Moral Cognition and its Basis in Social Cognition and Social Regulation

John Voiklis and Bertram F. Malle
Brown University

Abstract

What is the basis of moral cognition? Moral judgments are grounded in a number of cognitive and social-cognitive processes, which guide the social regulation of behavior and are, in turn, constrained by such regulation to be fair and evidence-based.

1. Introduction

Human beings live complex social lives, composed of various types of relationships across nested social hierarchies, all structured by rights, rules, and obligations. However, selfish goals persist, and keeping individuals' goals in line with community interests has become the primary challenge of modern morality. To meet this challenge human societies have developed two major social-cultural tools: a vast network of rules, norms, and values (Sripada & Stich, 2006; Ullmann-Margalit, 1977) and complex social practices of norm enforcement, such as blame, praise, apology, and reconciliation (Semin & Manstead, 1983).

This kind of social-cultural morality has to be taught, learned, and enforced by community members, even by the youngest among them (Göckeritz, Schmidt, & Tomasello, 2014). Acquiring norms likely benefits from early appearing preferences for prosocial agents over antisocial agents (Hamlin, 2014), but socially mature moral capacities rely heavily on nonmoral capacities: those of *social cognition* (Guglielmo, Monroe, & Malle, 2009).

Social cognition encompasses a hierarchy of interdependent concepts, processes, and skills that allow individuals to perceive, understand, and—most important for the present topic—evaluate one another. For example, norm enforcers infer the mental processes that generated a transgressive behavior (e.g., motive, belief, intention) before blaming the transgressor (Malle, Guglielmo, & Monroe, 2014). The transgressor must likewise infer the mental processes that generated the social act of blaming (e.g., the norm enforcer's goals, knowledge, and power of enforcing sanctions) when deciding to deny or admit, maintain or correct the norm-violating behavior.

These, then, are the phenomena this chapter aims to illuminate. We show how the elements of *social cognition* ground people's *moral cognition* and how social and moral cognition together guide the *social regulation of behavior by moral norms*. We aim to identify the concepts, mechanisms, and practices that go into forming various kinds of moral judgments, and the forms and functions of socially expressing those judgments.

2. Historical Context

The most prominent debates in moral philosophy grapple with dichotomies. Perhaps the oldest of these concerns the relative influence of *reason* and *passion* on human behavior (Hume, 1998; Kant, 2012). Moral *psychology*, too, has been heavily influenced by this dichotomy. During an early phase, scholars expressed great confidence in the human capacity to reason about moral matters—albeit a capacity that needs time to develop (Piaget, 1932; Kohlberg, 1981). During a later phase, scholars expressed sometimes fierce skepticism toward such reasoning capacities and offered emphatic claims about the primacy of affect in moral judgment (Alicke, 2000; Greene, 2008), about people's inability to access the cognitive basis of their judgments (Nisbett & Wilson, 1977; Haidt, 2001), and about the many biases from which these judgment suffer (Ditto, Pizarro, & Tannenbaum, 2009).

Dichotomies often suffer from exaggerations and simplifications. We hope to present a framework that goes beyond extreme positions and relies instead on theoretical analysis, existing empirical evidence, and predictions about new phenomena. We believe that drawing a line and designating two opposing sides—reason vs. passion, cognition vs. emotion, deliberation vs. intuition—is an unproductive way to tackle a multi-faceted phenomenon. We should rather survey the landscape and acknowledge the complex terrain of social life so as to discover the different psychological adaptations and social practices that have allowed people to navigate the terrain—imperfectly, but not as bumbling and blundering as they are sometimes portrayed. What enables such adaptive navigation, we will try to show, is the interactive system of moral cognition, social cognition, and social regulation.

This system is schematically illustrated in, in which the parts both inform one another (e.g., mental state inferences informing a blame judgment) but also justify one another (e.g., a wrongness judgment providing justification for an act of social regulation).

Two aspects of this schematic deserve comment. First, when we use the term "social cognition," we are not pitching our tent on the "reason" side of a dichotomy but rather conceive of social cognition as a large toolbox that contains both fast, automatic and slow, controlled mechanisms, intuition and deliberation, affect and thought. Second, whereas the relationship between social cognition and moral cognition has been discussed before (e.g., Shaver, 1985; Weiner, 1995), the appeal to *social regulatory* mechanisms of morality goes beyond the parameters of existing debates. Highlighting the social function of moral cognition can reveal much about how that process operates. More on both of these issues shortly.

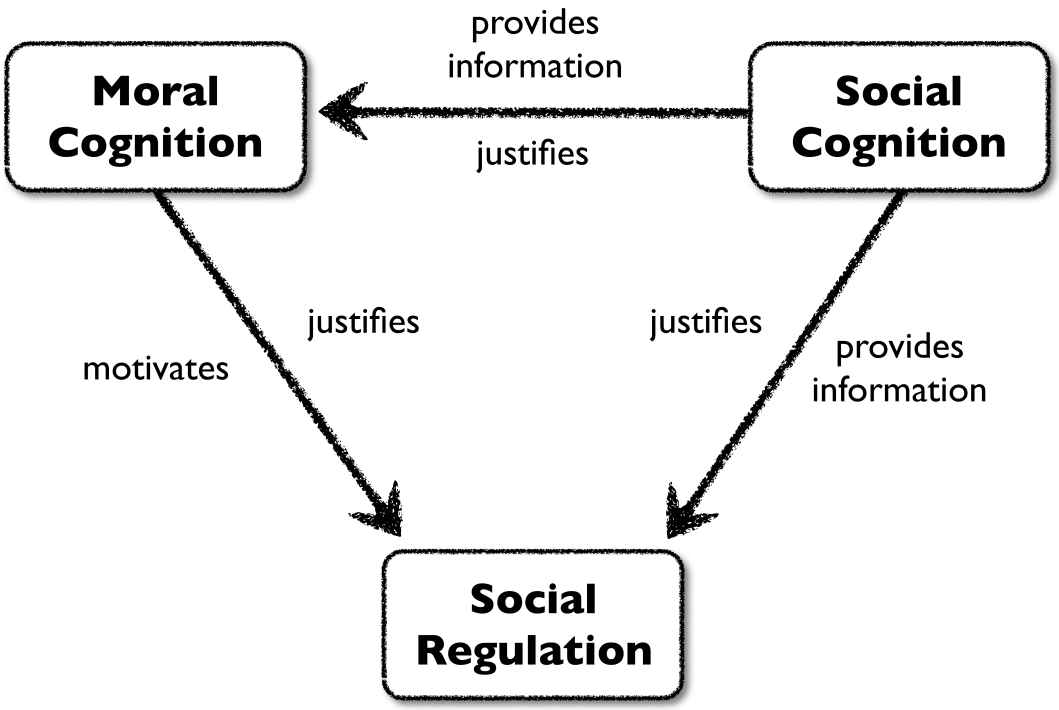

Figure 1. Schematic relationship between social cognition, moral cognition, and social regulation.

3. Theoretical Framework

To achieve the promised survey of the moral landscape, and to elucidate how moral cognition, social cognition, and social regulation are related, we address three questions: What makes moral cognition moral? What social-cognitive processes are involved in it? And how does moral cognition interact with social regulation?

What makes moral cognition moral?

The moral domain is that of regulating individual behavior in the context of community interests. Rules, norms, and values set the standards that, if fulfilled, serve community goals and make the community and its individuals succeed. The broader literature sometimes distinguishes moral from conventional rules or moral from social norms (Brennan, Eriksson, Goodin, & Southwood, 2013; Kohlberg, 1981). But for many purposes, it is best to assume a continuum of norms—

defined as standards and instructions that guide people in what they should do. We can then identify prototypes at each end. Moral norms, on one end, are part of a hierarchy in which moral "principles" and "values" are the most abstract instructions; social-conventional norms, at the other end, can often stand alone in regulating just one particular behavior or in solving a coordination problem. What the elements of this continuum have in common is that, in representing an instruction as a *norm* (as opposed to a goal or habit), people keenly take into account that (a) a sufficient number of individuals in the community in fact follow the instruction and (b) a sufficient number of individuals in the community expect and demand of each other to follow the instruction (and may be willing to enforce it through sanctions; Bicchieri, 2006; Brennan et al., 2013).

We can now conceptualize *moral cognition* as the set of capacities that allow people to properly engage with social and moral norms. People have to (a) *learn, store, activate, and deploy* norms; (b) make *judgments* (e.g., of permissibility, wrongness, blame) about these norms; (c) make *decisions* in light of these norms; and (d) *communicate* about the norms and their violations (e.g., prescribe, justify, apologize).[1]

How is social cognition involved in moral cognition?

What is social cognition? We endorse an inclusive definition that subsumes under the term all conceptual and cognitive tools that serve the overarching goal of making sense of other human agents. Figure 2 displays many of these tools arranged in an approximate hierarchy (for a glossary and detailed discussion, see Malle, 2008, 2015). On the bottom are those that have evolved earlier in phylogeny, develop earlier in ontogeny, and are generally simpler and faster processes; and on the top are those that have evolved more recently, develop later in childhood, and are generally more complex and slower processes. Tools higher up often rely on the output of tools further down, and in concert these tools perform important tasks in social life, such as explanation, prediction, and moral judgment (depicted outside the tree itself). Moreover, several of the processes at once presuppose and shape fundamental concepts, such as intentionality, belief, desire, and emotion categories.

[1] Perhaps a helpful term for this set of capacities would be *moral competence* (Malle, 2015a; Malle & Scheutz, 2014). A complete rendering of this competence would include both positive and negative behaviors, but here we focus, in keeping with the literature, on negative behaviors.

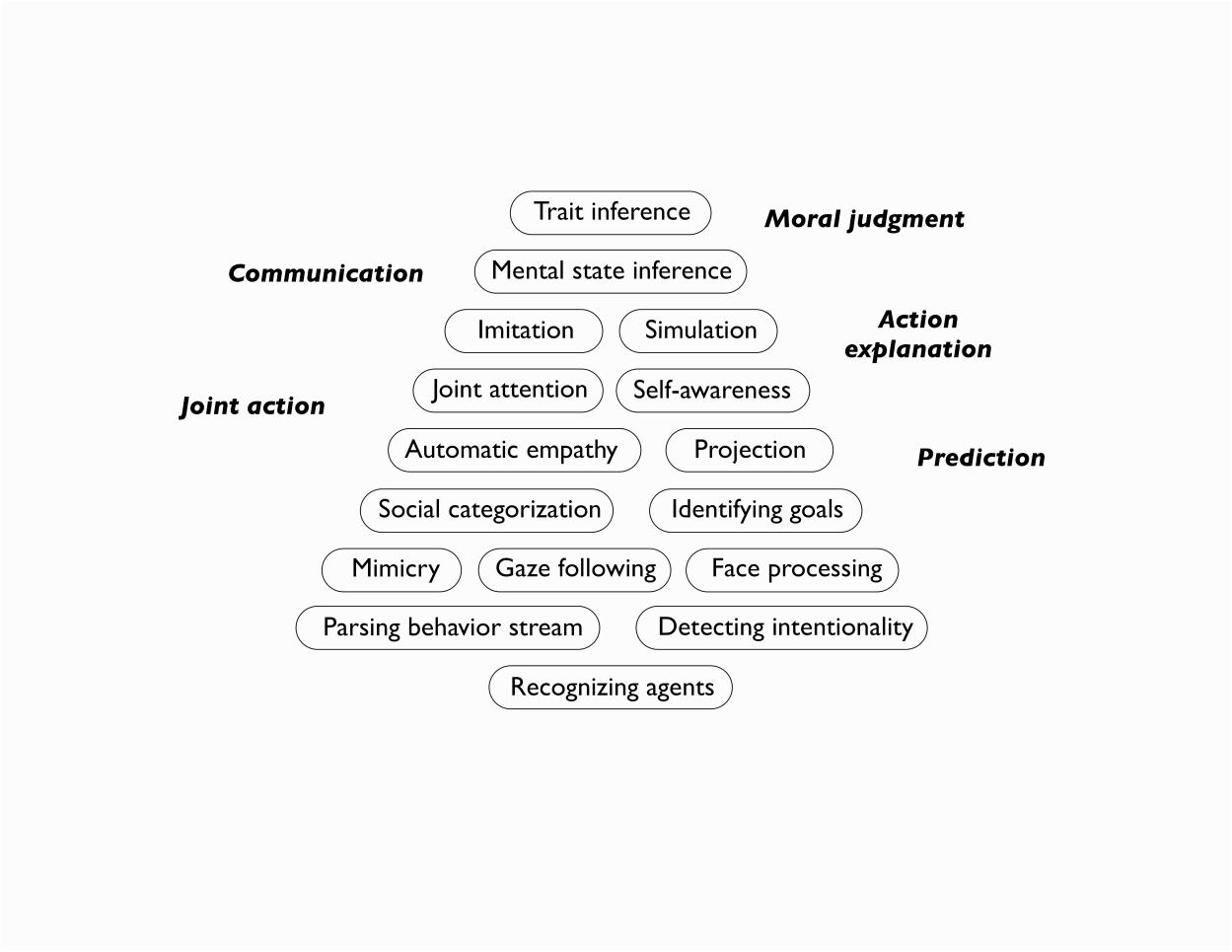

Figure 2. Broad conceptualization of social cognition as a tree-like, hierarchical collection of cognitive tools.

Against this background it is now easy to illustrate how social cognition supports and interacts with the four capacities of moral cognition.

In *norm learning*, social cognition contributes some of the learning mechanisms: mimicry and imitation provide powerful tools of adopting norms through action, while face processing and goal identification allow people to read others' evaluations of a given behavior and thereby infer the norms that the behavior conformed to or violated. For example, a scowl toward somebody who asks a new acquaintance too many private questions can teach and enforce norms of privacy and autonomy.

Moral judgment would be impossible without the basic tools of recognizing agents, parsing the behavior stream to identify (un)intentional norm violations, as well as empathy (often with the victim), simulation, and mental state inference to gauge the agent's specific reasons for committing the violation. Moreover, social categorization can influence judgments through prejudice (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006) and also help assign specific norms to people in particular roles, groups, and positions. Often overlooked is the fact that

different moral judgments require different levels of social-cognitive involvement: gauging the permissibility of an action is largely a matter of analyzing an action category relative to a norm system; the agent's specific mental states are less important. Wrongness appears to be judged more on the basis of the agent's mental state (Cushman, 2008), whereas blame incorporates all these information inputs (Malle et al., 2014).

Moral decisions and actions rely in part on self-awareness, simulation of one's own future guilt, empathy with potential victims, and others' moral sanctions. Such decisions and actions also involve social categorization of one's own roles and obligations, and accumulated trait inferences of one's own virtues (or lack thereof).

Moral communication, finally, includes such phenomena as expressing moral judgments either to the alleged violator or to another community member (Dersley & Wootton, 2000; Traverso, 2009); negotiating blame through justification and excuses (Antaki, 1994); and apology, compensation, or forgiveness to repair social estrangement after a norm violation (McKenna, 2012; Walker, 2006). People rely on mental-state inferences during communicative interactions and especially during social-moral interactions to accurately assess the other's goals and knowledge, because the stakes of maintaining relationships are high and under the threat of sanctions. Trait inferences may be formed through observation or gossip, especially when norm violators do not respond to social regulation attempts by their community. Also, low-level mechanisms of gaze and face processing, empathy, and goal inference are needed to gauge the honesty of justifications, the genuineness of apologies, and the seriousness of threatened sanctions.

How does social regulation interact with moral cognition?

We claimed earlier that heeding the social-regulatory function of moral cognition can benefit our understanding of how moral cognition itself operates. We now illustrate one such benefit by reconsidering the debate over how accurate or biased people are in forming moral judgments (Alicke, 2000; Ditto, 2009; Malle et al., 2014; Nadler & McDonnell, 2011).

The accuracy or bias of a given moral judgment is difficult to measure, because the laboratory rarely offers an objective criterion for the correct judgment. Typically researchers offer information to participants that they "should not" take into account, and when some of them do, a "bias" is diagnosed. For example, many researchers have argued that outcome severity, the motives and character of the norm violator, or the likeability of the victim must not be part of an unbiased moral judgment. But it is unclear who gets to decide, and on what basis, what people should or should not take into account (Malle et al., 2014; Nadler, 2012). Moreover, the potential arbiters, "philosophers, legal theorists and psychologists" (Alicke, 2008, p. 179), often do not agree with one another.

In the absence of objective criteria, an appealing alternative is to consider the moral judgment's function of regulating social behavior as a suitable standard—the socially shared criteria that people use to accept, question, criticize, or reject moral judgments. For example, what do people accept as the grounds of intense blame? They consider that the behavior violated

an important norm, that the violation was intentional, that the agent had no justifying reasons to perform the behavior, etc. (Malle et al., 2014). When would people reject intense blame? They do so when the behavior violated merely an insignificant norm, when the violation was unintentional and unavoidable but the norm enforcer treated it as if it was intentional, etc. Bias is then diagnosed when norm enforcers *overblame* or *underblame* relative to what is acceptable in the community (Kim, Voiklis, Cusimano, & Malle, 2015).

These standards of blame put pressure on people to keep their biases in check. Severe violations sometimes elicit powerful emotional responses that can lead to premature accusations or unfair punishment; and an observer's quick moral evaluation sometimes taints subsequent inferences about whether the violation was intentional, or justified, or preventable (Alicke, 2000; Knobe, 2010). Nevertheless, community members help correct these expressions of premature, biased, or inaccurate moral judgments by challenging those who blurt out allegations, by demanding warrant from those who overblame, thereby calming and slowing the processes of accusation and punishment.[2] Communities could not survive if their members blamed and punished one another without evidence or without differentiating between, say, mild and severe, intentional and unintentional violations. The regulatory functions of moral judgment, and the required warrant for such judgments, therefore push those judgments to be more reasonable, accurate, and fair, by the standards of the community in which they occur.[3]

Relating Social Cognition, Moral Cognition, and Social Regulation

We can now offer a more detailed schematic of the relationships between social cognition, moral cognition, and social regulation. In choosing the pictorial language of a flow diagram, we naturally leave out some complexity, but it forces us to make certain theoretical commitments, which can be tested experimentally.

The flow of processes begins with a negative event, which prompts the perceiver to assess whether the event was caused by an agent who violated norms. If yes, social-cognitive processes analyze the violator's mental states (including intentions and goals). This information feeds into moral cognition, which generates judgments about wrongness or blame. The outputs of moral and social cognition, along with preceding information about the event and the norms that were violated, feed into a decision about whether public moral criticism is warranted. If warrant exceeds threshold, the perceiver is likely to deliver public moral criticism (though many other considerations may inhibit criticism, such as role constraints, fear of retaliation, etc.). This moral criticism may prompt a timely change in the violator's behavior or, if not, the perceiver may consider renewed criticism or alternative responses, including gossip or retreat.

[2] These socially corrective strategies are not inventions of modern legal institutions; rather, they are successful informal practices that have persisted throughout history (Boehm, 1999; Pospisil, 1971).

[3] There are well-known limits to this shaping process: For example, members of a given group may demand fair and accurate norm enforcement for one another but not for members of disliked or lower status outgroups.

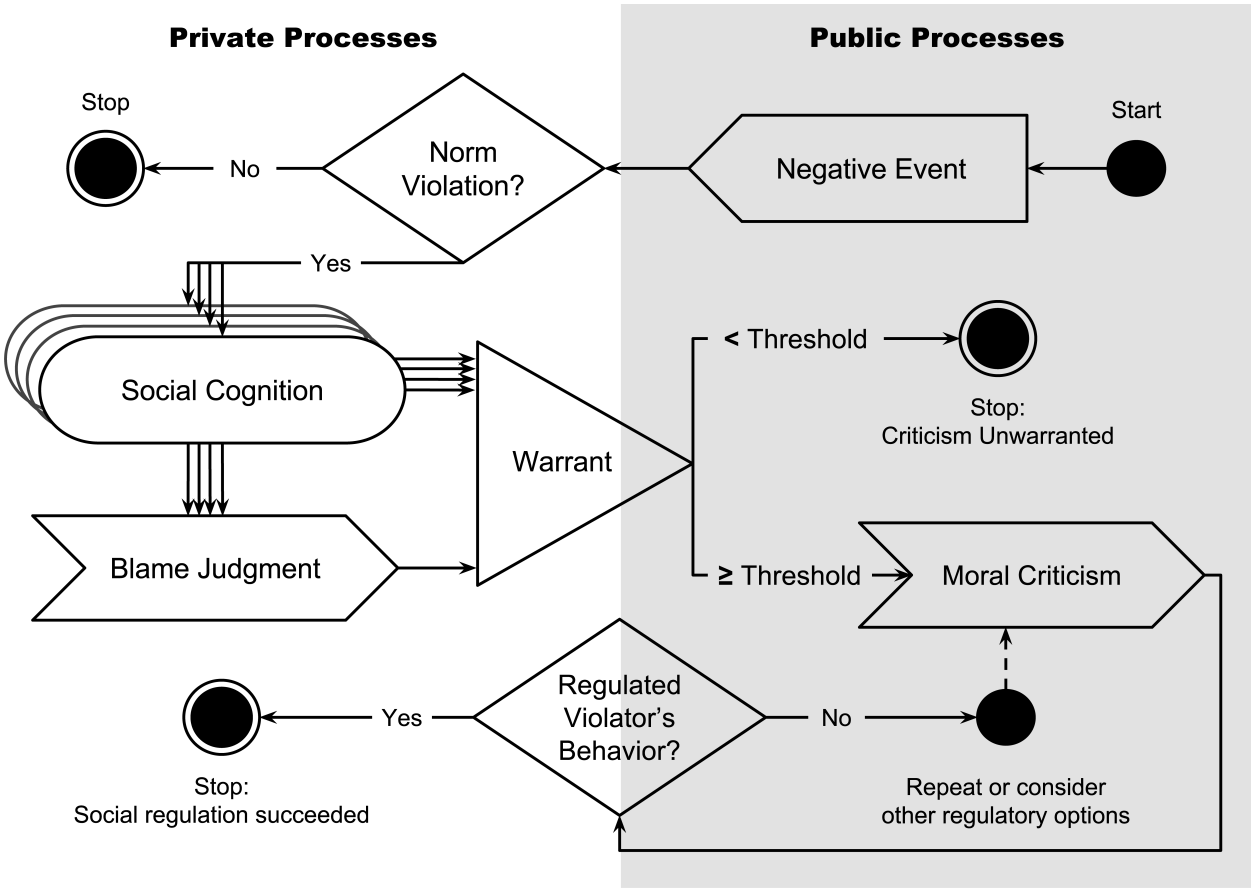

Figure 3. Flow diagram of processes of social and moral cognition in the context of social regulation of norm violations.

The full stop with which we break off the flow diagram conceals a more complex, finely tuned social dynamic between norm enforcers and norm violators: They negotiate levels of blame, meet accusation with justification, criticism with remorse, remorse with forgiveness, all in the service of rebuilding and maintaining social relationships (Walker, 2006).

4. Evidence

Empirical evidence for the social-cognitive basis of moral judgment has been accumulating over the past several years. In many studies, lay people clearly rely on social-cognitive inferences of intentionality when judging everyday moral actions (Lagnado & Channon, 2008) and when mastering fine distinctions between willingly, knowingly, intentionally, purposefully violating a norm (Guglielmo & Malle, 2010)—distinctions that also inform legal classifications of negligence and recklessness. Likewise, lay people judge goal-directed harm as less permissible and more often as wrong than harm as a side effect (Cushman & Young, 2011). Thus, moral and legal distinctions overlap with (and, perhaps, derive from) more general-purpose social-cognitive judgments. This derivative relationship is corroborated by results from functional magnetic

resonance imaging and lesion studies showing that the processing involved in either social or moral judgment activate many of the same regions in the prefrontal cortex (Forbes & Grafman, 2010).

People's cognitive system also makes distinctions between types of moral judgments that vary by the objects they judge: badness judges mere events, wrongness judges intentional actions, and blame judges an agent's specific relationship to a norm violation, whether intentional or unintentional (Malle et al., 2014; Monin, Pizarro, & Beer, 2007; Sher, 2006). These judgments also differ in their sensitivity to causal and mental-state information (Cushman, 2008; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015), but experiments on the detailed causal processes that flow between social cognition and these differing judgments remain lacking.

Experiments on social expressions of blame are also scarce. Nevertheless, initial work in our lab has demonstrated that, as with private judgments, people have a finely tuned map of public acts of moral criticism (Voiklis, Cusimano, & Malle, 2014). For example, the rich vocabulary used by English speakers to describe such acts—ranging from chiding violators to lashing out at them—do not merely represent linguistic variations but pick out systematic features of the underlying moral judgment and of the social context. When participants assessed twenty-eight acts (described by the most common verbs of moral criticism) on numerous properties of judgment and context, the first two dimensions of a principal components analysis were intensity of expression and direction of expression (towards the offender or towards others). Figure 4 depicts the quadrants of this space and four verbs that mark the prototypical acts in each quadrant. In a subsequent series of studies, we tested the hypothesis that people likely follow "norms of blaming" when scaling the intensity of moral criticism to the severity of transgressions (Kim et al., 2015). Indeed, when judging the appropriateness of various levels of moral criticism in response to a range of mild to severe transgressions, participants displayed a norm against "overblaming" (i.e., overly intense criticism for mild violations) but were more tolerant of "underblaming."

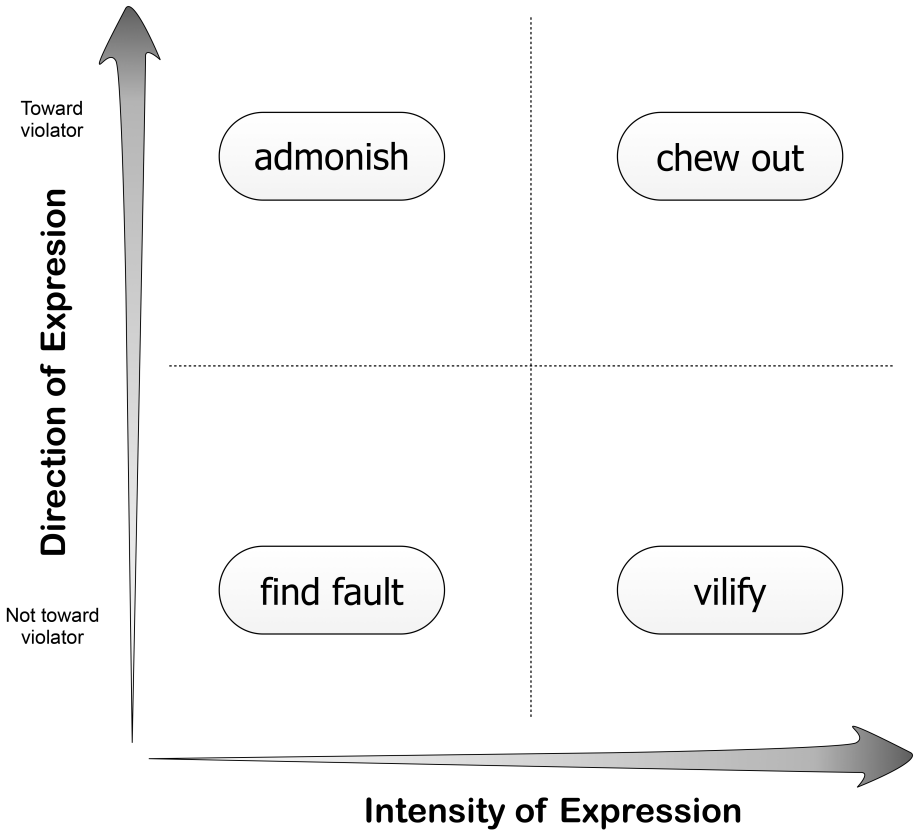

Figure 4. Four prototypes of public acts of moral criticism amidst variation of intensity of expression and direction of expression.

Individual and situational variability in social and moral cognition

So far, we have addressed social and moral cognition at the level of cognitive system components that exist in all neurotypical adults. Nevertheless, social-cognitive performance can vary as a function of maturation, neurological damage, and psychopathology (Frith & Frith, 2003), as well as due to motivation (Klein & Hodges, 2001) and task difficulty (Birch & Bloom, 2007). Often these deficits are presented as evidence that people are reflexively egocentric in their perception of other minds (Lin, Keysar, & Epley, 2010). An alternative interpretation is that people are dispositionally or situationally unprepared for attending to the full range of social information. In fact, preliminary evidence suggests that "warming up" social cognition with a practice task facilitates spontaneously unbiased predictions (in a mixed motive game) and spontaneously subtle assessments of intentions and intentionality (Knobe, 2003), especially for those scoring on the lower end of a social-cognitive performance measure (Voiklis, in preparation). So even though shallow processing and bias may predominate in states of disengagement, the correct situational cues can bring most individuals to their full social-cognitive potential. Among these situational cues, the community's demand for warrant in moral

criticism (especially blame) must rank very high, but direct tests of this hypothesis remain lacking.

There is, however, evidence for the malleability and the social shaping of moral reasoning more generally. As with other forms of (public) reasoning (Crowell & Kuhn, 2012; Kuhn, Zillmer, Crowell, & Zavala, 2013), moral judgment can improve with practice and feedback. Much as habitual reliance on heuristics (e.g., confirmation seeking) can be overcome with deliberate practice (Kuhn, 2011), people might likewise overcome any habitual neglect of social-cognitive information. Howe (1991), for example, showed in an experimental context that circuit judges adjusted their blame judgments to mitigating information twice as strongly as students. Applying one's social-cognitive abilities might also be a matter of mindset. When induced to believe in the malleability, as opposed to fixedness, of empathy, people appear more willing to expend empathic effort towards challenging targets (Schumann, Zaki, & Dweck, 2014). Moreover, people with a malleable mindset appear to seek out these challenges in order to improve their empathy; the challenge provides the learning opportunity, and the motivation to learn helps them meet that challenge.

Beyond skill learning, the vast developmental literature on changes in moral judgment and decision making support the claim of malleability. Gradual differentiation in moral cognition, according to our framework, is in good part the result of gradual differentiation in social cognition (Baird & Astington, 2004). For example, norm learning becomes more sophisticated as mental state inferences improve, and blame judgments become more sophisticated as the conceptual framework of mind grows. Specifically, as mental state concepts of belief and desire mature by age 4 to 5 (Wellman, 1990), outcome considerations in blame are balanced by mental-state considerations (Nelson-Le Gall, 1985). And as further differentiations of the intentionality concept emerge (Baird & Moses, 2001), the distinction between justified and unjustified violations and between preventable and unpreventable outcomes emerge as well (Fincham, 1982; Shaw & Sulzer, 1964).

What data would falsify our proposal?

The strongest evidence against our proposal would show that early moral evaluations or emotions in response to norm violations precede and swamp subsequent social-cognitive processing (Alicke, 2000; Knobe, 2010), a reversal to what our framework suggests. Confirmation of this claim requires methods for assessing temporal-causal relations between processes (millisecond by millisecond), but such methods have yet to be introduced into moral psychology. Furthermore, confirmation of this claim requires measuring a perceiver's affective responses after the perceiver recognizes an event as norm violating but before the perceiver determines the agent's causal involvement, intentionality, mental states, etc. Given the evidence for very early and automatic processing of agency and intentionality (Barrett, Todd, Miller, & Blythe, 2005; Decety, Michalska, & Kinzler, 2012), it would be difficult, both theoretically and experimentally, to fit any kind of graded affect into this tight early time window. Nevertheless, people are likely to perceive some kind of pre-conceptual badness before they process all the

details of a norm-violating event. Arguably, such an undifferentiated sense of badness does not represent a moral judgment (e.g., of blame), so, arriving at such a judgment would require additional social-cognitive processing. If this processing were systematically biased in favor of confirming the initial negative assessment (Alicke, 2000), moral judgments would still be fundamentally reliant on social cognition—but on less accurate social cognition.

A second major challenge to our proposal would be that social regulation of norm enforcement does not, as we propose, push social (and moral) cognition toward systematic information processing and accuracy. Evidence would have to show that the demand for warrant of moral judgments can be easily satisfied by biased and inaccurate social-cognitive information. It would not be enough to show that under some circumstances demand for warrant is ineffective but, rather, that widespread demand for warrant either does not exist or, even if it exists as a social practice, does not predict quality of social and moral processing.

5. Extension and Expansion

Social regulation of moral judgments

Our hypothesis that social regulation is not only the expression of moral judgment but a mechanism that keeps moral judgments honest has yet to be tested. A first requirement of testing it will be to design suitable experimental manipulations of community members putting demands for warrant on a perceiver who is expressing a moral judgment. A second requirement will be to devise reliable measures of accuracy in moral judgments and the perceiver's systematic responsiveness to evidence.

Our theoretical model predicts that the impact of demands for warrant varies by moral judgment type. Permissibility judgments are primarily reflections of shared norms, so the presence of a community member should simply increase reliability and collective agreement in these kinds of judgments, whereas the more complex blame judgments should become more deliberate and evidence-based (i.e., taking into account intentionality, mental states, etc.) in the presence of a community representative. There is also a reverse prediction—that an overwhelming need to be accepted by one's community can lead to more biased information processing if the community has strong expectations (e.g., about the guilt or innocence of a norm violator, or about the appropriate level of punishment). The fine balance between these different forces may be examined with agent-based modeling methods (Elsenbroich & Gilbert, 2014). "Societies" that balance socially demanded accuracy against socially demanded unanimity should be most successful because they keep the costs of false accusations and exaggerated punishment in check. However, stratified societies in which some subgroups have more power may shift these costs to the less powerful groups. The current incarceration rates of minorities in the U.S. is an example of such a dynamic. As a counterforce, however, recently increasing public scrutiny of aggressive policing of minorities signals a renewed demand for warrant for social-moral blame and punishment.

Institutional mechanisms of regulation, such as the state and the law, were long believed to be the dominant forms of regulation. But evidence from the fields of anthropology, psychology, sociology, and legal studies suggests that informal, interpersonal moral regulation is evolutionarily and culturally old, arises developmentally early, and is the predominant way, even today, of keeping individual community members in line with collective interests. Referring back to our flow diagram, ordinary social regulation sometimes fails; an enticing research direction might be to examine when institutional mechanisms take over social regulation and when these mechanisms are more effective than interpersonal ones.

Affect and emotion as social-moral signals

While the exact causal roles of affect and emotion in the information processing phase of moral cognition are still under debate, their involvement in public expressions of moral criticism may be more readily apparent (Wolf, 2011). Affect intensity—in words, face, and posture—scales such expressions (Voiklis et al., 2014) so that others recognize one's degree of outrage (McGeer, 2012; de Melo, Carnevale, Read, & Gratch, 2014). These expressions signal how important the violated norm is to the blamer, teach young community members about such importance rankings, and also communicate to norm violators what possible other sanctions might follow if they show no insight or atonement. Evidence for this social function of moral emotions might come from physiological studies that show a ramping up of negative arousal from early violation detection to late public expression. That is, the very opportunity to express one's judgment publicly may increase the involvement of affect that was previously, during mere "in the head" judgments, quite modest. Additional support might come from evidence that perceivers have less differentiated emotions when they cognitively form their moral judgments than when they publicly express them, because anticipating public scrutiny leads to more attentive information appraisals. Here too, perceivers' awareness of a community's strong expectations may sometimes unduly modulate their public judgments, such as offering exaggerated expressions of outrage; this, in turn, can fuel even stronger expressions by other community members and escalate collective moral condemnation beyond what perceivers felt in private.

Artificial morality

Work on moral psychology has recently expanded into artificial morality—the study and design of computational models of moral competence (Mao & Gratch, 2012; Tomai & Forbus, 2008) and implementation in social robots (Wallach & Allen, 2008; Malle & Scheutz, 2014). Social robots—embodied machines that are able to interact with humans—play an increasing role in contemporary society. Around a decade ago there were no robots in private homes, whereas in 2014, 4.7 million service robots for personal and domestic use were sold worldwide (IFR, 2015). These robots rarely posses extensive social-cognitive capacities but are improving rapidly (Nourbakhsh, 2013), and robots may soon function as social companions or assistants in health care, education, security, and emergency response. In such applications, however, robots will need to have basic moral competence to ensure physically and psychologically safe interactions

with humans (Malle & Scheutz, 2014). Designing such robots offers appealing new avenues for research, by testing more precise, formally specified models of both social-cognitive capacities (e.g., making intentionality inferences in live interactions) and moral capacities (e.g., recognizing norm violations and evidence-based responding). In addition, research will need to identify the conditions under which humans ascribe features such as intentionality, free will, or blame to artificial agents (Malle et al., 2015; Meltzoff, Brooks, Shon, & Rao, 2010; Monroe, Dillon, & Malle, 2014), because such ascriptions fundamentally alter human-machine interactions. Integrating robots into research will enable a better understanding of social and moral cognition, and integrating robots into society will require such understanding to achieve beneficial human-robot co-existence.

6. Summary

Returning from social-cognitive science fiction, we close by recapping our theoretical framework for understanding the processes of moral cognition. We argue that a hierarchy of social-cognitive tools ground moral cognition and that social and moral cognition together guide the social regulation of behavior. The practice of social-moral regulation, in turn, puts pressure on community members to engage in reasonably fair and evidence-based moral criticism. With the help of these cognitive adaptations and social practices, people are able to navigate the terrain of morality, accruing bumps and bruises along the way, but surviving as the most sophisticated social creature currently roaming the earth.

Acknowledgment

We are grateful for the collective work and insights by all members of our lab (http://research.clps.brown.edu/SocCogSci/Personnel/personnel.html). This project was supported by grants from the Office of Naval Research, No. N00014-14-1-0144 and N00014-13-1-0269.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574. doi:10.1037//0033-2909.126.4.556

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*, 179–186. doi:10.1163/156770908X289279

Antaki, C. (1994). *Explaining and arguing: The social organization of accounts*. London: Sage.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, *2004*, 37–49. doi:10.1002/cd.96

Baird, J. A., & Moses, L. J. (2001). Do preschoolers appreciate that identical actions may be motivated by different intentions? *Journal of Cognition & Development*, *2*, 413–448. doi:10.1207/S15327647JCD0204_4

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*, 313–331. doi:10.1016/j.evolhumbehav.2004.08.015

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382–386. doi:10.1111/j.1467-9280.2007.01909.x

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

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. New York, NY: Oxford University Press.

Crowell, A., & Kuhn, D. (2012). Developing dialogic argumentation skills: A 3-year intervention study. *Journal of Cognition and Development*, *15*, 363–381. doi:10.1080/15248372.2012.725187

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380. doi:10.1016/j.cognition.2008.03.006

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*, 1052–1075. doi:10.1111/j.1551-6709.2010.01167.x

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*, 209–220. doi:10.1093/cercor/bhr111

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*, 73–88. doi:10.1037/a0034251.supp

Dersley, I., & Wootton, A. (2000). Complaint sequences within antagonistic argument. *Research on Language and Social Interaction*, *33*, 375–406. doi:10.1207/S15327973RLSI3304_02

Ditto, P. H. (2009). Passion, reason, and necessity: A quantity-of-processing view of motivated reasoning. In T. Bayne & J. Fernández (Eds.), *Delusion and self-deception: Affective and motivational influences on belief formation.* (pp. 23–53). New York, NY: Psychology Press.

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making* (pp. 307–338). San Diego, CA: Elsevier Academic Press.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383–386. doi:10.1111/j.1467-9280.2006.01716.x

Elsenbroich, C., & Gilbert, G. N. (2014). *Modelling norms*. Dordrecht, Netherlands: Springer.

Fincham, F. D. (1982). Moral judgment and the development of causal schemes. *European Journal of Social Psychology*, *12*, 47–61. doi:10.1002/ejsp.2420120104

Forbes, C. E., & Grafman, J. (2010). The role of the human prefrontal cortex in social cognition and moral judgment. *Annual Review of Neuroscience*, *33*, 299–324. doi:10.1146/annurev-neuro-060909-153230

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*, 459–473. doi:10.1098/rstb.2002.1218

Göckeritz, S., Schmidt, M. F. H., & Tomasello, M. (2014). Young children's creation and transmission of social norms. *Cognitive Development*, *30*, 81–95. doi:10.1016/j.cogdev.2014.01.003

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development.* (pp. 35–80). Cambridge, MA: MIT Press.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*, 1635 –1647. doi:10.1177/0146167210386733

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. doi:10.1080/00201740903302600

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834. doi:10.1037/0033-295X.108.4.814

Howe, E. S. (1991). Integration of mitigation, intention, and outcome damage information, by students and circuit court judges. *Journal of Applied Social Psychology*, *21*, 875–895. doi:10.1111/j.1559-1816.1991.tb00448.x

Hume, D. (1998). *An enquiry concerning the principles of morals*. (T. L. Beauchamp, Ed.) (Scholarly edition, orig. publ. 1751.). Oxford, UK: Oxford University Press.

IFR (2015). *World robotics: Service robots 2015*. (Available at http://www.ifr.org/service-robots/statistics/) Frankfurt, Germany: International Federation of Robotics (IFR).

Kant, I. (2012). *Groundwork of the metaphysics of morals*. (M. Gregor & J. Timmermann, Trans.) (Revised edition, orig. publ.1785.). Cambridge: Cambridge University Press.

Kim, B., Voiklis, J., Cusimano, C., & Malle, B. F. (2015, February). *Norms of moral criticism: Do people prohibit underblaming and overblaming?* Poster presented at the Annual meeting of the Society of Personality and Social Psychology, Long Beach, CA.

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin*, *27*, 720–730. doi:10.1177/0146167201276007

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190 – 194. doi:10.1093/analys/63.3.190

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329. doi:10.1017/S0140525X10000907

Kohlberg, L. (1981). *The philosophy of moral development: Moral stages and the idea of justice*. San Francisco: Harper & Row.

Kuhn, D. (2011). What people may do versus can do. *Behavioral and Brain Sciences*, *34*, 83–83. doi:10.1017/s0140525x10002864

Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cognition and Instruction*, *31*, 456–496. doi:10.1080/07370008.2013.830618

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*, 754–770.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556. doi:10.1016/j.jesp.2009.12.019

Malle, B. F. (2008). The fundamental tools, and possibly universals, of social cognition. In R. M. Sorrentino & S. Yamaguchi (Eds.), *Handbook of motivation and cognition across cultures* (pp. 267–296). New York, NY: Elsevier/Academic Press.

Malle, B. F. (2015a). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*. doi:10.1007/s10676-015-9367-8

Malle, B. F. (2015b). Social robots and the tree of social cognition. In Y. Nagai & S. Lohan (Eds.), *Proceedings of the Workshop "Cognition: A bridge between robotics and interaction" at HRI'15, Portland, Oregon* (pp. 13–14). Available at http://www.macs.hw.ac.uk/~kl360/HRI2015W/proceedings.html.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*, 147–186. doi:10.1080/1047840X.2014.877340

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. *IEEE International Symposium on Ethics in Engineering, Science, and Technology* (pp. 30–35). Chicago, IL: IEEE.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15 (pp. 117–124). New York, NY: ACM.

Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgment in multi-agent interactions. *Journal of Artificial Intelligence Research*, *44*, 223–273.

McGeer, V. (2012). Civilizing blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 162–188). New York, NY: Oxford University Press.

McKenna, M. (2012). Directed blame and conversation. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 119–140). New York, NY: Oxford University Press.

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*, 966–972. doi:10.1016/j.neunet.2010.09.005

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, *11*, 99–111. doi:10.1037/1089-2680.11.2.99

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, *27*, 100–108. doi:10.1016/j.concog.2014.04.011

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems*, *75*, 1–31.

Nadler, J., & McDonnell, M.-H. (2011). Moral character, motive, and the psychology of blame. *Cornell Law Review*, *97*, 255–304.

Nelson-Le Gall, S. A. (1985). Motive-outcome matching and outcome foreseeability: Effects on attribution of intentionality and moral judgments. *Developmental Psychology*, *21*, 323–337. doi:10.1037//0012-1649.21.2.332

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. doi:10.1037/0033-295X.84.3.231

Nourbakhsh, I. R. (2013). *Robot futures*. Cambridge, MA: MIT Press.

Piaget, J. (1932). *The moral judgment of the child*. London, UK: Kegan Paul, Trench, Trubner and Co., 1932.

Pospisil, L. (1971). *Anthropology of law: A comparative theory*. New York: Harper & Row.

Schumann, K., Zaki, J., & Dweck, C. S. (2014). Addressing the empathy deficit: Beliefs about the malleability of empathy predict effortful responses when empathy is challenging. *Journal of Personality and Social Psychology*, *107*, 475–493. doi:10.1037/a0036738

Semin, G. R., & Manstead, A. S. R. (1983). *The accountability of conduct: A social psychological analysis*. London: Academic Press.

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

Shaw, M. E., & Sulzer, J. L. (1964). An empirical test of Heider's levels in attribution of responsibility. *Journal of Abnormal and Social Psychology*, *69*, 39–46. doi:10.1037/h0040051

Sher, G. (2006). *In praise of blame*. New York, NY: Oxford University Press.

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

Tomai, E., & Forbus, K. (2008). Using qualitative reasoning for the attribution of moral responsibility. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Traverso, V. (2009). The dilemmas of third-party complaints in conversation between friends. *Journal of Pragmatics*, *41*, 2385–2399. doi:10.1016/j.pragma.2008.09.047

Ullmann-Margalit, E. (1977). *The emergence of norms*. Clarendon library of logic and philosophy. Oxford: Clarendon Press.

Voiklis, J. (in preparation). A little "mindreading" practice facilitates later social-cognitive inferences.

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 Meeting of the Cognitive Science Society* (pp. 1700–1705). Austin, TX: Cognitive Science Society.

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

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. New York, NY: Oxford University Press.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York, NY: Guilford Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wolf, S. (2011). Blame, Italian style. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reason and recognition: Essays on the Philosophy of T. M. Scanlon* (pp. 332–347). New York, NY: Oxford University Press.