The Moral-Conventional Distinction: A Harm-Based Hypothesis

Philip Petrov

Abstract: Research both older and more recent suggests that people's tendency to distinguish between morality and convention depends on their tendency to react in a particular way to the perception of harm. However, partly because defining "harm" is not straightforward, researchers have proposed disparate ways of understanding the concept for purposes of describing the moral-conventional distinction. This essay presents an updated harm-based account of the moral-conventional distinction. On this account, in contrast to conventional rules, moral rules are those the violation of which tends to produce in observers the mental representation of one type of entity bad-harming another type, where "bad-harm"—which demarcates the sub-set of set-backs to interest that people perceive as morally relevant—is a semantic prime that resists verbal definition but exhibits typicality effects. This account efficiently explains why the moral-conventional distinction is simultaneously pan-cultural and culturally variant, better grounds the study of the moral-conventional distinction in the computational theory of mind, and identifies additional ways in which Elliot Turiel's foundational research can absorb or rebut criticisms.

Keywords: cognition; harm; moral-conventional distinction; moral philosophy; moral psychology; Turiel

Introduction

Researchers in the social and cognitive sciences and in philosophy continue to disagree about how to describe the relation between *morality* and *convention*, and by extension the relation between *moral* and *conventional rules*. Some hold that moral and conventional rules are meaningfully distinct,[1] whereas others question the empirical reality or the conceptual utility of the proposed distinction.[2] Moreover, those who agree that moral and conventional rules are meaningfully distinct often disagree about how to describe the distinction. In short, debate about whether the *moral-conventional distinction* exists, and about how to describe it if it does, persists.

This essay defends three general positions (to be explicated below): (1) the moral-conventional distinction exists; (2) the best way to describe the distinction is in terms of harm;

[1] *See, e.g.*, Jessica Bregant et al., *Crime Because Punishment? The Inferential Psychology of Morality and Punishment*, 2020 UNIVERSITY OF ILLINOIS LAW REVIEW 1177, 1186–87 (2020); Audun Dahl & Elliot Turiel, *Using Naturalistic Recordings to Study Children's Social Perceptions and Evaluations*, 55 DEVELOPMENTAL PSYCHOLOGY 1453 (2019); Bryce Huebner et al., *The Moral-Conventional Distinction in Mature Moral Competence*, 10 JOURNAL OF COGNITION AND CULTURE 1 (2010); Melanie Killen & Audun Dahl, *Moral Reasoning Enables Developmental and Societal Change*, 16 PERSPECTIVES ON PSYCHOLOGICAL SCIENCE 1209, 1214 (2021); Victor Kumar, *Moral Judgment as a Natural Kind*, 172 PHILOSOPHICAL STUDIES 2887, 2892–95 (2015); Ayelet Lahat et al., *An Event-Related Potential Study of Adolescents' and Young Adults' Judgments of Moral and Social Conventional Violations*, 84 CHILD DEVELOPMENT 955 (2013); JOHN MIKHAIL, ELEMENTS OF MORAL COGNITION: RAWLS' LINGUISTIC ANALOGY AND THE COGNITIVE SCIENCE OF MORAL AND LEGAL JUDGMENT 104, 143 (2011); James G. Quigley, *Moral Psychology and the Unity of Morality*, 27 UTILITAS 119, 128 (2015); Stuart F. White et al., *Neural Correlates of Conventional and Harm/Welfare-Based Moral Decision-Making*, 17 COGNITIVE, AFFECTIVE, AND BEHAVIORAL NEUROSCIENCE 1114 (2017).

[2] *See, e.g.*, Joseph Heath, *Morality, Convention and Conventional Morality*, 20 PHILOSOPHICAL EXPLORATIONS 276 (2017); Daniel Kelly et al., *Harm, Affect, and the Moral/Conventional Distinction*, 22 MIND AND LANGUAGE 117 (2007); Neil Levy, *Psychopaths and Blame: The Argument From Content*, 27 PHILOSOPHICAL PSYCHOLOGY 351, 352–58 (2014); Heidi Maibom, *What Experimental Evidence Shows Us about the Role of Emotions in Moral Judgement*, 5 PHILOSOPHY COMPASS 999, 1003 (2010); Joshua May, *The Limits of Emotion in Moral Judgment*, *in* THE MANY MORAL RATIONALISMS 286, 294 (Karen Jones & François Schroeter eds., 2018); Shivam Patel & Edouard Machery, *Do the Folk Need a Meta-Ethics?*, 41 BEHAVIORAL AND BRAIN SCIENCES e109 (2018); K. J. P. Quintelier & D. M. T. Fessler, *Confounds in Moral/Conventional Studies*, 18 PHILOSOPHICAL EXPLORATIONS 58 (2015); Joshua Rottman & Liane Young, *Mechanisms of Moral Development*, *in* THE MORAL BRAIN: A MULTIDISCIPLINARY PERSPECTIVE 123, 126–28 (Jean Decety & Thalia Wheatley eds., 2015); David W. Shoemaker, *Psychopathy, Responsibility, and the Moral/Conventional Distinction*, 49 SOUTHERN JOURNAL OF PHILOSOPHY 99, 105–10 (2011).

and (3) in describing the distinction in terms of harm, it is helpful to define the relevant notion of harm as a *semantic prime* (to be defined below) that resists verbal definition but exhibits typicality effects. In short, the essay builds on or proposes modifications to several earlier accounts of the moral-conventional distinction to present an updated account thereof. This account efficiently explains why the moral-conventional distinction is simultaneously pan-cultural and culturally variant, better grounds the study of the moral-conventional distinction in the computational theory of mind,[3] and identifies additional ways in which Elliot Turiel's foundational research can absorb or rebut criticisms.

Many people find intuitively compelling the idea that the moral-conventional distinction depends in some significant way on harm. However, partly because "harm" is difficult to define, researchers[4] have advanced disparate views about how harm relates to the moral-conventional distinction. This essay argues for the empirical plausibility and conceptual usefulness of a particular harm-based account of the moral-conventional distinction. The following is a preliminary summary of this account. The moral-conventional distinction obtains because people react in a particular way to rule violations that tend to produce in observers the computation

[3] The computational theory of mind is the view, widespread in the cognitive sciences, that mentation can be usefully understood in terms of the performance of computations on mental representations. For one overview, see Michael Rescorla, *The Computational Theory of Mind*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2020). Most research on the moral-conventional distinction, being closer to psychology than to the cognitive sciences, does not use this approach.

[4] *Compare, e.g.*, Paulo Sousa et al., *The Morality of Harm*, 113 COGNITION 80 (2009), *with, e.g.*, Stephen Stich et al., *On the Morality of Harm: A Response to Sousa, Holbrook, and Piazza*, 113 COGNITION 93 (2009).

AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT.[5] People across cultures are predisposed to react negatively to the mental representation of *an agent bad-harming a victim*, where "bad-harm" is a semantic prime that resists verbal definition but exhibits typicality effects.[6] In this formulation, "agent," "bad-harm," and "victim" denote cognitive variables that are highly plastic and can be filled with a variety of contents.[7] For example, in a culture in which people tend to represent laughing at one's mother as a bad-harm, seeing a child (AGENT) snicker at (BAD-HARM) her mother (VICTIM) can generate empathy or sympathy toward the mother and anger or blame toward the child. Let the mental state that is generated by the mental representation of an agent bad-harming a victim be denoted *moral judgment*. On the present account, moral rules are those the violation of which tends to produce moral judgment in observers. For example, in a culture in which people tend to represent sorcery as a bad-harm, people are likely to perceive a rule against sorcery (e.g., "do not hex others") as a moral rule. By contrast, conventional rules are those that provide people with guidance about how to behave in situations of an identifiable type but whose violation does not tend to produce moral judgment in observers. For example, in many U.S. office buildings, people recognize a convention to the effect that, upon entering a crowded elevator, one should turn to face the elevator doors. If a person violates this convention, most observers will *not* perceive the person as having violated a moral rule, because, for most people, seeing someone fail

[5] The view that moral judgment is due to the perception of an active type of entity bad-harming a passive type of entity is old and general. *See, e.g.*, ARTHUR SCHOPENHAUER, THE BASIS OF MORALITY 165–71 (Arthur Brodrick Bullock trans., 2d ed. 1915). The best-known and most developed specification of this view in contemporary psychology is that of Kurt Gray and Chelsea Schein, who have argued in detail that an agent harming a victim constitutes the archetype of a moral wrong. *See, e.g.*, Chelsea Schein & Kurt Gray, *The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm*, 22 PERSONALITY AND SOCIAL PSYCHOLOGY REVIEW 32 (2018). In its contention that moral judgment depends on the perception of an agent bad-harming a victim, the present analysis builds on Gray's and Schein's argument.

[6] I am grateful to Jonathan Bendor for helping me to recognize the relevance of the concept of bad-harm to the study of moral psychology.

[7] For the view that "agent," "harm," and "victim" are cognitive variables, see especially Chelsea Schein & Kurt Gray, *The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm*, 22 PERSONALITY AND SOCIAL PSYCHOLOGY REVIEW 32, 32–51 (2018).

to face the elevator doors upon entering a crowded elevator is not sufficient to produce the representation of an agent bad-harming a victim. On the present account, then, the moral-conventional distinction is a consequence of the mind's tendency to compute rule violations that tend to produce moral judgment differently from the way in which it computes rule violations that do not have this effect.

The essay has three parts. **Part I** briefly describes Elliot Turiel's foundational (c. 1970s–90s) account of the moral-conventional distinction and proposes a friendly amendment to it. Because Turiel's is the oldest and best-known explicit account of the moral-conventional distinction, and because contemporary researchers continue to build on and to challenge his work, it remains useful to begin a study of the moral-conventional distinction with Turiel. **Part II** fast-forwards to the present and examines several recent studies of the moral-conventional distinction. It concludes that recent research on the subject both supports and can benefit from an account that describes the moral-conventional distinction in terms of a semantic prime that resists verbal definition but exhibits typicality effects. **Part III** presents such an account.

Before proceeding, it is helpful to clarify the argument in at least two respects. First, the argument proceeds from naturalistic and sentimentalist starting points: it assumes that moral phenomena are explicable in ordinary natural-scientific terms (naturalism) and that human moral judgment both requires the capacity for, and consists partly of, affect or emotion (sentimentalism). This means that some researchers in moral psychology—those who reject naturalism or sentimentalism—would not entirely endorse this essay's argument. I hope that even readers who reject these commitments will find the essay valuable, if only for clarifying and adding exactitude to a perspective with which they disagree.

Secondly, although this essay studies mental phenomena, it is not—and does not aim to be—a publication in psychology. Nearly all research in psychology today focuses on presenting experimental results (in this regard, psychology differs from fields such as political science, which still has a dedicated "theory" wing). This essay, by contrast, is theoretical: instead of conducting a

new experiment, it analyzes existing experimental results to clarify and to sharpen the understanding of the moral-conventional distinction. The reader whose primary field is psychology may find it helpful to conceive of the essay as a philosophical analysis. Conceiving of the essay in this way should emphasize that its success criteria are not those of a paper reporting an experiment. Relatedly, as several researchers[8] have argued, the study of the moral-conventional distinction is, and should be, inter-disciplinary. Given that studying the moral-conventional distinction often requires studying how people categorize social artifacts,[9] the moral-conventional distinction is an obvious topic for psychology. At the same time, given that properly interpreting people's categorizations often requires attending to how subjects and experimenters understand and use words,[10] the moral-conventional distinction is an equally obvious topic for philosophy.

I. The Moral-Conventional Distinction: Turiel's Foundational Formulation

This part briefly reviews Turiel's foundational account of the moral-conventional distinction and proposes a friendly amendment to it.

Turiel is a U.S. developmental psychologist who has done as much as anyone else systematically to study the moral-conventional distinction. Beginning in the late 1970s, Turiel

[8] *See, e.g.*, Andrew W. Delton & Max M. Krasnow, *Adaptationist Approaches to Moral Psychology*, *in* THE MORAL BRAIN: A MULTIDISCIPLINARY PERSPECTIVE 19, 19 (Jean Decety & Thalia Wheatley eds., 2015); John M. Doris, *Introduction*, *in* THE MORAL PSYCHOLOGY HANDBOOK 1, 1 (John M. Doris & Moral Psychology Research Group eds., 2010).

[9] Justin Landy has developed this point in greater detail. *See, e.g.*, Justin F. Landy, *Representations of Moral Violations: Category Members and Associated Features*, 11 JUDGMENT AND DECISION MAKING 496, 497–98 (2016).

[10] Nicholas Southwood has developed this point in greater detail. *See, e.g.*, Nicholas Southwood, *The Moral/Conventional Distinction*, 120 MIND 761, 762–63 (2011).

(working with several collaborators[11]) reported that, by the age of five,[12] the vast majority of children develop the ability to distinguish between morality and convention, and that they distinguish between the two on the basis of a small number of discrete dimensions.[13] More precisely, extrapolating from many (by now, dozens) of experiments that he and his collaborators had conducted, Turiel reported that young children tend to respond in systematically different ways to actions such as "Mary hit William" (which Turiel classified as a moral violation) versus actions such as "Mary came to school wearing pajamas" (which he classified as a conventional violation). Turiel also reported that people continue to draw the moral-conventional distinction throughout their lives, and that the tendency to distinguish between morality and convention holds across cultures.

The following discussion provides a sense of Turiel's experimental methodology. Turiel presented young children with two types of vignettes. The first type—Turiel's "moral" category—included events such as:

> "Mary hit William."
> "Mary destroyed William's toy."
> "Mary made fun of William until he started to cry."

The second type—Turiel's "conventional" category—included events such as:

> "Mary came to school wearing pajamas."
> "Mary left William's toy on the floor after playing with it."

---

[11] Turiel's collaborators have included, *inter alia*, Charles Helwig, Larry Nucci, Judith Smetana, Marie Tisak, and Donna Weston.

[12] This is a safe estimate. Consider, for instance, Ha Na Yoo's and Judith Smetana's statement that "children's moral and conventional distinctions in criterion judgments emerge by age 3, and the distinction effects get stronger with age." Ha Na Yoo & Judith G. Smetana, *Distinctions Between Moral and Conventional Judgments From Early to Middle Childhood: A Meta-Analysis of Social Domain Theory Research*, 58 DEVELOPMENTAL PSYCHOLOGY 874, 883 (2022).

[13] For a recent overview of Turiel's moral-conventional distinction, see, e.g., id. For an older overview, see, e.g., Judith G. Smetana, *Social-Cognitive Domain Theory: Consistencies and Variations in Children's Moral and Social Judgments*, *in* HANDBOOK OF MORAL DEVELOPMENT 119 (Melanie Killen & Judith G. Smetana eds., 2006).

"Mary called her teacher by his first name."[14]

Turiel asked children to answer closed-ended questions about these vignettes (e.g., "would the act still be wrong if people in another place thought that it was okay?"). He also asked them to provide open-ended explanations of why, if at all, the actions described in the vignettes were wrong. Consider one of Turiel's and Larry Nucci's experiments with Amish-Mennonite children.[15] Turiel and Nucci set out to compare the children's reactions to hitting another person (which the authors classified as a moral violation) to their reactions to working on the sabbath (which they classified as a conventional violation). First, Turiel and Nucci asked the children whether hitting another person was wrong. Nearly all of the children responded that it was. Turiel and Nucci then asked the children whether hitting another would still be wrong if god said that it was okay. Most children responded that hitting would still be wrong. Next, Turiel and Nucci asked the children whether working on the sabbath was wrong. Again, nearly all of the children responded that it was. However, when Turiel and Nucci asked the children whether working on the sabbath would still be wrong if god okayed it, most of the children responded that it would no longer be wrong. From this response pattern, Turiel and Nucci inferred that, whereas most of the children perceived hitting as an "authority-independent" violation (wrong regardless of whether an authority forbids it), they perceived working on the sabbath as an "authority-dependent" one (wrong only if an authority forbids it).

On the basis of many such experiments, Turiel concluded that, by the age of five, the vast majority of children distinguish between moral and conventional violations, and that they do so

[14] These are variations on some of the many vignettes that Turiel has used. For examples of Turiel's own wording, see, e.g., Philip Davidson et al., *The Effect of Stimulus Familiarity on the Use of Criteria and Justifications in Children's Social Reasoning*, 1 BRITISH JOURNAL OF DEVELOPMENTAL PSYCHOLOGY 49, 52 (1983).

[15] *See* Larry Nucci & Elliot Turiel, *God's Word, Religious Rules, and Their Relation to Christian and Jewish Children's Concepts of Morality*, 64 CHILD DEVELOPMENT 1475 (1993). The children in this study were older (10–16).

along a small number of discrete dimensions. Turiel's summations of his moral-conventional distinction have changed somewhat over time. Thus, anyone seeking to summarize his position must make editorial decisions. Many researchers[16] hold that his research emphasizes four main dimensions of difference between moral and conventional violations: *harmfulness*, *seriousness*, *authority-independence*, and *generalizability* (hereafter, "HSAG") (see Figure 1). This essay follows this interpretive trend.

	moral violation	conventional violation
harmfulness	often described as harmful	rarely described as harmful
seriousness	described as more serious	described as less serious
authority-independence	described as a violation regardless of whether an authority forbids it	described as a violation only if an authority forbids it
generalizability	described as a violation regardless of the culture in which it occurs	described as a violation only if it occurs in certain cultures

Figure 1: Turiel's Moral-Conventional Distinction: HSAG

In summary, Turiel reported that, by the age of five, the vast majority of children develop the ability to distinguish between morality and convention; that children distinguish between the

[16] *See, e.g.*, Joseph Heath, *Morality, Convention and Conventional Morality*, 20 PHILOSOPHICAL EXPLORATIONS 276, 279 (2017); Bryce Huebner et al., *The Moral-Conventional Distinction in Mature Moral Competence*, 10 JOURNAL OF COGNITION AND CULTURE 1, 2 (2010); Daniel Kelly et al., *Harm, Affect, and the Moral/Conventional Distinction*, 22 MIND AND LANGUAGE 117, 118 (2007). For an interpretation of Turiel's research that describes it as emphasizing only two main dimensions of difference between moral and conventional violations, see, e.g., Paulo Sousa & Jared Piazza, *Harmful Transgressions qua Moral Transgressions: A Deflationary View*, 20 THINKING AND REASONING 99, 103 (2014).

two on the basis of HSAG; that people continue to draw the moral-conventional distinction throughout the life-cycle; and that morality and convention are "distinct conceptual domains."

Today, although many researchers[17] accept Turiel's moral-conventional distinction, many others[18] doubt or reject it. As one team of authors has put it, "[t]he moral-conventional distinction has a contested history."[19] It is difficult to estimate what percentage of researchers in any given field accept versus reject Turiel's moral-conventional distinction, but one can safely state that *many but far from all* researchers concur in it.

Given that some contemporary researchers reject Turiel's formulation of the moral-conventional distinction, I propose a friendly modification to the latter, one that I believe may assuage some critics' resistance to Turiel's position. Turiel's supporters need not and should not defend the view that HSAG are "criteria" that demarcate the "moral domain" (both Turiel[20] and his supporters[21] have often used both of these concepts). Both of these concepts—"criteria" and "moral domain"—generate more costs than benefits for Turiel's position. The concept of a domain is highly ambiguous in the present context, and skeptics[22] have reasonably criticized Turiel's reliance on it. Indeed, given that Turiel has advanced a psychological or *mind-dependent* account of the moral-conventional distinction, it is not clear that the concept of a domain—which has a *mind-independent* connotation—is useful to Turiel. Moreover, describing HSAG as "criteria" risks

[17] *See supra* note 1.

[18] *See supra* note 2.

[19] Audun Dahl & Talia Waltzer, *Constraints on Conventions: Resolving Two Puzzles of Conventionality*, 196 COGNITION 104152, 1 (2020).

[20] *See, e.g.*, Elliot Turiel, *The Development of Morality*, *in* CHILD AND ADOLESCENT DEVELOPMENT: AN ADVANCED COURSE 473, 493 (William Damon & Richard M. Lerner eds., 2008).

[21] *See, e.g.*, Cameron B. Richardson et al., *Extending Social Domain Theory With a Process-Based Account of Moral Judgments*, 55 HUMAN DEVELOPMENT 4, 5 (2012).

[22] *See, e.g.*, Stephen Stich, *The Moral Domain*, *in* ATLAS OF MORAL PSYCHOLOGY 547, 552–53 (Kurt Gray & Jesse Graham eds., 2018).

creating the unfortunate impression that, if people do not tend to perceive a rule as a moral rule along *all* of Turiel's dimensions, then the rule cannot be usefully described as a moral rule. Skeptics have reasonably criticized Turiel on this basis, too.[23] Instead of describing HSAG as criteria that demarcate the moral domain, Turiel and his supporters should, I propose, describe HSAG as *some of the experiential or phenomenological correlates of the negative reaction that people often automatically experience upon mentally representing an agent bad-harming a victim*. For example, seeing *a* steal *v*'s property may generate sympathy toward *v* and blame toward *a*, and these mental representations may motivate the observer to characterize *a*'s behavior as (1) serious and (2) wrong regardless of whether an authority forbids it. From this perspective, the four HSAG dimensions are not on a par, because harmfulness is the most causally important one of them. Put differently, it is not that moral violations (as opposed to conventional violations) trigger HSAG, but that moral violations (as opposed to conventional violations) trigger *harmfulness* (H), which in turn often but not always triggers *seriousness*, *authority-independence*, and *generalizability* (SAG). This modification softens Turiel's moral-conventional distinction and deprives it of some of its ideational content. However, the modification has an important benefit: it emphasizes that what at bottom generates the moral-conventional distinction is the particular effect that perceptions of certain types of harm exert on the mind.

II. The Moral-Conventional Distinction: Recent Reported Findings

This part has two sections. Section II.A discusses recent reported findings that suggest that the moral-conventional distinction depends on harm. Section II.B discusses recent conceptualizations of the role of harm in the moral-conventional distinction.

[23] *See, e.g.*, Edouard Machery & Ron Mallon, *Evolution of Morality*, *in* THE MORAL PSYCHOLOGY HANDBOOK 3, 34 (John M. Doris & Moral Psychology Research Group eds., 2010).

Since Turiel's foundational work of the 1970s, 80s, and 90s, researchers[24] have greatly expanded the literature on the moral-conventional distinction, making it fragmented, heterogenous, and difficult quickly to review. Rather than trying to summarize this literature, this part analyzes a small number of highly relevant studies therein. My selection of these studies as opposed to others is not an implicit criticism of other recent work on the moral-conventional distinction; given more space, I would review more studies.[25]

II.A The moral-conventional distinction's dependence on harm

Several recent studies converge to suggest, in different ways and using different experiment designs, that the moral-conventional distinction depends in some significant way on harm. Consider three examples.

In one study, Audun Dahl and Talia Waltzer[26] reported that, in many or most cases, although people believe that more than one convention can satisfactorily resolve a given coordination problem, people's concern about harm to others constrains their beliefs about what

[24] For examples, see *supra* notes 1 and 2.

[25] For other relevant studies, *see, e.g.*, R. J. R. Blair, *Emotion-Based Learning Systems and the Development of Morality*, 167 COGNITION 38 (2017); Carly Giffin & Tania Lombrozo, *An Actor's Knowledge and Intent Are More Important in Evaluating Moral Transgressions Than Conventional Transgressions*, 42 COGNITIVE SCIENCE 105 (2018); Susanne Hardecker et al., *Young Children's Behavioral and Emotional Responses to Different Social Norm Violations*, 150 JOURNAL OF EXPERIMENTAL CHILD PSYCHOLOGY 364 (2016); Marina Josephs & Hannes Rakoczy, *Young Children Think You Can Opt Out of Social-Conventional but Not Moral Practices*, 39 COGNITIVE DEVELOPMENT 197 (2016); Elizabeth B. Kim et al., *Does Children's Moral Compass Waver Under Social Pressure? Using the Conformity Paradigm to Test Preschoolers' Moral and Social-Conventional Judgments*, 150 JOURNAL OF EXPERIMENTAL CHILD PSYCHOLOGY 241 (2016); Ayelet Lahat et al., *Cognitive Processing of Moral and Social Judgements: A Comparison of Offenders, Students, and Control Participants*, 68 QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY 350 (2015); Francesco Margoni, *The Distinction Between Morality and Convention in Older Adults*, 53 COGNITIVE DEVELOPMENT 100840 (2020); Kelly Lynn Mulvey, *Evaluations of Moral and Conventional Intergroup Transgressions*, 34 BRITISH JOURNAL OF DEVELOPMENTAL PSYCHOLOGY 489 (2016).

[26] Audun Dahl & Talia Waltzer, *Constraints on Conventions: Resolving Two Puzzles of Conventionality*, 196 COGNITION 104152, 10–11 (2020).

types of conventions are appropriate. To use Dahl's and Waltzer's[27] main example, consider a basketball team that must choose a team uniform for competition. Dahl and Waltzer reported that, although people tend to believe that both a blue and a red jersey can constitute a satisfactory convention, people also tend to believe that a jersey that contains a message that would severely offend spectators is not. Notice a basic implication of Dahl's and Waltzer's study: it suggests that, in choosing a solution to a coordination problem, once people begin to perceive that a proposed convention may cause harm to others, they begin to use moral considerations to limit the menu of acceptable proposed conventions. Dahl's and Waltzer's study thus illustrates that the representation of harm to others is likely to play a significant role in any accurate description of the moral-conventional distinction.

In another study, Antonia Langenhoff, Audun Dahl, and Mahesh Srinivasan[28] showed children (3–5) and adults a puppet performing two different actions: (1) pressing on the hand of another puppet who subsequently displays indications of pain and (2) pressing on a button on a box after an authority figure has disallowed touching the box. In line with Turiel, the authors classified (1) as a moral violation and (2) as a conventional violation. Importantly, (1) and (2) were physical-structurally similar actions: in both, a puppet approached an object and pressed on it. Langenhoff et al. reported that, despite the physical-structural similarity between the two actions, subjects' behavioral and verbal reactions to (1) and (2) indicated that they perceived (1) as a moral violation and (2) as a conventional violation. Langenhoff et al.'s study suggests that, when people categorize a rule violation along the moral-conventional continuum, they automatically look beyond the physical-structural character of the violation (e.g., "does the underlying action involve a push or a pull?") and attend to the violation's effect on others' welfare—to whether the violation causes harm. A simple but notable implication of this is that a purely physical description of an

[27] *See id.*

[28] Antonia F. Langenhoff et al., *Preschoolers Learn New Moral and Conventional Norms From Direct Experiences*, 215 JOURNAL OF EXPERIMENTAL CHILD PSYCHOLOGY 105322 (2022).

action (e.g., "Mary pressed on William's hand") is not necessarily sufficient to determine whether people would tend to categorize the action as moral or as conventional.

In a third study, Meltem Yucel, Robert Hepach, and Amrisha Vaish showed young children (ages 3–4) and adults two vignettes: (1) an agent destroying another's belongings and (2) an agent playing a board game with a friend and defying the game's rules.[29] In line, again, with Turiel, the authors classified (1) as a moral violation and (2) as a conventional violation. The authors reported that both the children and the adults exhibited greater *pupil dilation* (a measure of affective arousal) upon seeing (1) than they did upon seeing (2). The authors also reported that both groups exhibited greater *looking* (a measure of attention) toward the person whose belongings the agent destroyed than they did toward the friend who played the board game with the agent. This result suggests that perceiving violations of moral rules involves concern about harm to others in a way that perceiving violations of conventional rules does not, as well as that this concern is partly affective.

Taken together, the three preceding studies exemplify that much recent research on the moral-conventional distinctions converges on the conclusion that, in categorizing rules as moral versus conventional, people rely heavily on perceptions of harm.

II.B The meaning of "harm" in the moral-conventional distinction

Stating that the moral-conventional distinction depends on harm leaves unanswered the question of how an account of the moral-conventional distinction should conceptualize or understand "harm." This section considers three conceptualizations of harm in recent research on the moral-conventional distinction, focusing especially on their limits. Part III below then presents an alternative such conceptualization.

[29] *See* Meltem Yucel et al., *Young Children and Adults Show Differential Arousal to Moral and Conventional Transgressions*, 11 FRONTIERS IN PSYCHOLOGY 548 (2020).

In describing the moral-conventional distinction, one way to conceptualize harm is to leave it undescribed, in which case readers are likely to assign the word its ordinary-language or colloquial meaning. One limitation of this approach is that, because people differ greatly in their understandings of harm,[30] it is liable to cause confusion. Moreover, in much colloquial usage, ordinary speakers restrict "harm" to physical injury and to emotional distress,[31] but this conception of harm is too narrow for purposes of describing the moral-conventional distinction. This is because, among other things, people often categorize as *moral* violations that they do not perceive as involving physical injury or emotional distress, such as violations of a rule against belittling a dead person. To describe the moral-conventional distinction in terms of harm successfully, it is necessary to give "harm" at least some definitional contour—to depart from colloquiality to at least some degree. This means that leaving "harm" undescribed is unlikely to be the best path forward for researchers who seek to provide a harm-based account of the moral-conventional distinction.

For another and more contoured conceptualization of harm, consider Paulo Sousa's and Jared Piazza's proposal that, in contrast to conventional violations, moral violations are those that involve infringements of "basic rights."[32] Defining harm as "pain or suffering," Sousa and Piazza argued that, because many harmful actions involve infringements of basic rights, people often categorize harmful actions as moral violations. On Sousa's and Piazza's account, the distinction between morality and convention can be usefully described in terms of infringements of basic

[30] *See, e.g.*, Nick Haslam, *Concept Creep: Psychology's Expanding Concepts of Harm and Pathology*, 27 PSYCHOLOGICAL INQUIRY 1 (2016).

[31] Jurists also often restrict "harm" in this way. Specialists in U.S. tort law, for instance, often conceive of harm as being limited to bodily injury, property damage, and emotional distress. *See, e.g.*, AMERICAN LAW INSTITUTE, 1 RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM 1 (2010).

[32] Paulo Sousa & Jared Piazza, *Harmful Transgressions qua Moral Transgressions: A Deflationary View*, 20 THINKING AND REASONING 99, 104, 122–26 (2014).

rights, and, because harmful actions often involve such infringements, the presence of harm in a rule violation often causes people to categorize the violation as moral.

Without questioning the conceptual utility or the intellectual verdancy of Sousa's and Piazza's account, I propose two friendly modifications to it. First, their account uses two significant sources of vagueness: vagueness as to the meaning of "harm," and vagueness as to the meaning of "basic right." If it is possible to produce an account that consolidates these two sources of vagueness into one, then it is likely to be advantageous to do so (in theory construction, minimizing the number of sources of vagueness is a general methodological desideratum). Thus, an alternative account might try to use only one, not two, significantly vague terms. Secondly, Sousa's and Piazza's account indexes the moral-conventional distinction to the concept of basic rights:

> Since we hypothesise that the folk concept of basic-rights violation/injustice implies transgression as well as authority independence and generality, *all actions interpreted as involving basic-rights violation/injustice are to be understood as moral transgressions*, independent of other properties, such as being harmful, that one may attribute to the action . . . .[33]

The concept of basic rights may be too cognitively elaborate and too culturally contingent to be successful in describing the moral-conventional distinction. For example, people who live in cultures that do not have a folk concept of basic-rights—or even a folk concept of rights in general—presumably nonetheless distinguish between morality and convention.[34] This suggests that it is not so much the perception of a basic-rights violations but rather some other perception—perhaps a less cognitively elaborate one—that causes people to perceive certain violations as moral

[33] *Id.* at 105 (emphasis added).

[34] Relatedly, researchers in political theory who find the concept of moral or non-legal rights useful often disagree about which such rights are basic. *See, e.g.*, Jeppe von Platz, *Are Economic Liberties Basic Rights?*, 13 POLITICS, PHILOSOPHY, AND ECONOMICS 23 (2014).

versus conventional. Thus, an alternative account might try to avoid the concept of basic rights altogether.

Consider, now, a third conceptualization of harm. Edward Royzman, Robert Leeman, and Jonathan Baron proposed that:

> [H]umans are biologically prepared to moralize (accord social transcendence to) normative principles that concern acts of interpersonal harm, with harm being understood broadly enough to capture a multitude of ways in which one person's behavior may (culpably) reduce or constrain the utility (good) of others. . . . This capacity is, by stipulation, emotion-independent and separate from general-purpose reasoning. . . . [T]he process of selective moralization is effected by a system oriented towards a particular rule content and . . . this content is largely defined by acts or dispositions deemed intrinsically harmful to others (in the general sense of an act's having a negative impact on the utilities of others, the meaning that need not correspond to and is unlikely to be fully captured by the relatively narrow lay meaning of *harm* or *hurt*).[35]

Royzman et al.'s account holds that people tend to treat rule violations that involve intrinsic harm to others as moral rules, and that people's tendency to do this does not involve affect or emotion.

Although I largely concur in Royzman et al.'s conceptually useful and intellectually verdant account, I propose two friendly modifications to it. First, in conceptualizing harm, it may be helpful to avoid the concept of intrinsic-ness,[36] because many actions that people tend to categorize as moral violations cannot usefully be described an intrinsically harmful. Consider, for instance, a study whose authors reported that three-year-old-children can understand that, if hitting an animal of an unfamiliar species provides pleasure to the animal, then it is not morally

---

[35] Edward B. Royzman et al., *Unsentimental Ethics: Towards a Content-Specific Account of the Moral–Conventional Distinction*, 112 COGNITION 159, 165–73 (2009) (internal citations omitted). Note Royzman et al.'s observation that the conception of harm as physical injury or emotional distress is too narrow for purposes of describing the moral-conventional distinction.

[36] For a criticism of the use in research in moral psychology of the related concept of "essence," see Andrew E. Monroe et al., *Morality Goes Beyond Mind Perception*, 23 PSYCHOLOGICAL INQUIRY 179, 180 (2012).

wrong to do so (at first glance, hitting would seem to be an excellent example of an intrinsically harmful action).[37] More generally, to the extent that the moral-conventional distinction is mind-dependent, the concept of intrinsic-ness may not be apt in describing it. With respect to how people categorize rule violations as moral versus conventional, what matters is not so much the (mind-independent) nature of the violation but rather how observers (mind-dependently) perceive the violation. Secondly, it may be too strong to characterize people's capacity to moralize harmful acts as "emotion-independent," because many or most negative moral evaluations of harmful acts (e.g., "Mary should be castigated for what she did") involve the activation of emotion[38] (this is true whether one understands emotion in "basic emotions" or in "constructivist" terms[39]). Royzman et al. may well be correct that the human capacity and tendency to moralize a certain type of harm is at least partly modular (in their terms, "emotion-independent and separate from general-purpose reasoning"), but it is not clear why the modularity of moralization would preclude the involvement of affect in the latter.

III. The Moral-Conventional Distinction: An Updated Harm-Based Account

Part I above reviewed Turiel's foundational formulation of the moral-conventional distinction. Part II above reviewed several recent studies that either suggest that the moral-conventional distinction depends on harm or conceptualize harm for purposes of the moral-conventional distinction. Building on or modifying earlier research, this part presents an updated harm-based account of the moral-conventional distinction.

---

[37] Philip David Zelazo et al., *Intention, Act, and Outcome in Behavioral Prediction and Moral Judgment*, 67 CHILD DEVELOPMENT 2478, 2488 (1996).

[38] *See, e.g.*, Jean Decety et al., *The Contribution of Emotion and Cognition to Moral Sensitivity: A Neurodevelopmental Study*, 22 CEREBRAL CORTEX 209, 216–17 (2012).

[39] For an overview of the basic emotions-constructivism debate in the study of affect, see Kristen A. Lindquist et al., *The Hundred-Year Emotion War: Are Emotions Natural Kinds or Psychological Constructions?*, 139 PSYCHOLOGICAL BULLETIN 255 (2013).

It is a very general feature of moral cognition that people across cultures are predisposed to react negatively to the mental representation of *an agent bad-harming a victim*.[40] For example, seeing a child (AGENT) snicker at (BAD-HARM) her mother (VICTIM) can generate empathy or sympathy toward the mother and anger or blame toward the child. The relevant computation can be denoted, in abbreviated form, as:

AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT

In this formulation, AGENT denotes "actor" or "causer," VICTIM denotes "patient" or "recipient," and MORAL JUDGMENT denotes a mental state characterized by components such as affective arousal, negative valence toward the AGENT, and motivation to punish the latter. BAD-HARM, which is a more subtle concept, requires at least some elaboration. One way to understand "bad-harm" is in terms of the idea in normative moral philosophy that a harm is *a set-back to an interest*.[41] If a harm is a set-back to an interest, then "bad-harm" demarcates the sub-set of set-backs to interest that people perceive as morally relevant. For example, to most people, *a* defeating *v* in tennis—and thereby setting back *v*'s interest in winning in the game—is not a bad-harm, whereas *a* negligently driving into and destroying *v*'s home is. Another way to understand "bad-harm" is in terms of the idea of a *semantic prime* in linguistics. Cliff Goddard and Anna Wierzbicka have reported that "bad" is one of a relatively small number of semantic primes—words or word-meanings that exist in all natural languages and that, because one cannot paraphrase them in simpler terms, resist verbal definition.[42] Examples of semantic primes, in Goddard's and

[40] For some of the extensive evidence for this proposition, see *infra* note 49. The best-known and most developed "agent-harm-victim"-based account of moral judgment in contemporary psychology is that of Kurt Gray and Chelsea Schein. *See supra* note 5.

[41] For the best-known and most developed conceptualization of harm as set-back to interest in contemporary philosophy, see, e.g., JOEL FEINBERG, THE MORAL LIMITS OF THE CRIMINAL LAW: HARM TO OTHERS 31–64 (1984).

[42] *See, e.g.*, CLIFF GODDARD & ANNA WIERZBICKA, WORDS AND MEANINGS: LEXICAL SEMANTICS ACROSS DOMAINS, LANGUAGES, AND CULTURES 12 (2014).

Wierzbicka's sense, are "bad," "good," "me," "you," "big," "small," "see," "hear," "live," and "die." A core feature of semantic primes is that one cannot verbally define them without circularity.[43] That "bad" is a semantic prime explains why "bad-harm" resists verbal definition (and why nobody has yet succeeded in formulating an uncontroversial normative moral theory that explains which harms or set-backs to interest are morally relevant (bad) and which are not morally relevant (not-bad)). Given the mental technology of moral judgment that arises in the ordinary course of human ontogeny, as well as often culturally variable beliefs about the range of objects that qualify as bad, people automatically encode certain harms, and not others, as bad-harms. These bad-harms are the harms representing which causes people to make moral judgments. That philosophers and other writers would struggle to produce a verbal definition of "bad-harm" reflects not a lack of intellectual acuity but the partly inaccessible nature of the moral judgment computation. To borrow a term from ethology,[44] the difficulty of verbally defining "bad-harm" seems to be a *species-typical* cognitive constraint, one that affects not only amateurs such as most ordinary speakers but also relevant specialists such as people with advanced degrees in philosophy. One more way to understand "bad-harm" is in terms of the *prototype theory of concepts* in psychology and in philosophy.[45] Although "bad-harm" resists verbal definition, people tend to perceive certain events (e.g., physical attacks) as more prototypical instances of bad-harms than they do others (e.g., verbal insults).[46] Indeed, the concept of bad-harm—like the concept of

---

[43] *See, e.g.*, *id.*

[44] *See, e.g.*, Michael R. Murphy et al., *Species-Typical Behavior of Hamsters Deprived From Birth of the Neocortex*, 213 SCIENCE 459 (1981).

[45] *See, e.g.*, Hans Kamp & Barbara Partee, *Prototype Theory and Compositionality*, 57 COGNITION 129 (1995) (psychology); Eric Margolis, *How to Acquire a Concept*, 13 MIND AND LANGUAGE 347 (1998) (philosophy).

[46] *See, e.g.*, Isobel A. Heck et al., *"There Are No Band-Aids for Emotions": The Development of Thinking About Emotional Harm*, 57 DEVELOPMENTAL PSYCHOLOGY 913 (2021); *see also, e.g.*, Brodie C. Dakin et al., *Broadened Concepts of Harm Appear Less Serious*, 14 SOCIAL PSYCHOLOGICAL AND PERSONALITY SCIENCE 72 (2023).

morality—is most likely a prototype concept, one that people do not typically conceive of in terms of, for instance, individually necessary and jointly sufficient conditions. All in all, a bad-harm is a harm or set-back to interest that people perceive as morally relevant and the representation of which tends to produce moral judgment. Although "bad-harm" is partly a semantic prime and thus resists verbal definition, it exhibits typicality effects. This partly explains why people can often agree about what types of events qualify as important instances of morally relevant harm.

In one sense, AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT is simply a formalization of the very familiar fact that perceiving one type of entity set back an interest of another type can cause the observer to make a negative moral evaluation. All of us are familiar with this computation; we have performed it many times. In a more scientifically interesting sense, AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT is what some psychologists[47] would call a "cognitively natural" or "cognitively intuitive" psychological mechanism: it often occurs automatically, quickly, and irresistibly; it requires minimal or no education or training (although education and

[47] *See, e.g.*, ROBERT N. MCCAULEY, WHY RELIGION IS NATURAL AND SCIENCE IS NOT 3 (2011).

training can affect it in behaviorally important ways[48]); it occurs in people in all cultures; and the wholesale inability to perform it is evidence of psycho-pathology.[49]

The preceding account of moral judgment—while condensed, incomplete, and schematic—is sufficient to generate a useful harm-based account of the moral-conventional distinction. According to this account, moral and conventional rules have distinct prototypes. Moral rules are those the violation of which tends to trigger the computation AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT in observers. Consider the rule forbidding battery (in the colloquial and non-legal sense of "hitting intended to cause physical injury"). People tend to perceive this rule as moral, because the action that it forbids (e.g., "she hits him") closely resembles the "agent bad-harm victim" prototype. Consider, by contrast, the rule prescribing that white moves first in chess. People tend to perceive this rule as conventional, not moral, because the action that it forbids is too distant from the preceding prototype. In contrast to moral rules, conventional rules are those the violation of which does not tend to trigger the computation AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT in observers.

[48] Hanno Sauer has developed this point in greater detail. *See, e.g.*, HANNO SAUER, MORAL JUDGMENTS AS EDUCATED INTUITIONS 51–83 (2017).

[49] The proposition that people across cultures are predisposed to react negatively to the mental representation of an agent bad-harming a victim is supported by a large body of evidence, including: (1) research in developmental psychology reporting that children under one year of age can exhibit avoidance behavior toward agents whom they have seen interfering with others' interests, *see, e.g.*, J. Kiley Hamlin et al., *Social Evaluation by Preverbal Infants*, 450 NATURE 557 (2007); (2) research in animal psychology reporting that, in some non-human species, individuals can "share" in the suffering of conspecifics, *see, e.g.*, Inbal Ben-Ami Bartal et al., *Empathy and Pro-Social Behavior in Rats*, 334 SCIENCE 1427 (2011); and (3) research in semantics reporting that the syntactic construction "someone does something to someone else"—which lexicalizes the agent harm victim representation—seems to exist in all natural languages, *see, e.g.*, CLIFF GODDARD & ANNA WIERZBICKA, WORDS AND MEANINGS: LEXICAL SEMANTICS ACROSS DOMAINS, LANGUAGES, AND CULTURES 13–14 (2014). Each of these research programs is open to challenge and is subject to more than one reasonable interpretation. Nonetheless, taken together, research programs such as these support the proposition that the computation AGENT BAD-HARM VICTIM ⟶ MORAL JUDGMENT is cognitively intuitive.

Although moral and conventional rules have distinct prototypes, many rules have the feature that it is difficult to determine whether they are closer to the one prototype or to the other. Put differently, many rules have the feature that it is difficult to predict whether people would tend to perceive them as moral or as conventional. Consider the rule prescribing black dress at funerals. If I attend a Western funeral in a beige suit, some guests, representing me as setting back the deceased's interest in self-respect, may perceive that I have violated a moral rule (e.g., "he's insulting the dead"). Others, representing me as being eccentric, may perceive that I have violated a conventional one (e.g., "he's eclectic"). Without relatively fine-grained knowledge about the guests, it would be difficult to predict which of them would be more likely to exhibit the one reaction as opposed to the other.

One way to render the present hypothesis more precise is via the concept of *stimulus degradation*.[50] In cognitive psychology, stimulus degradation is the process whereby a stimulus becomes increasingly more difficult to perceive—increasingly less "intact"—due to disruption in its ability to transmit information. For example, making a written word on a page or a computer screen appear increasingly blurry or small is a way to degrade the stimulus. The present hypothesis can be usefully stated in terms of the concept of stimulus degradation: in the perception of a rule violation, the less (more) degraded is the AGENT BAD-HARM VICTIM stimulus, the more (less) likely is the observer to perceive the rule as a moral rule. Consider, again, the rule forbidding battery, the "white-first" rule in chess, and the rule prescribing black dress at funerals. In violations of the rule forbidding battery, the AGENT BAD-HARM VICTIM stimulus tends to be intact. By contrast, in violations of the "white-first" rule in chess, the AGENT BAD-HARM VICTIM stimulus tends to be degraded beyond recognition. The rule prescribing black dress at funerals is somewhere intermediate between the first two rules: upon perceiving a person who does not wear black to a funeral, many people recover the AGENT BAD-HARM VICTIM stimulus, but many others

[50] *See, e.g.*, Marcin Szwed et al., *The Role of Invariant Line Junctions in Object and Visual Word Recognition*, 49 VISION RESEARCH 718, 719 (2009).

do not. This logic yields a simple prediction: if asked to arrange the preceding rules in moral-to-conventional order, most people would array them as follows: rule forbidding battery > rule prescribing black dress at funerals > "white-first" rule in chess.

The present hypothesis has several implications for how to describe and to understand the moral-conventional distinction. First, the moral-conventional distinction is fuzzy and plastic, not sharp or rigid, and very many rules are, like the rule about black dress at funerals, border-line cases. The fuzziness and plasticity of the moral-conventional distinction follow quite directly from the conceit that both morality and convention are most likely prototype-concepts (in research on the prototype theory of concepts, many authors[51] treat the vagueness of a concept as evidence that the concept has a prototype representation). If people represented morality and convention in terms of, for instance, individually necessary and jointly sufficient conditions, then one would observe far fewer border-line cases than one in fact does.

Secondly, with respect to the question of cultural variation, the moral-conventional distinction is "locally variable" but "globally uniform": cultures differ in how they classify rules as moral versus conventional, but every culture recognizes a moral-conventional distinction. For comparison, another example of a locally variable but globally uniform mental-behavioral phenomenon may be people's tendency to create *status hierarchies*.[52]

Thirdly, it is not helpful to conceive of rules as *inherently* moral or as *inherently* conventional. With the possible exception of a small number of rules (e.g., "do not cause wanton pain to in-group members"),[53] whether a rule is moral or conventional depends on how people

---

[51] *See, e.g.*, Benjamin Cohen & Gregory L. Murphy, *Models of Concepts*, 8 COGNITIVE SCIENCE 27, 30 (1984); James A. Hampton, *Concepts as Prototypes*, 46 PSYCHOLOGY OF LEARNING AND MOTIVATION 79 (2006); Edouard Machery, *Concepts: A Tutorial*, *in* CONCEPTS AND FUZZY LOGIC 13, 19 (Radim Belohlavek & George J. Klir eds., 2011).

[52] *See, e.g.*, Cameron Anderson et al., *Is the Desire for Status a Fundamental Human Motive? A Review of the Empirical Literature*, 141 PSYCHOLOGICAL BULLETIN 574, 592 (2015).

[53] *See, e.g.*, Christopher Boehm, *Purposive Social Selection and the Evolution of Human Altruism*, 42 CROSS-CULTURAL RESEARCH 319, 331 (2008).

tend to represent violations of it. This suggests that the moral-conventional distinction cannot be usefully understood as a distinction between two lists of rules, one moral and the other conventional. This simple point is worth emphasizing because some researchers[54] have argued or implied that the difficulty of producing uncontroversial lists of moral and of conventional rules undermines the moral-conventional distinction's empirical reality. From the perspective of the present hypothesis, this inference is mistaken, because the premise that people struggle to agree about which rules are moral and which are conventional does not support the conclusion that people do not distinguish between morality and convention. At any rate, it is more helpful to conceive of a rule's moral or conventional status as depending on how people represent violations of the rule. For a roughly comparable conclusion by a team of experimentalists, consider Francesco Margoni's and Luca Surian's statement that, in evaluating rule violations, people can easily "switch between a 'moral construal' . . . and a 'conventional construal.'"[55] Margoni and Surian reported that whether subjects perceive a rule violation as moral or as conventional can depend on whether they are considering (1) whether it would be acceptable for *the subjects themselves* to violate the rule or (2) whether it would be acceptable for *others* to do so.[56] This finding generalizes. Whether people perceive a given rule violation as moral or as conventional can also depend on, for instance, what Cecilia Wainryb called people's "informational

[54] *See, e.g.*, Mordecai Nisan, *Moral Norms and Social Conventions: A Cross-Cultural Comparison*, 23 DEVELOPMENTAL PSYCHOLOGY 719, 724 (1987); Tage Shakti Rai & Alan Page Fiske, *Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality*, 118 PSYCHOLOGICAL REVIEW 57, 58–59 (2011). For one of Turiel's responses to this line of argument, see Elliot Turiel et al., *A Cross-Cultural Comparison About What? A Critique of Nisan's (1987) Study of Morality and Convention*, 24 DEVELOPMENTAL PSYCHOLOGY 140 (1988).

[55] Francesco Margoni & Luca Surian, *Question Framing Effects and the Processing of the Moral-Conventional Distinction*, 34 PHILOSOPHICAL PSYCHOLOGY 76, 94 (2021).

[56] *See id.* at 92–93.

assumptions" (causal and factual beliefs).[57] For example, a person who is unacquainted (acquainted) with the germ theory of disease may perceive a violation of the rule forbidding spitting in public as conventional (moral).[58]

It is now possible meaningfully to elaborate on the three positions that the Introduction said this essay would defend. (1) The moral-conventional distinction is due to the fact that people tend to respond differently to rule violations that trigger the perception of an agent bad-harming a victim versus rule violations that do not have this property. (2) It is best to describe the moral-conventional distinction in terms of harm because it is the mental representation of a certain type of harm—bad-harm—that causes people to delineate a sub-set of rules as *moral* rules. (3) In describing the moral-conventional distinction in terms of harm, it is helpful to define the relevant notion of harm—bad-harm—as a semantic prime that resists verbal definition but exhibits typicality effects. This notion of harm can be useful to researchers who study the moral-conventional distinction or moral psychology more generally. For example, it can help to account for why people often struggle to describe the border between morality and convention (because "bad-harm" resists verbal definition), as well as for why people can often agree about what types of events qualify as important instances of morally relevant harm (because "bad-harm" exhibits typicality effects).

The foregoing is not, of course, the only plausible account of the moral-conventional distinction. That said, I hope that even readers who reject this essay's account will have found it valuable, if only for showing how a particular harm-based understanding of moral cognition can explain and unify many of the experimental and theoretical results that researchers studying the moral-conventional distinction have recently reported.

---

[57] *See* Cecilia Wainryb, *Understanding Differences in Moral Judgments: The Role of Informational Assumptions*, 62 CHILD DEVELOPMENT 840 (1991).

[58] For one study of people's perceptions of the morality of public spitting, see Shaun Nichols, *Norms With Feeling: Towards a Psychological Account of Moral Judgment*, 84 COGNITION 221, 227–32 (2002).

* * *

Readers may also be interested to note that a harm-based account of the moral-conventional distinction such as this essay's yields unique responses to several critics of Turiel's foundational formulation of the distinction. Consider two examples.

As Part I above noted, Turiel reported that harmfulness is one of the dimensions along which people distinguish between morality and convention. Jonathan Haidt, Silvia Helena Roller, and Maria Dias set out to challenge Turiel by reporting an experiment in which some subjects perceive rule violations as moral violations despite not perceiving them as harmful. Haidt et al. presented subjects with what the researchers characterized as "harmless" actions, such as: (1) a woman using her country's flag to clean her bathroom; (2) two siblings kissing; (3) a son promising his dying mother that he will visit her grave and then failing to do so.[59] After reporting that some subjects perceived these putatively harmless actions as morally wrong, Haidt et al. concluded that, *contra* Turiel, some people hold a "non-harm-based morality."[60] Although Haidt et al. did not try to define "harm," it is clear from their phrasing that they understood it in an objective, mind-independent sense.[61] They thus presumed that there is no harm in (e.g.) a son's breaking a promise to his dead mother. However, once one understands harm not in a mind-independent sense but as a mental representation, it becomes easy to see that, *contra* Haidt et al., people can and often would perceive the son's behavior as a harm to the mother. Indeed, although

[59] Jonathan Haidt et al., *Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?*, 65 JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 613, 617 (1993).

[60] *E.g.*, *id.* at 626.

[61] *See, e.g.*, *id.* at 613 (approving the view that an action is harmless if it "violate[s] no interests of others").

Haidt continues to hold that his vignettes describe "harmless taboo violations,"[62] even a third of Haidt et al.'s subjects explicitly said that the son's behavior involved a discrete harm.[63]

In a different criticism, Daniel Kelly and co-authors interpreted Turiel as arguing that HSAG exhibit a "law-like tendency" to co-occur.[64] Kelly et al. thus set out to challenge Turiel by reporting an experiment in which this tendency fails to obtain. In one of their vignettes, for instance, Kelly et al. sought to show that *harmfulness* and *authority-independence* can fail to co-occur. In the relevant vignette, a military sergeant conducts survival training for special forces trainees by putting them through a simulated interrogation during which, as part of the training, they are physically abused.[65] In the course of this training, the trainees "often end[] up with bruises or injuries that last[] for a week or more."[66] Kelly et al. presented subjects with two versions of this vignette (paraphrased in V_1 and V_2 below) and then asked the subjects to state whether the sergeant's behavior was "okay."[67]

> V_1 (prohibited by an authority): High command has just prohibited physical abuse in survival training, and the sergeant's superior has ordered him to obey high command. However, in violation of orders, the sergeant is continuing to expose his trainees to physical abuse in the course of survival training.
>
> - 91% of subjects said that the behavior was "not okay."

[62] Jonathan Haidt & Fredrik Bjorklund, *Social Intuitionists Answer Six Questions about Moral Psychology*, *in* MORAL PSYCHOLOGY: THE COGNITIVE SCIENCE OF MORALITY: INTUITION AND DIVERSITY 181, 196–97 (Walter Sinnott-Armstrong ed., 2008).

[63] *See* Jonathan Haidt et al., *Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?*, 65 JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 613, 618, 620 (1993).

[64] Daniel Kelly et al., *Harm, Affect, and the Moral/Conventional Distinction*, 22 MIND AND LANGUAGE 117, 119 (2007).

[65] *See id.* at 125.

[66] *Id.*

[67] *See id.* at 122.

> V_2 (not prohibited by an authority): There are no such orders, and the sergeant's superior has allowed him to use physical abuse in the course of survival training. The sergeant is exposing his trainees to physical abuse in the course of survival training.
>
> - 42% of subjects said that the behavior was "not okay."[68]

Kelly et al. interpreted this response pattern as follows: the sergeant's behavior in both V_1 and V_2 was "clearly harmful,"[69] but many subjects' evaluations of whether this behavior was "okay" depended on whether it was prohibited by an authority. Thus, *harmfulness* and *authority-independence* can fail, Kelly et al. concluded, to co-occur.

From the perspective of this essay's account of the moral-conventional distinction, a few nuances are relevant in properly evaluating Kelly et al.'s criticism of Turiel. First, whether exposing special forces trainees to physical abuse in the course of survival training is a harm is a question about which many people disagree; this behavior is difficult accurately to characterize as a "clear harm." In fact, one may very reasonably believe that *refraining from* using physical abuse in the course of survival training harms special forces trainees by decreasing the probability that they will survive if they are captured in real combat. Secondly, some people perceive a sergeant's violation of orders as a harm (e.g., to the effectiveness of the military organization), so, from the perspective of harmfulness, V_1 and V_2 are not the same. Thirdly, some people treat the fact that a behavior is prohibited by an authority as evidence that the behavior is a harm, as when a citizen infers that, because a person violated one or another law or ordinance, the person is morally blameworthy. Taking account of these nuances suggests that Kelly et al.'s result does not necessarily undermine Turiel's core conclusion that most people distinguish between morality and convention. Consider the following interpretation. In V_2, where the violent survival training was not banned, many subjects did not perceive the sergeant's behavior as harmful, and they thus

[68] For Kelly et al.'s own wording, see *id.* at 125. For the percentages, see *id.* at 127.

[69] *Id.* at 124.

said that the sergeant's behavior was "okay." But in V_1, where the training was banned, and where the sergeant's decision nonetheless to continue it constituted a violation of orders, many of the subjects that had not perceived the sergeant's behavior as harmful in V_2 now did perceive it as harmful.

Conclusion

Research both older and more recent suggests that people's tendency to distinguish between morality and convention depends on their tendency to react in a particular way to the perception of harm. However, partly because defining "harm" is not straightforward, researchers have proposed disparate ways of understanding harm for purposes of describing the moral-conventional distinction. This essay has presented an updated harm-based account of the moral-conventional distinction. On this account, in contrast to conventional rules, moral rules are those the violation of which tends to produce in observers the mental representation of one type of entity bad-harming another type, where "bad-harm"—which demarcates the sub-set of set-backs to interest that people perceive as morally relevant—is a semantic prime that resists verbal definition but exhibits typicality effects. This account efficiently explains why the moral-conventional distinction is simultaneously pan-cultural and culturally variant, better grounds the study of the moral-conventional distinction in the computational theory of mind, and identifies additional ways in which Elliot Turiel's foundational research can absorb or rebut criticisms.