can calculate how distant or close the scenarios are from this template. This allows me to directly test the proposition that moral cognition occurs through template matching. The results from Study 2 show that distance from the template is indeed informative of how long it takes individuals to classify events as harmful. Interestingly, these findings also suggest how this relationship should be understood, a point that Gray and Schein (2018) overlook. The relationship is quadratic: as distance increases reaction time increases, but only up to a point. Individuals are quick to classify events that are very close or far apart from the template. Events that lie at a middling distance are more ambiguous and take longer to be perceived as harmful. These findings, then, provide empirical evidence for the notion that template matching mediates moral cognition, but also give a clearer idea of how this process occurs.

These studies also shed light on one of the most central debates within the study of moral cognition. A great deal of research within this line of work has revolved around the question of whether there are such things as harmless wrongs (Tepe and Aydinli-Karakulak 2019). These scenarios tend to be associated with the “purity” foundation: they often involve sexual or dietary taboos. As noted above, scholars have recently argued that these acts are still perceived through the lens of harm. Indeed, the results of Study 1 echo this idea: the so-called purity transgressions ranked amongst the most harmful. Yet, the findings of Study 2 compel us to rethink the distinctiveness of these scenarios. They are quite far away from the prototype and, appropriately, they tend to be classified fairly quickly. However, the classification goes contrary to what the model would predict: they tend to be considered harmful. These violations, then, break the patterns that our models uncover in the data. Both sides of the debate, then, seem to have pointed out important aspects of this type of transgressions. Gray and Schein (2018) are right in insisting that that purity scenarios are understood through the lens of harm. However, these transgressions are distinctive insofar as they considered very harmful and yet are quite far away from the prototypical moral wrong. The challenge, then, is to explain this incongruence. Our findings suggest that we should move away from the debate about “harmless” wrongs, but it compels us to consider, then, how practices that are distant from the prototype come to be regarded as harmful. One alternative, for instance, is that there are different processes of “harmification” (Schein and Gray 2016). While understanding the harmfulness of events that are very close to the prototype might be easy, doing the same for practices that are far away might be more cognitively strenuous. Therefore, it might be necessary for groups to build additional moral scaffolding – such as taboos and mechanisms of public shame – in order to bring events that are very far away from the template into the realm of harm. These are only conjectures, but they serve to highlight an important point: our findings suggest that what is distinctive about purity scenarios is not that they are harmless, but rather that their semantic structure does not match their perceived harmfulness. Research on moral cognition should seek to understand, then, how these violations come to be perceived through the lens of harm.

## Limitations and Future Directions

An important limitation of this study concerns the number and type of fictional vignettes used. The key independent variables in these studies are related to the semantic structures of the events. Therefore, given that I only use 25 vignettes, I rely on limited variation to produce my estimates. This might be why some of the coefficients have large confidence intervals. Furthermore, the events are all practices that are relatively immoral. This might sound obvious but effectively what I am doing here is a form of selecting on the dependent variable. I am making an argument about a general process of template matching using only a particular kind of practice. Now, this is a deliberate choice: this helps root this study within wider research on moral cognition and shows that there are different ways of analyzing these vignettes. This analysis, then, should be regarded as a first step. Future work should try to systematically vary the connotative attributes of the scenarios, thereby exploring the semantic space more rigorously.

Another potential shortcoming of this analysis is related to the sensitivity of one of the instruments. Reaction time studies are fickle to begin with and here I am adding complexity to this task. I am asking participants to react to full sentences and to do so remotely, through a crowdsourcing website. Though the data show consistent and encouraging patters, these limitations should be taken into account when drawing conclusions from these studies. In future work, it would be useful to find ways of measuring moral cognition more precisely. It could be possible, for instance, to use eye-tracking measures or to adapt novel techniques – such as Affect Misattribution Procedures – to the analysis of moral cognition.

## Conclusion

These studies represent, to my knowledge, the first concrete attempt to test one of the most plausible contemporary theories of how moral cognition operates. Current evidence supports a more unified vision of moral judgement, where perceptions of harm are the main drivers of this process. Gray and Schein have outlined a model of moral cognition that is consistent with this vision; one where immorality is ascribed through template matching. The evidence to support this position, however, remains limited. Here, I device a technique for directly examining their conceptualization. I leverage existent repositories of cultural meanings to look at the underlying semantic structures of moral transgressions. This, in turn, allows me to accomplish two things: to produce a rigorous formal definition of a prototypical moral wrong, and to examine whether deviations from that template affect how individuals perceive transgressions. The results suggest that distance from the prototypical moral trangression is indeed informative of the time it takes individuals to classify scenarios as harmful. This evidence adds plausibility to the idea that the attribution of immorality happens through template matching.

Furthermore, these studies shed light on the ongoing debates about the existence of “harmless” wrongs. They show that even the so-called purity trangressions are understood through the lens of harm and that events that are more distant from the prototypical violation are less likely to be classified as harmful. However, these analyses also suggest that these transgressions are indeed distinctive: they are quite distant from the template and yet they are considered very harmful. The question, then, should not be how harmless wrongs become immoral but rather how events that are far away from prototypical moral trangressions are “harmified”.

This paper should be regarded as an initial foray into a concrete empirical examination of the role that template matching plays in the attribution of immorality. As noted above, there are important limitations that compel me to be cautious about the claims that can be advanced with my evidence. The variation between the semantic structures of the events is limited and reaction time data can be unreliable. Nonetheless, I have taken the necessary steps to make this analysis as riguruous as possible. Thus, despite the limitations, this study represents a novel attempt to provide empirical grounding to a theory that, though popular, remains largely untested. By analyzing the semantic structures of moral trangressions, I present evidence that suggests that the attribution of immorality occurs through template matching. Future work should continue to analyze connotative regularities amongst moral violations in order to further our understanding of how moral cognition operates and how it varies systematically across groups.

### References:

Clifford, Scott. 2017. “Individual Differences in Group Loyalty Predict Partisan Strength.” *Political Behavior* 39 (3): 531–52. <https://doi.org/10.1007/s11109-016-9367-3>.

Clifford, Scott, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. “Moral Foundations Vignettes: A Standardized Stimulus Database of Scenarios Based on Moral Foundations Theory.” *Behavior Research Methods* 47 (4): 1178–98.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. “Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism.” In *Advances in Experimental Social Psychology*, 47:55–130. Elsevier.

Gray, Kurt, and Jonathan E. Keeney. 2015a. “Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality.” *Social Psychological and Personality Science* 6 (8): 859–68.

———. 2015b. “Disconfirming Moral Foundations Theory on Its Own Terms: Reply to Graham (2015).” *Social Psychological and Personality Science* 6 (8): 874–77. <https://doi.org/10.1177/1948550615592243>.

Gray, Kurt, Liane Young, and Adam Waytz. 2012. “Mind Perception Is the Essence of Morality.” *Psychological Inquiry* 23 (2): 101–24. <https://doi.org/10.1080/1047840X.2012.651387>.

Greene, Joshua D. 2014. “Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics.” *Ethics* 124 (4): 695–726. <https://doi.org/10.1086/675875>.

———. 2015. “The Rise of Moral Cognition.” *Cognition* 135 (February): 39–42. <https://doi.org/10.1016/j.cognition.2014.11.018>.

Haidt, Jonathan, and Craig Joseph. 2004. “Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues.” *Daedalus* 133 (4): 55–66. <https://doi.org/10.1162/0011526042365555>.

———. 2007. “The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules.” *The Innate Mind* 3: 367–91.

Haidt, Jonathan, Silvia Helena Roller, and Maria G Dias. n.d. “Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?” 16.

Harnad, Stevan. 2017. “To Cognize Is to Categorize: Cognition Is Categorization.” In *Handbook of Categorization in Cognitive Science*, 21–54. Elsevier.

Heise, David R. 2007. *Expressive Order: Confirming Sentiments in Social Actions*. Springer Science & Business Media.

Heise, David R. 2014. “Cultural Variations in Sentiments.” *SpringerPlus* 3 (1). <https://doi.org/10.1186/2193-1801-3-170>.

Hitlin, Steven, and Stephen Vaisey. 2013. “The New Sociology of Morality.” *Annual Review of Sociology* 39 (1): 51–68. <https://doi.org/10.1146/annurev-soc-071312-145628>.

Hunzaker, M. B. Fallin. 2016. “Cultural Sentiments and Schema-Consistency Bias in Information Transmission.” *American Sociological Review* 81 (6): 1223–50. <https://doi.org/10.1177/0003122416671742>.

Hunzaker, MB Fallin, and Lauren Valentino. 2019. “Mapping Cultural Schemas: From Theory to Method.” *American Sociological Review* 84 (5): 950–81.

Martin, John Levi, and Matthew Desmond. 2010. “Political Position and Social Knowledge.” *Sociological Forum* 25 (1): 1–26. <https://doi.org/10.1111/j.1573-7861.2009.01154.x>.

Monk, Ellis P. 2019. “The Color of Punishment: African Americans, Skin Tone, and the Criminal Justice System.” *Ethnic and Racial Studies* 42 (10): 1593–1612.

Moore, Rick. 2017. “Fast or Slow: Sociological Implications of Measuring Dual-Process Cognition.” *Sociological Science* 4: 196–223. <https://doi.org/10.15195/v4.a9>.

Osgood, Charles E. 1969. “On the Whys and Wherefores of E, P, and A.” *Journal of Personality and Social Psychology* 12 (3): 194–99. <https://doi.org/10.1037/h0027715>.

Osgood, Charles Egerton, William H. May, Murray Samuel Miron, and Murray S. Miron. 1975. *Cross-Cultural Universals of Affective Meaning*. Vol. 1. University of Illinois Press.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. “Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research.” *Journal of Experimental Social Psychology* 70 (May): 153–63. <https://doi.org/10.1016/j.jesp.2017.01.006>.

Rosch, Eleanor. 1999. “Principles of Categorization.” *Concepts: Core Readings* 189.

Schein, Chelsea, and Kurt Gray. 2015. “The Unifying Moral Dyad: Liberals and Conservatives Share the Same Harm-Based Moral Template.” *Personality and Social Psychology Bulletin* 41 (8): 1147–63. <https://doi.org/10.1177/0146167215591501>.

———. 2016. “Moralization and Harmification: The Dyadic Loop Explains How the Innocuous Becomes Harmful and Wrong.” *Psychological Inquiry* 27 (1): 62–65.

———. 2018. “The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm.” *Personality and Social Psychology Review* 22 (1): 32–70. <https://doi.org/10.1177/1088868317698288>.

Tepe, Beyza, and Arzu Aydinli-Karakulak. 2019. “Beyond Harmfulness and Impurity: Moral Wrongness as a Violation of Relational Motivations.” *Journal of Personality and Social Psychology* 117 (2): 310–37. <https://doi.org/10.1037/pspi0000169>.