The Meanings of Moral Wrongs

Nicolas Restrepo

Introduction

Individuals and groups must constantly classify and categorize moral transgressions; they produce legal and moral codes that group certain types of violations together and, more importantly, that sort these categories according to their perceived severity. There is widespread debate regarding the processes that underpin the categorization of moral wrongdoing. Currently, the most plausible account contends that we attribute moral wrongdoing through prototypical association. In other words, we compare violations to a mental template of what a typical trangression looks like, and immorality is ascribed to the extent that the perceived act resembles that template. Here, I contend that by looking at the connotative meanings associated with different moral vignettes, we can get a clearer idea of what this prototypical moral transgression looks like. Having established the features that identify exemplary moral trangressions, we can calculate how closely certain events resemble these prototypical violations. This gives us a rare opportunity: it allows us to rigurously test whether proximity to a mental exemplar mediates the attribution of immorality.

Background

Attribution of Immorality

Research around moral decision-making has changed throughout the last few decades, shifting emphasis from issues of logical reasoning to questions about deep-seated cognitive processes. The discussion about how people decide what is right and what is wrong had mostly focused on the rational calculations that actors face when presented with a conundrum. The trolley problem has become the main example of this type of research. At heart, this work deals with whether people make moral decisions based on non-negotiable ethical principles or whether they focus on the consequences of their actions. Recently, however, researchers have pointed out that the bulk of our moral decision-making does not look like a trolley problem at all. In fact, those dilemmas are so interesting precisely because how dissimilar they are to our everyday experiences. When we encounter a potentially immoral event, we tend to intuitively know what is and is not permissible. We attribute immorality, not through chains of logical reasoning, but rather through automatic gut feelings;

some events just appear morally dubious even if we are not quite capable of articulating why. Research on moral attribution, then, has moved to analyzing the underlying cognitive processes that permit these automatic gut reactions; scholars have started to explore the cognitive mechanisms that make us such swift arbiters of right and wrong. From this line work, two theories have arisen as the main contenders: one posits that moral wrongs belong to different categories and that humans are adept at picking up these differences; the other puts forward a more unified conception of moral cognition.

The most prominent theory that advances a view of morality as consisting of differentiated domains is known as moral foundations theory (henceforth MFT). This framework emerges as a response to an empirical puzzle: the fact that there are events that we find impermissible, but which do not involve harm or injustice. A typical example is that of two siblings having consensual sex with no consequences. Although no entity is being harmed or wronged, most people still view this scenario as immoral. To explain this gap, MFT scholars argue that the attribution of immorality responds to different underlying logics. In other words, murder and incest are both moral transgressions but they are fundamentally different types of misdeeds and, thus, we assess their wrongness based on different criteria. While the former is judged on the basis of harm, the latter is understood through the lens of what MFT scholars call "purity". These two codes – along with concerns for care, loyalty, and authority – make up the foundations upon which morality is constructed. MFT, then, puts forward a pluralist vision of moral decision-making, which seeks to explain the empirical breadth and diversity of moral prohibitions.

There is an underlying argument here about how moral cognition operates. MFT presumes that discrete categorization is the main process through which individuals attribute immorality. Each foundation has certain stimuli and emotions related to it. We have mental faculties that are adept at identifying those stimuli and associating them with the right category. Once we know what "foundation" is being violated, we can produce appropriate moral judgements. For example, when we see an athlete cheating, we identify the event as a fairness violation and feel outraged. When confronted with incest, we correctly categorize it as a purity violation and tend to experience disgust. The bottom line is that, according to this theory, attributing immorality entails correctly identifying what set of moral logics an event appears to breach and then reacting appropriately.

This position has been recently criticized by scholars who are pushing for a more integrated conception of morality. Recent empirical evidence shows that the five moral foundations are all highly correlated. Different moral prohibitions, then, seem to be cut from the same cloth, and cannot be neatly parceled out into distinct groups. According to Gray and Schein, the unifying dimension that underpins moral decision-making is harm. We judge actions to be immoral insofar as we can perceive them as harmful. Interestingly, the language of

harm even creeps into our understanding of moral breaches that do not seem to directly affect any entities. Thus, incestual marriages are often described as harming the health – the gene pool - of the group and breaching religious dietary taboos is often depicted as harming one's soul or that of the ancestors. The argument of harm being the central engine for moral decision-making is certainly not new, but it has accrued significant empirical validation in the last decade. The evidence suggests that perceptions of harm shape how we determine immorality and how we assess severity. Rather than being organized around different logics, then, morality seems to be underpinned primarily by perceptions of harmfulness.

This position, in turn, implies a more continuous vision of moral cognition. Transgressions share an underlying structure: they involve a cognizant agent directing a harmful behavior towards a vulnerable patient. This framework represents a cognitive template against which we compare events to gauge how immoral they are. Situations that closely resemble this template are quickly flagged as immoral. The farther away an event is from this prototypical structure, however, the harder it will be for us to understand it as immoral. This explanation relies heavily on the idea that category membership is ascribed, not through a set of exclusionary criteria, but rather due to proximity to a salient exemplar. Moral cognition, then, does not entail sorting an event into discrete categories to assess what kind of moral code it breaches. Instead, it entails placing a situation on a continuum, more or less distant from our mental template of a moral wrong. There is empirical evidence that supports this argument: individuals tend to label actions as immoral more quickly if they also think they are harmful. Thus, the most cognitively plausible account of the attribution of immorality suggests that we classify moral violations based on their proximity to a fuzzy mental template – one that depicts an intentional agent causing harm to a vulnerable victim.

Though convincing, this account still leaves crucial questions unanswered. Although proponents argue persuasively for the existence of a cognitive template of moral wrongdoing, the characteristics of this exemplar remain opaque. For instance, it is unclear what makes an action seem prototypically harmful or a victim obviously vulnerable. Thus, we do not know what the prototypical moral transgression looks like, let alone how to go about calculating whether actions are distant from it. Relatedly, it is key to further specify exactly what role distance from the prototype plays in the attribution of immorality. We know that proximity to the template is not necessarily related to the severity of transgressions; there are violations – such as incest – that, while distant from the prototypical moral dyad, are considered very harmful. Therefore, distance from the prototype should not necessarily be related to perceived severity. What this distance should predict, then, is the cognitive difficulty that it takes individuals to reach their assessment. In other words, proximity to the exemplary moral wrong should not necessarily reflect how severe we think actions are but rather how easily we can categorize them as immoral. In this paper, I aim to address these two gaps. Drawing from sociological

work, I seek to clarify what a prototypical moral transgression looks like and to show that distance from it affects how easily individuals can categorize events as immoral.

The Social Meanings of Moral Transgressions

A good first step towards providing a clearer description of what a prototypical moral wrong might look like is to recognize that mental templates are eminently social. The cultural landscapes in which we are embedded shape how we envision categories, and the exemplars around which they are organized. In other words, our readiness to say who counts as African American (Monk, 2014) or as poor (Valentino and Hunzaker, 2019) changes depending on the social positions in which we are situated. This is a crucial point for this analysis: it means that we can examine the cultural meanings that are associated with different transgressions to begin to piece together an outline of a prototypical moral wrong. The central challenge is to be able is to map a rigorous measure of cultural meaning onto moral transgressions. Luckily, there are sociological traditions that have collected large repositories of semantic structures that we can use to accomplish this. In this analysis, I will be using the dictionaries of affective meanings collected by Affect Control Theory (henceforth ACT) scholars.

Drawing from its symbolic interactionist roots, ACT starts from the premise that individuals attach affective meanings to social concepts and that those semantic structures can be reduced to a set of measurable dimensions. Building on Osgood et al.'s (1957) work, it contends that there are three basic dimensions of meaning: evaluation, potency, and activity. Evaluation pertains to characteristics such as goodness and badness. Potency, in turn, captures notions of power and weakness, while Activity relates to issues of liveliness and quietness. Osgood et al.'s (1975) extensive research demonstrates the cross-cultural validity of these dimensions of affective meaning, and their utility has been broadly recognized.

Using these dimensions, it is possible to collect rigorous measures of the affective meanings that individuals attach to particular concepts. The measuring technique used is called semantic differentials: respondents are asked to rate a series of concepts on a scale ranging from, for example, very weak to very strong. The averages of all three dimensions then are computed (Heise, 1979) and each concept is assigned an EPA profile. This means that it is possible to place identities and behaviors in a three-dimensional space and to compare their underlying semantic structures; babies, for instance, are described as very good, very weak, and somewhat lively, while murderers are depicted as very bad, very powerful, and slightly active. Importantly, these values have proved to be both remarkably accurate and generalizable. Here, then, we have a framework that makes meaning measurable and formalizable. By locating social concepts in a three-dimensional semantic space, we

can shed light on the symbolic structures through which cultures organize the social world.

We can leverage these large repositories of affective meanings to explore the underlying semantic structures of moral transgressions. Research about moral cognition has primarily used fictional scenarios to examine how respondents assess immorality. These fictional scenarios tend to share a grammatical structure (actor-behavior-object) which is, incidentally, the main unit of analysis of ACT. This means that we can take advantage of their dictionaries of affective meanings to translate moral scenarios into a format that allows us to examine their underlying semantic structure. For example, a commonly used fictional scenario is:

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

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

> Teacher [2.5; 2.31; 0.32] hits [-2.66; 1.30; 2.12] lazy student [-0.42; -0.76; -1.56]

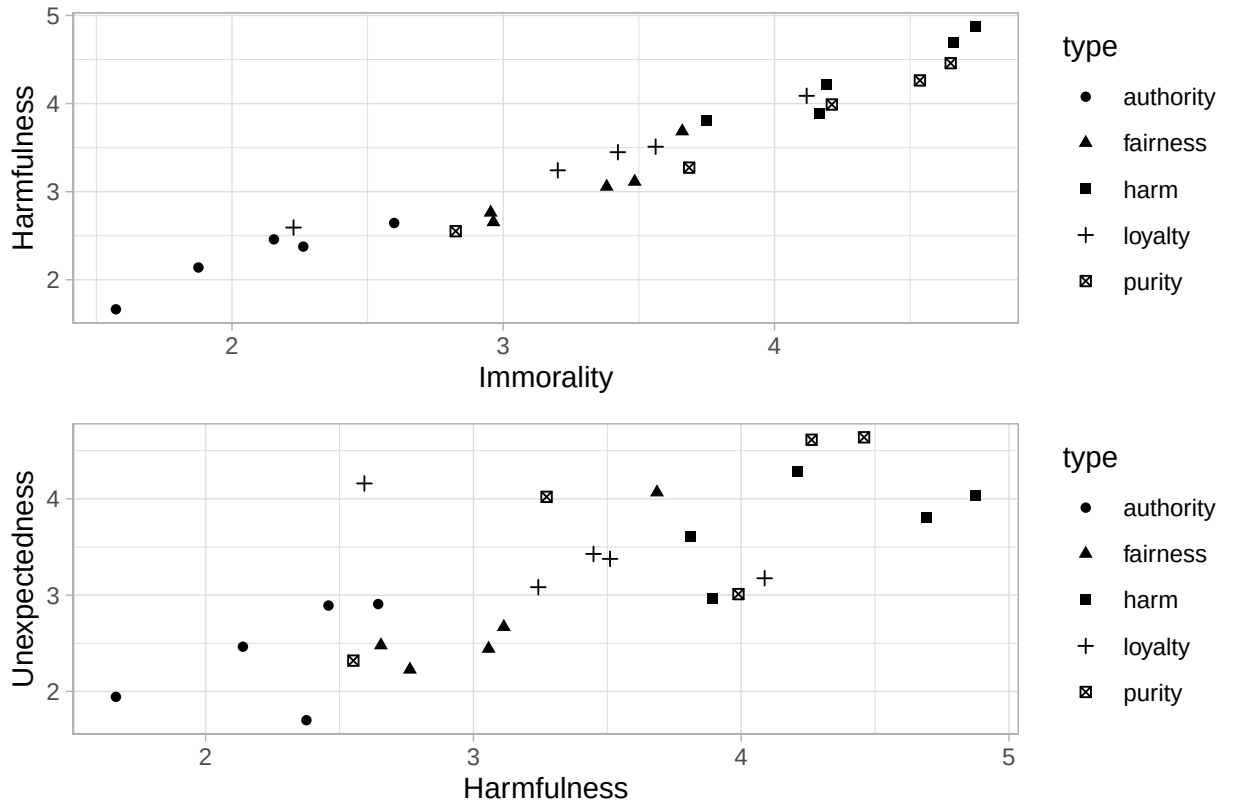
These formal translations, in turn, help us explore moral transgressions more systematically. Specifically, we can examine the shared and recurrent semantic features of moral violations. In other words, we can analyze the affective meanings associated with prototypically harmful acts such as "murder" or with characteristically vulnerable patients such as "children". Furthermore, we can measure the differences and overlaps between the semantic structures of transgressions, examining what makes violations more or less prototypical. As a result, we would be able to quantify deviations from prototypicality, which would allow us to shed light on whether this distance is indeed informative for understanding the attribution of immorality. This is precisely what the following two studies seek to accomplish.

Study 1

In the first study, I aim to validate whether respondents interpret the translated scenarios consistently and to examine what are the semantic features that have the biggest impact in their opinions about the immorality violations. For this, I translated 25 of the most common vignettes in this literature into the EPA semantic space. Using the Prolific platform, I asked participants (n = 205) about the extent to which they considered each scenario harmful, immoral, and unexpected. After filtering for missed attention checks, there is a total of 194 usable responses. Figure 1 shows the relationship between the average immorality of each scenario and its average harmfulness and unexpectedness. The results resonate with previous empirical evidence: immorality is positively related to both harmfulness and unexpectedness. The first result further reaffirms the relationship between harm and immorality. The latter, in turn, lends credence to the notion that severe transgressions tend to be perceived as more unexpected because they entail the breaching of social norms.

For the purposes of the study, the most important point is that the translated scenarios are being interpreted consistently; this means that there was not a significant loss of information when the vignettes were translated into the new format.

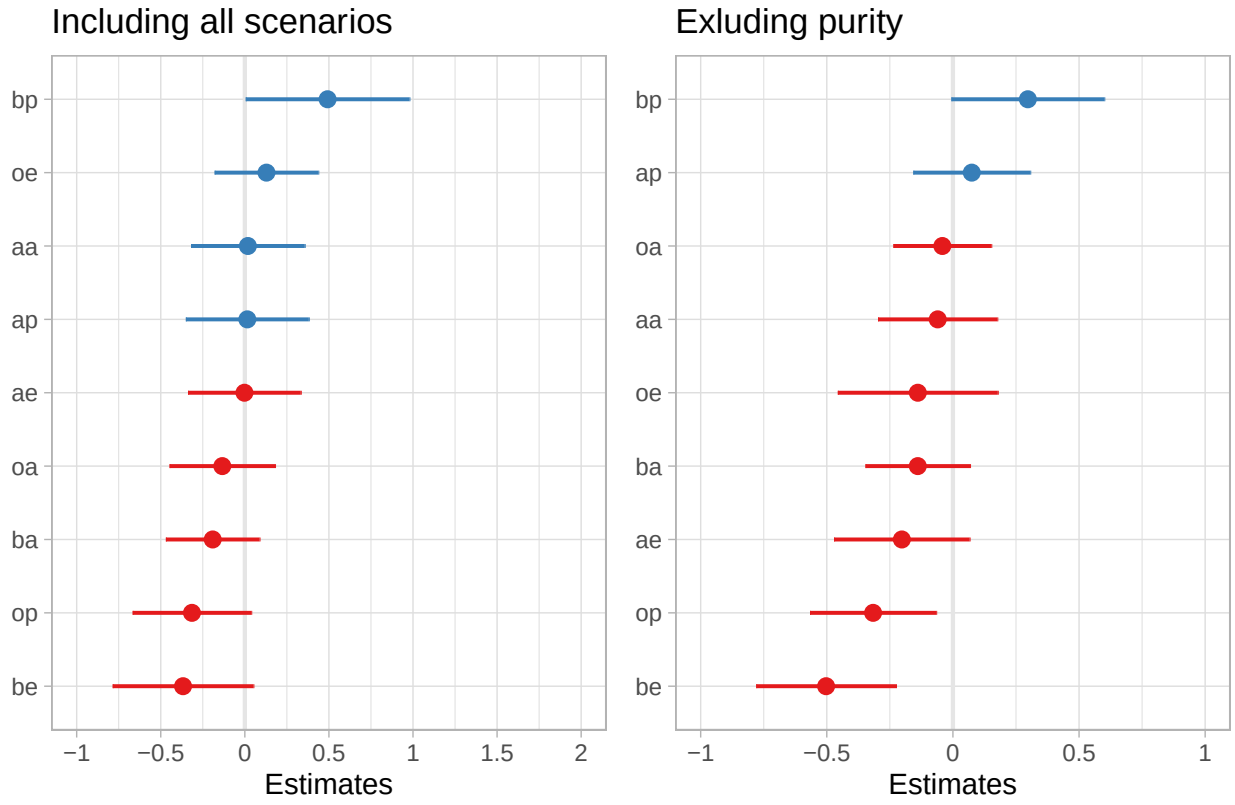
Figure 1



In order to examine how the underlying semantic structures of violations affect perceptions of immorality, I conduct statistical analyses with immorality ratings as the dependent variable and the EPA values of each component as the main independent variables. Here, I build cross-classified models, giving each respondent and each scenario its own intercept. In order to select the most appropriate model for the current analysis, I use different measures of information criteria. Across the board, the most informative model for the current data seems to be a cross-classified model that includes the EPA values but that foregoes additional covariates, such as gender and political ideology. Figure 2 shows the coefficient plots for two models; the second one excludes the so-called 'purity' scenarios from the analysis. Both models show that there are three semantic features that seem particularly informative when trying to understand the attribution of immorality: the behavior's evaluation, the behavior's potency, and the object's potency. These results resonate with the theory of dyadic morality. Negative and potent behaviors are perceived as more immoral. In turn, as the object's potency increases, the immorality of the act tends to decrease. This finding speaks directly theory's idea of vulnerability. As actors appear less potent, the behaviors directed at them are considered more immoral.

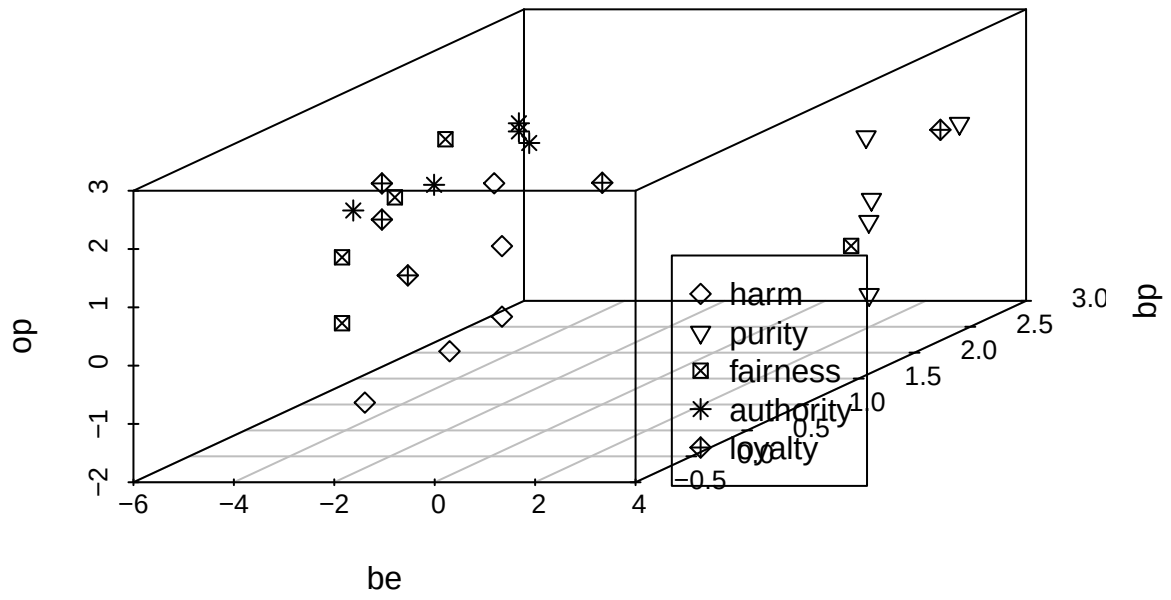
This adds plausibility to the notion that the vulnerability of the victim is key when assessing immorality.

Figure 2



Though informative, the coefficients in the first model do not reach conventional levels of significance. However, in the model that excludes the 'purity' transgressions, the coefficients for the object's potency and the behavior evaluation become significant. This shows that there might be something about the excluded violations that does not quite fit with the semantic patterns evidenced in the rest of the scenarios. When we plot the scenarios in three-dimensional space, this difference becomes evident. In Figure 3, the axes are the three semantic features we identify as particularly important. We notice that the so-called 'purity' transgressions are grouped away from most of the other scenarios. They do not seem to share the same underlying semantic structure as other scenarios, but they are amongst those rated as most immoral and most harmful. We need to think, then, about what distance in this semantic space means. Study 2 seeks to answer this question.

Figure 3



Study 2

In this study, I ask respondents (n = 92) to tell me whether they categorize the 25 scenarios as immoral or not immoral, harmful or harmless. In each question, the scenario appears in the middle of the screen and the dichotomous categories are shown at the bottom, on opposite sides of the page. Using their keyboards, the participants indicate the category to which the event belongs. The variable of interest, here, is how long it takes individuals to categorize the violations. By using this method, I root this analysis in a longstanding tradition of studies that have implemented reaction time data to explore issues surrounding social cognition.

This analysis seeks to test whether the attribution of immorality does occur through the association of the situation at hand with a prototypical mental template. As discussed above, this theory does not necessarily predict that proximity to the exemplar necessarily leads to higher levels of perceived immorality. Our data confirm this: while violations involving sexual taboos are quite distinct from the rest of the transgressions, some of them rank amongst the most severe. Thus, distance form a prototypical transgression should not predict severity. If the theory is correct, however, it should predict how cognitively difficult it is to attribute immorality. The further away an event is from our mental template, the harder it should be for us to

understand it as immoral. This is the idea I aim to test in this study. I aim to analyze whether distance from a prototypical moral wrong predicts how long it takes respondents to categorize a scenario as immoral.

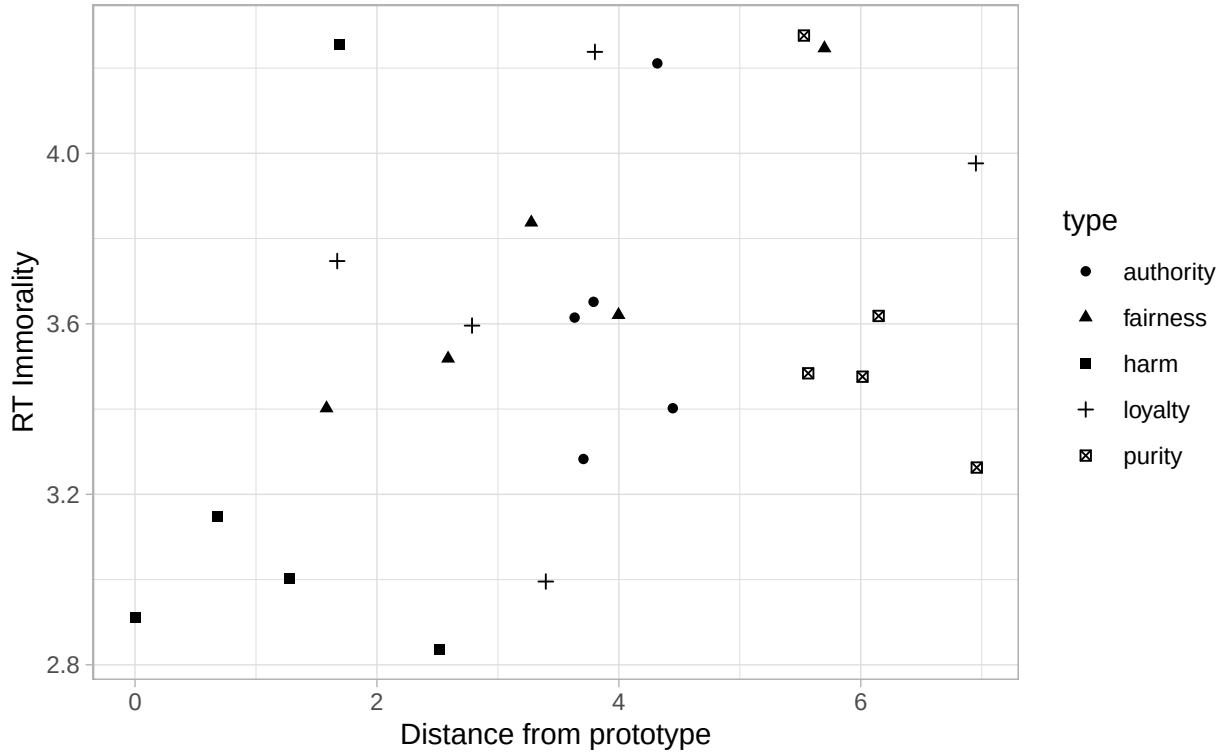
To undertake this analysis, I need to choose which one of the translated scenarios to use as the exemplary transgression. The data from the previous analysis helps inform this decision: a prototypical moral violation should involve a potent and bad behavior directed towards a weak victim. From the available vignettes, the one that best reflects this semantic structure is: "a person hurts a child". This scenario will be used as the prototype – the reference point against which semantic distance will be calculated. Distance from the prototypical transgression will be defined as Euclidean distance in the three-dimensional space presented in Figure 3. If the prototypical moral transgression is p, then distance for scenario i will be defines as:

$$D_p = \sqrt{(BP_p - BP_i)^2 + (BE_p - BE_i)^2 + (OP_p - OP_i)^2}$$

Figure 4 shows the relationship between distance from the prototype and reaction time. We notice a slight positive relationship, which resonates with the theory that is being tested. Note, however, that reaction time seems to be highly correlated with the length of the vignette. Scenarios like "a married person has sex with an adulterer" and "an employee conspires with a competitor" have the highest reaction times. Reaction type data is quite sensitive to the length of the prompts that respondents have to categorize and, therefore, in the analyses I need to make sure I account for this factor.

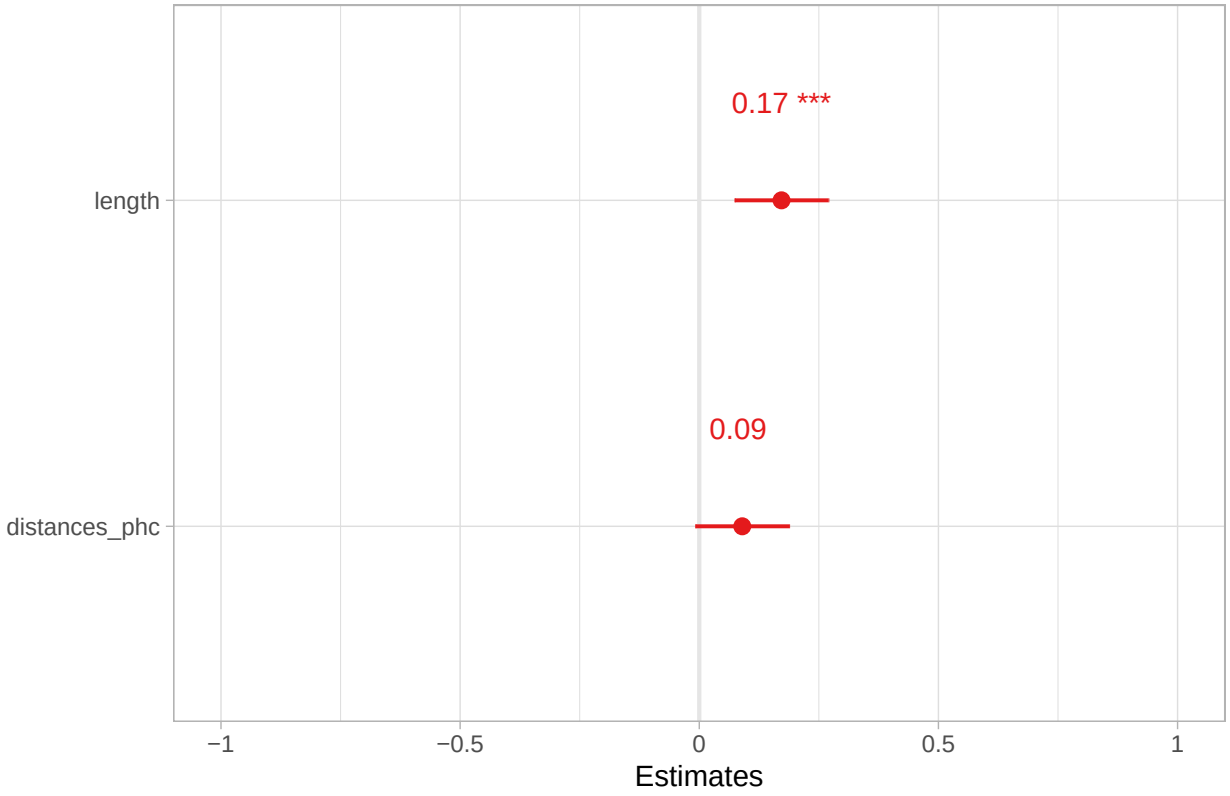
Figure 4

Reaction time by distance from the prototype



In the statistical analysis, I will use reaction time as my dependent variable and distance from the prototype as my independent variable. Here, I use a cross-classified model that adds random intercepts for individuals and for scenarios. I also make sure to control for the length of the vignette. Length here is defined as the number of characters in the sentence. Figure 5 shows the coefficients of the model. Under the assumptions of the model, distance from the prototype is positively related with reaction time, even after controlling for length. A one standard deviation increase in the distance from the prototype is predicted to increase reaction time by 0.09 standard deviations. This coefficient, however, does not meet the criteria for conventional significance. However, the confidence interval [from -.01 to 0.19] suggests that the coefficient is likely positive.

Figure 5



Discussion

The studies above show that there is much to be gained by studying the underlying semantic structures of moral transgressions. Research about moral decision-making has relied heavily on the use of fictional vignettes to elicit perceptions of harmfulness and immorality. Here, I leverage existent repositories of cultural meanings to examine the underlying semantic meanings that are associated with the different components of these scenarios. This technique helps me accomplish two main things: it allows me to bring greater rigor to existent empirical evidence and it lets me test current presuppositions about how moral cognition unfolds. Overall, this research shows that examining the connotative meanings of moral violations can help us shed light on the process through which individuals come to decide what is permissible and what is not.

At a general level, this analysis presents a novel approach to a well-trodden method in the study of moral cognition. I take advantage of the fact that sociologists have collected large dictionaries of affective means that contain many of the identities and behaviors that feature in the vignettes used by researchers interested in moral decision-making. This coincidence allows me to translate these fictional scenarios into a syntax that formalizes their semantic properties. In other words, this technique allows us me to measure and contrast

the cultural meanings associated with moral scenarios. This, in turn, yields insightful findings. Consistent with widely accepted theories of morality, I find that transgressions tend to have a patterned structure: they involve a bad and powerful behavior directed at a vulnerable victim. Previous work highlights the importance of the evaluation of the actor and of the victim, but my findings point out the significance of the patient's vulnerability. This semantic feature is a theoretical cornerstone of the dyadic theory of morality. By uncovering this underlying semantic structure, we can place moral transgressions in a cartesian plane and contrast their positions. This technique opens new opportunities related to what we can do with fictional vignettes. It helps us look at the formal properties of scenarios and to examine whether underlying relationships exist at a connotative level.

The ability to measure distances between transgressions allows me to provide a rigorous test of how moral cognition operates. Currently, the most plausible account of moral cognition presupposes that individuals evaluate situations by comparing them with a cognitive template of a prototypical moral transgression. By analyzing the semantic structures of violations, I can make an informed decision about what constitutes an exemplary violation and calculate distances between the prototype and other scenarios. The results of Study 2 suggest that distance from the prototypical moral violation is indeed positively associated with how long it takes individuals to label an event as immoral. This lends credence to the idea that moral cognition occurs as agents compare an event with a cognitive template of what a typical moral violation looks like. While not definitive, this analysis helps adjudicate one of the central ongoing debates in the study of moral reasoning. It provides rigurous evidence that supports the vision of moral cognition as a continuous process, not underpinned by discrete categories but rather organized around fuzzy cognitive templates.

Additionally, these studies also inform the debates that have unfolded around the existence of "harmless" wrongs. While the findings reaffirm the idea that moral violations are not underpinned by different logics, they show that the scholarly fascination around "harmless" transgressions might not be entirely unfounded. When the transgressions are placed in a common plane, we notice that the so-called "purity" transgressions – the ones often identified as "harmless" wrongs – are all grouped in one corner, away from the majority of the events. Thus, while these events rank amongst the most immoral, they are distant from our cognitive template of what a moral violation should look like. This seemingly paradoxical tension might explain why these transgressions have been ascribed such distinctiveness. As study 2 shows, this tension can be explained by noting that distance from exemplary violations does not necessarily affect a transgression's severity but rather how long it takes agents to label it as immoral. This account, in turn, suggest a plausible explanation of why "purity" transgressions tend to be shrouded in taboo and fascination: given that it is more difficult to attribute immorality to these atypical violations, we might have to build a scaffolding of social prohibitions

around them to better signal their wrongness. Now, these are only conjectures, but they provide a set of hypotheses and plausible mechanisms that can be tested. A semantic approach to moral cognition, then, might help us make progress in solving one of the central puzzles in this line of research.