Voiklis, J., Cusimano, C., & Malle, B. F. (2014). A social-conceptual map of moral criticism. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1700-1705). AustinTX: Cognitive Science Society.

A Social-Conceptual Map of Moral Criticism

John Voiklis (john_joiklis@brown.edu)
Corey Cusimano (corey_cusimano@brown.edu)
Bertram Malle (bertram_malle@brown.edu)
Brown University, Department of Cognitive, Linguistic & Psychological Sciences
Box 1821, 190 Thayer Street, Providence, RI 02912 USA

Abstract

Moral criticism is both a social act and the result of complex cognitive and conceptual processes. We demonstrate consensual features of various acts of moral criticism and locate them within a higher-order feature space. People showed consensus in judging 28 verbs of moral criticism on 10 features, and the judgment patterns formed a two-dimensional space, defined by an intensity axis and an interpersonal engagement axis. Subsets of verbs formed well-defined clusters roughly corresponding to the four quadrants of this space. The marker verbs of these clusters were *lashing out* (intense, public), *pointing the finger* (mild, public), *vilifying* (intense, private), and *disapproving* (mild, private).

Keywords: social cognition; moral psychology; verb semantics

Introduction

What is the core of moral psychology? The literature focuses on judgment, reasoning, and emotion—all internal states in the social-moral perceiver. But if morality were nothing more than a complex private state then it would be wholly puzzling how it could serve its major function—to regulate human behavior. So an obvious, but understudied, question emerges: how social acts of moral criticism do the work of regulating behavior. A "how" question naturally subsumes many sub-questions, of which we address two: First, what is the space of moral criticism—that is, what are the candidate acts of moral criticism that allow people to regulate others' behavior (should one blame the other, admonish, scold, or castigate him)? Second, along which dimensions do these acts of moral criticism differ? Is it their emotionality, deliberation, their persuasive potential?

Our approach takes seriously that almost all acts of moral criticism are expressed in language; we therefore begin our investigation by charting the social-conceptual space of moral criticism: verbs that depict such criticism and the rich implications they carry (Fillmore, 1971). Imagine overhearing someone say "He admonished her for what she had done." This sentence, though seemingly cryptic, still carries a lot of implied meaning. The choice of the verb *admonished*—instead of, for example, *berated*—suggests a certain intensity of the moral critic's emotion, hints at his relative status, intimates the severity of her bad behavior, and conveys a certain likelihood that the two will discuss and reconcile their differences.

In this contribution we do not aim to document the full set of implied meanings of every possible verb of moral criticism. Instead, we select a subset of implied features and test them on a representative sample of verbs. This way we hope to chart out a map of moral criticism, even if not all points of interest will be filled in.

We will proceed as follows. First, we derive a number of candidate features of moral criticism from previous psychological and sociological work (Drew, 1998; Laforest, 2009; Malle, Guglielmo, & Monroe, 2014). We then select a set of verbs that depict acts of moral criticism and populate the feature space of moral criticism. Next, we empirically test how well these features explain the variability among the set of selected verbs. Finally, we integrate our results into a preliminary model of the space of moral criticism.

Deriving Features of Moral Criticism

Blame, writes Beardsley (1979, p. 176), "has a power and poignancy for human life unparalleled by other moral concepts." Indeed, blame is arguably the paradigmatic moral criticism. Therefore, some of the fundamental properties of blame (Malle et al., 2014) can guide us in deriving features of moral criticism more generally.

The features we studied can be grouped into three categories (see Table 1):

1. Features of the social act
2. Features of the underlying judgment
3. Semantic landmarks

Features of the social act. Evolved instincts for belonging, caring, and shared experience do some of the work of motivating individuals to behave in ways that sustain social relations (Churchland, 2012; Deigh, 1996; Joyce, 2006; Rai & Fiske, 2011). However, complex social life would be impossible without norms and values for sharing and reciprocity, self-control, and mutual recognition of rights—all of which keep an individual's behavior in line with community interests (Sripada & Stich, 2006; Sunstein, 1996). This kind of cultural morality has to be taught, learned, and enforced by community members. Praising and blaming people for their behaviors, and occasionally punishing them, enforces the norms and values of cultural morality (Cushman, 2013).

This social-regulatory property of blame requires that a moral critic actually perform a public and communicative act (FEATURE: *Public Act* in Table 1). Several other features follow from this public communication.

Moral criticism varies by whether the critic directly addresses the norm violator or talks to others about the norm violator (FEATURE: *2nd- vs. 3rd-Person*) (Dersley & Wootton, 2000; Laforest, 2009; Traverso, 2009).

Table 1: Assessed features of moral criticism and their formulations in the study.

Features of the Social Act	
Public Act	Was this more like a private thought or more like a public action?
2nd vs. 3rd Person	Did he act directly toward her or did he express this to other people?
Conversation	How likely is it that, right after, he and she will talk about what happened?
Reform	Given his action, how likely is she to improve her future behavior?
Features of the Underlying Judgment	
Offense Severity	How bad was what she [the offender] had done?
Thoughtfulness	How thoughtful or impulsive was he in doing that?
Emotionality	How intense was the emotion he felt?
Acceptability	How socially acceptable was what he [moral critic] did?
Semantic Landmarks	
Like *blame*	How similar in meaning is this [statement] to "He blamed her for the bad thing she had done."?
Like *punish*	How similar in meaning is this [statement] to "He punished her for the bad thing she had done"?

If the moral critic does directly address the norm violator, behavior regulation varies as a function of whether the criticism invites further conversation (FEATURE: *Conversation*) about the norm violation and its possible repair (McGeer, 2012; McKenna, 2012; Newell & Stutman, 1991).

Finally, in response to the moral criticism violators may intend to repair the damage to their social standing and promise adherence to the violated norms in the future (FEATURE: *Reform*) (Bennett, 2002; Walker, 2006).

Features of the underlying judgment. Moral criticism carries at least temporary costs for the offender, be they emotional or social (Bennett, 2002; McKenna, 2011). Imposing such costs on another community member demands warrant: the blamer must be able to offer grounds for his or her act of blaming (Bergmann, 1998; Coates & Tognazzini, 2012). Warrant lies, first and foremost, in the offending act. Just as the law has a proportionality principle (Engle, 2012) we expect that people's moral criticism is finely attuned to the severity of the offense (FEATURE: *Offense Severity*) (Fillmore, 1971).

In addition, a number of social-cognitive assessments normally enter blame (e.g., of intentionality, reasons, knowledge; Cushman, 2008; Guglielmo, Monroe, & Malle, 2009; Shaver, 1985). If blame is grounded in such careful assessments, moral criticism may be considered thoughtful rather than impulsive (FEATURE: *Thoughtfulness*).

Despite significant social-cognitive work, moral judgment is often accompanied by affective states, from simple disapproving feelings to more complex states of indignation or outrage (Alicke, 2000; Prinz, 2006). Social acts of criticism may then express this affective tone to different degrees (FEATURE: *Emotionality*).

Thoughtfulness and emotional intensity, arising from the judgment relative to the severity of the offense, may determine how appropriate particular degrees of moral criticism are—ranging between mildly disapproving of and chastising the offender. In dealing with interpersonal criticism and complaints, people welcome thoughtful, clear, and constructive criticism whereas they dislike yelling and personal attacks as expressions of disapproval (Alberts, 1989). Especially when publicly expressed, moral criticism thus is likely to vary in its degree of social acceptability (FEATURE: *Acceptability*).

Semantic landmarks. Having derived a number of features from a theory of blame, we decided to treat *blame* as one of the landmarks in the space of moral criticism, comparing all other moral action verbs to blame. As perhaps the superordinate term of moral criticism, blame may well occupy a center spot in the dimensions of *Thoughtfulness* and *Emotionality*, summed into *Acceptability*, and appear both as a private thought and *Public Act*, expressed in *2nd*- and *3rd-person* communication.

In addition, because blame is often equated with or treated as parallel to punishment, we added acts of *punishing* the offender as the second landmark. Acts akin to punishment, we can expect, will more often follow from *Severe Offenses*, be accompanied by more *Emotionality*, and leave less room for *Conversation* with the offender.

Hypotheses. If these derived features of moral judgment and moral communication help characterize the greater social-conceptual space of moral criticism[1], we should expect representative speakers of a given language to be able to assess acts of moral criticism relative to these features. Overhearing someone say "He [admonished, berated, rebuked, etc.] her for the bad thing she had done" should easily allow the person to rate the moral act for each of the eight features: whether the specific verb (e.g., *rebuke*) implies a private thought or public action, whether the action was addressed to the violator or some third person, whether the episode permitted continued communication, and so on. Likewise, if not just lexicons but speakers of a language have a differentiated vocabulary of moral criticism, they should easily assess how similar each considered moral verb is to acts of blaming and acts of punishment. Most important, if the conceptual space of moral criticism is truly a reflection of ordinary social-moral practice, and not just the fiction of cognitive scientists, a considerable degree of consensus must exist among people making such assessments.

[1] There are other candidate features of moral criticism we have not yet examined in detail, including role, relationship, and context. We reserve these complex features for a future study.

Verbs of moral criticism. We tested these hypotheses by asking participants to make inferences about twenty-eight stimulus verbs that denote varieties of moral criticism (henceforth MC verbs). We selected these verbs with the goal to fully represent the social-conceptual space of moral criticism, while avoiding rare and obsolete verbs. Roget's Thesaurus provided the initial set of forty verbs; these included synonyms of the verb "blame" and their own synonyms. The WordNet database corroborated a subset of these synonyms. The online edition of the Oxford English Dictionary provided one filter; based on the definitions and contextual sentences, we excluded verbs with too many unrelated meanings, as well as verbs outside of current and common usage. The online database for the Corpus of Contemporary American English (COCA) provided another filter. This corpus contains more than 425 million words of text from spoken dialogue, fiction, popular magazines, newspapers, and academic texts that appeared between 1990 and 2012. Based on the COCA frequency counts for each verb in its various tense forms (e.g. blame, blames, blamed, blaming) and contextual sentences, we excluded infrequent verbs and those used mainly in the passive voice. The remaining twenty-eight verbs appear in Table 2.

Table 2: Commonly used verbs denoting moral criticism.

Verb	Frequency	Verb	Frequency
accuse	2181	find fault with	182
admonish	2614	lash out at	515
attack	1091	let X have it	333
berate	706	object to	1961
blame	12770	point the finger	282
castigate	397	rebuke	656
censure	498	reprimand	629
chastise	809	reproach	575
chew out	100	revile	419
chide	959	scold	405
condemn	2119	slander	496
criticize	3439	tell X off	152
denounce	957	vilify	633
disapprove	785		

Method

Materials

Inference Probes. The 28 MC verbs were inserted into the [verbed] placeholder in the sentence, "He [verbed] her for the bad thing she had done." To probe inferences about these verb phrases, we developed questions to which participants responded on a seven-point rating scale. (They could also respond "I don't know this word/phrase.") The questions for each feature are displayed in Table 1. Importantly, any given participant responded to only one probe, for all 28 verbs. This way, any correlations between features are not driven by participants' hypotheses or response biases, but by the consensual rating profiles that independent samples of participants produced.

Participants and Procedure

We recruited 300 fluent English speakers (female = 164, and unreported = 2; mean age approximately 32 years) through Amazon's Mechanical Turk. Participants were assigned to one of ten groups (each N = 30), in which they repeatedly responded to the same inferential question for each of the 28 MC verbs. Each group of participants was redirected to a different page in an external Web application. There, all participants completed an English competency test (a sentence-completion test based on a standard eighth-grade literary text) and provided the reported demographic information. Participants in each group then received condition-specific instructions. For example, the *Emotionality* inference condition was introduced as follows (see supplemental materials Web page for variations used in other conditions):

Please read carefully!

You will read a number of sentences that are related in meaning to "He blamed her for what she had done." (We never specify exactly what she had done; just assume it was something bad.)

We are interested in the way people interpret these kinds of sentences. In particular, we'd like to know how much emotional intensity is implied by these sentences (e.g., that "He blamed her").

Some sentences may imply that the agent (= "He") felt an intense emotion; some may imply that he felt no emotion. It will depend on the specific words in the sentence. Please read each sentence, then indicate on the rating scale how intense the emotion was that he felt. We will always ask you about his emotion.

Please do this task from memory -- do not look up any of the words or phrases. It's okay if there are some words you are not familiar with.

In two training trials participants answered their condition-specific question from Table 1 (e.g., for *Emotionality*, "How intense was the emotion he felt?") for two verbs: "He [yelled at / spoke out against] her for the bad thing she had done." They then proceeded to answer the same question for the 28 MC verbs on 7-point scales, with condition-appropriate anchors (for *Emotionality,* "Not at all" to "Extremely intense").

Results

We organize the results around our two goals: to examine whether people show consensus on the selected features of moral criticism, and to characterize the social-conceptual space of moral criticism spanned by those features.

Consensus About the Features of Moral Criticism

To assess participant consensus about the feature values they inferred from the range of verbs we used Cronbach's standardized coefficient α and the average correlations of each rater with the group $\bar{r}_{i*N}$ (see Table 3). The results support show that people strongly agree on how the 28 verbs of moral criticism are arranged along each feature dimension.

Table 3: Social consensus of feature inferences

	α	$\bar{r}_{i*N}$
Features of Social Act		
Public Act	0.94	0.58
2nd vs. 3rd Person	0.96	0.63
Conversation	0.84	0.36
Reform	0.71	0.21
Features of Underlying Judgment		
Offense Severity	0.93	0.53
Thoughtfulness	0.94	0.59
Emotionality	0.95	0.63
Social Acceptability	0.97	0.79
Semantic Landmarks		
Like Blame	0.91	0.51
Like Punish	0.95	0.63

The Feature Space of Moral Criticism

We had identified four features of moral criticism as social acts and four features of underlying moral judgments, complemented by two semantic landmarks. If this set of 10 variables at least partially constitutes the social-conceptual space of moral criticism we should be able to recover dimensions of this space from a principal components analysis of the verb × feature correlation matrix. That is, we are looking to capture the higher-order properties that account for systematic differences among MC verbs.

Table 4: Principal components of inferred features

Features	PC 1	PC 2	PC 3
Emotionality	0.91		
Acceptable	-0.90		
Severity	0.87		
Conversation	-0.78	0.43	
Punish	0.69	0.47	0.43
Blame	-0.44		
Public		0.84	
ThirdPers		-0.80	-0.30
Thoughtful	-0.50	-0.69	
Reform			0.92

Note: Varimax rotated solution; loadings < 0.30 not shown

Three components accounted for 78% of the variance. We see in Table 4 and Figure 1 that the first and strongest component unites the four judgment features with the similarity to punishment. We label this the "Intensity" dimension of moral criticism, anchored on the high end by acts that respond to severe offenses and come with strong emotions and on the opposing end by acts that are socially acceptable and have potential for further conversation. The second component captures the interpersonal nature of acts of criticism—whether they are public and directed at the offender or more private, perhaps even just in thought.

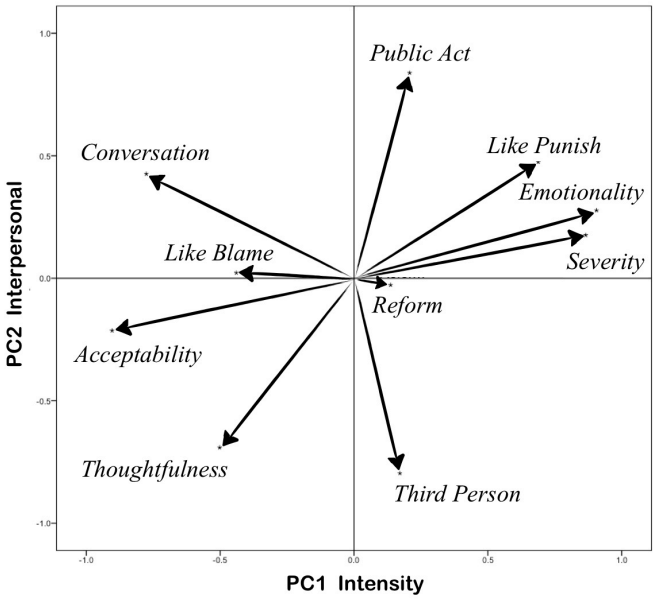

Figure 1. How inferred features constitute the first two principal components of the moral criticism feature space.

Two features seem rather distinct from the rest: *Reform* formed its own component, perhaps because it was too difficult a judgment to make without information about the offender or the specific offending act. In addition, the semantic landmark *Like Blame* loaded modestly on the Intensity component but would have formed its own component in a four-factor solution. Perhaps features such as status, role, and relationship (which were not assessed here), determine more precisely whether an act of moral criticism is "like blame."

Kinds of Moral Criticism in the Feature Space

Now that we have established a feature space that is primarily defined by judgment intensity and interpersonal address, we can plot the 28 MC verbs within this feature space and subject the underlying "factor scores" to clustering algorithms (Reynolds, Richards, Iglesia, & Rayward-Smith, 2006). Four clusters emerged repeatedly when partitioning around four to six medoids[2], and they corresponded well to the four quadrants defined by the two-component feature space (see Figure 2): intense acts to the person's face (attack, lash out, berate), intense acts in the person's absence (revile, vilify, slander), milder acts to the person's face (accuse, criticize, blame), and milder acts in the person's absence (disapprove, find fault, object to). The remaining verbs, near the center of the plot, made up a large residual cluster, exhibiting no clear differentiation given the current set of features. Once more, additional features such as context, role, or relationship might provide further differentiation.

[2] Medoids are the set of MC verbs that minimized the sum of the dissimilarities between each verb and its closest representative verb.

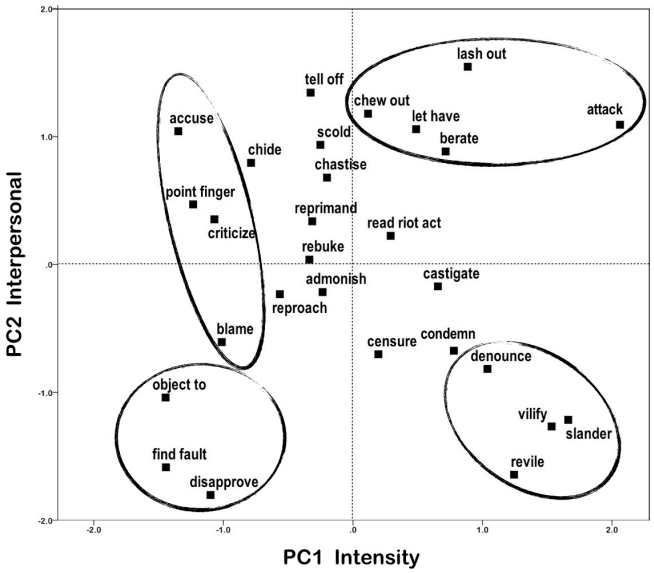

Figure 2. Verbs of moral criticism plotted within the first two principal components of the feature space.

Discussion

In the present paper, we report on our progress towards two sub-goals of understanding how social acts of moral criticism do the work of regulating behavior: discovering consensual features of such acts and locating the acts within a higher-order feature space of moral criticism. We isolated four reliable features of moral criticism as a social act, four reliable features of the judgments underlying moral criticism, and two semantic landmarks of moral criticism. A principal components analysis on these 10 variables (applied to the 28 verbs of moral criticism) yielded two major dimensions of morally critical acts: their intensity and their interpersonal engagement.

Within this two-dimensional feature space we then mapped the various verbs of moral criticism and identified four clusters with well-differentiated feature patterns. The clustered verbs formed roughly a 2 × 2 classification of prototypes of moral criticism: intense vs. mild criticism that is either directed publicly at the offender or kept largely private. The marker verbs (medoids) of these prototypes are *lashing out* (intense, public), *vilifying* (intense, private), *pointing the finger* (mild, public), and *disapproving* (mild, private).

Implications for moral psychology. The results reported here represent an initial step towards bridging the gap between the internal judgments of moral perceivers and their regulatory influence over other people's behavior. We have provided an existence proof for an accessible and coherent representation of social acts of moral criticism. Future research can build on this foundation and examine in more detail whether the different types of moral criticism are grounded in distinct information processing. For example, is *lashing out at* someone not only more intensely expressed, but also based on more sloppy information processing? Do public acts of moral criticism come with stronger warrant because the costs of false accusations are higher than in private condemnation? Finally, does the two-dimensionality of social acts of moral criticism also apply to nonverbal expressions? Do certain gestures and facial expressions "code for" the social prototypes we have identified? How exactly does a social perceiver of a certain gesture detect that another person is not just *blaming* an offender, but actually *denouncing* him?

Implications for artificial moral agents. Whatever the promise of studying nonverbal moral expressions, language remains the dominant interface for social moral interactions in human communities. Our present study is part of a larger project on the necessary ingredients of a "moral robot," and the verbal channel of expression will also be paramount for any near-term artificial moral agent. The moral competence of a robotic car, for example, will have to involve producing and comprehending verbal exchanges with passengers about norm-violating behaviors by the operator and other drivers. A robotic partner in police patrolling will have to be able to detect potential norm violations and express those observations to its human partner. A search-and-rescue robot that needs to decide whether to save a crying baby or a screaming adult would ideally later explain to its supervisor why it made its decision; and depending on those "reasons," the supervisor may need to make an adjustment to the robot's "value system." Learning verbs of moral criticism may be a small ingredient in designing such a morally competent robot, but understanding and representing the underlying social-conceptual space of such verbs will be essential.

Conclusion

The present research illustrates, we hope, the strength of a broad cognitive-science approach to morality: using tools and ideas from linguistics, psychology, and statistics to consider simultaneously both mental and social processes, both how people talk about the world and how they conceptualize the world, and both as individual information processors and as members of complex social communities.

Acknowledgments

This project was supported by a grant from the Office of Naval Research, No. N00014-13-1-0269. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

References

Alberts, J. K. (1989). A descriptive taxonomy of couples' complaint interactions. *Southern Communication Journal*, *54*, 125–143.

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

Beardsley, E. L. (1979). Blaming. *Philosophia*, *8*, 573–583.

Bennett, C. (2002). The varieties of retributive experience. *The Philosophical Quarterly*, *52*, 145–163.

Bergmann, J. R. (1998). Introduction: Morality in discourse. *Research on Language & Social Interaction*, *31*, 279–294.

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.

Churchland, P. S. (2012). *Braintrust: What neuroscience tells us about morality*. Princeton, NJ: Princeton University Press.

Coates, D. J., & Tognazzini, N. A. (2012). The contours of blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 3–26). New York, NY: Oxford University Press.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Cushman, F. (2013). The functional design of punishment and the psychology of learning. In R. Joyce, K. Sterelny, B. Calcott, & B. Fraser (Eds.), *Psychological and environmental foundations of cooperation.*, Signaling, commitment and emotion (Vol. 2). Cambridge, MA: MIT Press.

Deigh, Jo. (1996). Morality and personal relations. *The sources of moral agency* (pp. 1–17). New York, NY: Cambridge University Press.

Dersley, I., & Wootton, A. (2000). Complaint sequences within antagonistic argument. *Research on Language and Social Interaction*, *33*, 375–406.

Drew, P. (1998). Complaints about transgressions and misconduct. *Research on Language & Social Interaction*, *31*, 295–325.

Engle, E. (2012). The history of the general principle of proportionality: An overview. *Dartmouth Law Journal*, *10*, 1–11.

Fillmore, C. J. (1971). Verbs of judging: An exercise in semantic description. In C. Fillmore J. & D. T. Langendoen (Eds.), *Studies in Linguistic Semantics* (pp. 272–289). New York, NY: Holt, Rinehard and Winston.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Philosophy*, *52*, 449–466.

Joyce, R. (2006). *The evolution of morality*. MIT Press.

Laforest, M. (2009). Complaining in front of a witness: Aspects of blaming others for their behaviour in multi-party family interactions. *Journal of Pragmatics*, *41*, 2452–2464.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*.

McGeer, V. (2012). Civilizing blame. *Blame: Its nature and norms* (pp. 162–188). New York, NY: Oxford University Press.

McKenna, M. (2011). *Conversation and responsibility*. New York, NY: Oxford University Press.

McKenna, M. (2012). Directed blame and conversation. *Blame: Its nature and norms* (pp. 119–140). New York, NY: Oxford University Press.

Newell, S. E., & Stutman, R. K. (1991). The episodic nature of social confrontation. In J. A. Anderson (Ed.), *Communication yearbook* (Vol. 14, pp. 359–413). Thousand Oaks, CA: Sage.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*, 29–43.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*, 57–75.

Reynolds, A. P., Richards, G., Iglesia, B. de la, & Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*, 475–504.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer Verlag.

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind (Volume 2: Culture and cognition)* (pp. 280–301). New York, NY: Oxford University Press.

Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, *96*, 903–968.

Traverso, V. (2009). The dilemmas of third-party complaints in conversation between friends. *Journal of Pragmatics*, *41*, 2385–2399.

Walker, M. U. (2006). *Moral repair: Reconstructing moral relations after wrongdoing*. New York, NY: Cambridge University Press.