The neuroscience of moral cognition: from dual processes to dynamic systems
Jay J Van Bavel, Oriel FeldmanHall and Peter Mende-Siedlecki

Prominent theories of morality have integrated philosophy with psychology and biology. Although this approach has been highly generative, we argue that it does not fully capture the rich and dynamic nature of moral cognition. We review research from the dual-process tradition, in which moral intuitions are automatically elicited and reasoning is subsequently deployed to correct these initial intuitions. We then describe how the computations underlying moral cognition are diverse and widely distributed throughout the brain. Finally, we illustrate how social context modulates these computations, recruiting different systems for real (vs. hypothetical) moral judgments, examining the dynamic process by which moral judgments are updated. In sum, we advocate for a shift from dual-process to dynamic system models of moral cognition.

Address
New York University, United States

Corresponding author: Van Bavel, Jay J (jay.vanbavel@nyu.edu, jay.vanbavel@gmail.com)

In June 2003, a runaway freight train sped toward downtown Los Angeles at 70 miles per hour [1]. Railway officials estimated that if the train hit downtown LA, it would kill dozens. However, diverting the train to a less populous, blue-collar neighborhood could potentially reduce the cost of damage and the death toll. After deliberating, the officials decided to divert the train, which crashed into several homes, injuring a dozen residents, including several small children. Every day, medical researchers, insurance adjusters, and automobile manufacturers are forced to address a similar moral dilemma: how do we evaluate the costs of human life against other considerations?

Moral questions like these have long been debated within the philosophical community, often through the use of thought experiments designed to probe intuitions of right and wrong. This approach has gained considerable traction in experimental research, where thought experiments like the Trolley dilemma — in which people must decide to kill one person to save five — have become some of the most popular methods in psychology. These ubiquitous moral dilemmas illustrate that changing one simple factor (e.g., pushing a man vs pulling a lever) can profoundly alter moral judgment [2], and offer evidence that emotion and reason can act as dueling psychological inputs. Taken together, these results have provided the empirical foundation for dual-process models of morality (e.g. [3,4]). Yet these hypothetical moral dilemmas fail to capture many of the psychological and neural processes that underpin real moral behavior.

To better understand the mental computations underlying moral judgment and decision-making, recent theories of morality have begun to integrate ancient philosophical debates with rich psychological and biological models. We review research from the dual-process tradition, in which intuitions are automatically elicited and reasoning is subsequently deployed to rationalize or correct for these initial intuitions. We next describe how the mental computations underlying moral cognition are diverse (including personal goals and identities, representations of others’ mental states and character, and social norms) and widely distributed throughout the brain. These findings illustrate why models of moral cognition require more than two processes. We also describe how social context modulates these computations, recruiting different systems for real (vs. hypothetical) moral judgments. Finally, we examine the dynamic process by which moral judgments are updated based on new information. Taken together, we argue for a shift away from dual-process models in favor of a more dynamic model of moral cognition.

Dual-process morality
Over the past fifteen years, the field of moral cognition has focused heavily on the roles of intuition and reason in moral judgment. For example, the Social Intuitionist Model [5] states that these intuitions provide an automatic, internal signal that guides moral judgment. According to this model, the human capacity for reason is largely relegated to the role of post hoc rationalizing, merely serving to justify initial moral intuitions. Indeed, the human brain responds to harmful acts within a hundred milliseconds — suggesting that moral intuitions emerge very rapidly [6]. Likewise, the physical act of harming someone feels intuitively wrong to many people,
even if it might save the lives of many others. Accordingly, people are less willing to sacrifice one life to save several others when they have to physically harm someone than when they are emotionally distant from the situation [7]. Early neuroimaging work suggested that these more ‘personal’ actions evoke a strong response in brain regions associated with affective valuation — like the ventromedial prefrontal cortex (vmPFC) — dovetailing with the theory that automatic intuitions can overpower more deliberate, utilitarian calculations [2]. Similarly, damage to the vmPFC is associated with increased support for a more rational, utilitarian response (e.g. [8,9]). These studies have firmly established the role of intuition in moral cognition.

It remains a source of contention, however, whether or not people can effectively regulate these intuitions. According to many dual-process models, corrective control plays a major role in moral judgment [3,4]. People can override their initial aversion to harming someone in order to save several others, and this utilitarian decision is associated with activation in several brain regions implicated in cognitive control — such as the anterior cingulate cortex (ACC) and lateral prefrontal cortex [10]. This is especially true during high-conflict dilemmas, when strong emotional intuitions clash with utilitarian computations — such as pushing a man off a bridge or smothering a baby to save several other lives. Moreover, cognitive load [11] and temporary disruption of the lateral prefrontal cortex [12,13] impair the ability to enact these utilitarian decisions. Consistent with dual-process models, intuition and reason often work in opposition to generate moral judgments.

**Dual-process models: a false dichotomy?**

Debates about the precise interplay between intuition and reason have dominated the study of morality. However, developments in cognitive neuroscience lay bare a major problem underlying this debate: there are more than two processes guiding moral judgment (Figure 2; see [14,15]). For instance, one influential review of the neural systems underlying dual process models of moral cognition actually described at least eight distinct brain regions, each of which implements a discrete cognitive process [16]. The complexity of moral cognition is compounded by the temporal dynamics of moral judgment, which unfold over time courses ranging from milliseconds to days. For example, initial reactions regarding harm can be modulated depending on the demands of the situation and as information about the target and normative context is revealed [6*]. These rapidly emerging, highly dynamic cognitive processes involve more than simple corrective control and are critical to understanding moral judgment.

We are not the first to suggest that morality is better characterized as a dynamic system [17,18], nor is this feature specific to morality: the evaluative system is comprised of widely distributed component processes that are highly interactive and sensitive to frames and mindsets ([19]; see also [20]). The field of neuroscience underscores the need to move beyond the simple dual-process dichotomy and embrace models of moral cognition that capture the rich, dynamic nature of human psychology and neuroscience. It would be a lost opportunity if we reduced the rich neuroscience findings into pre-existing, over-simplified psychological models. Instead, we argue that psychological models can benefit from incorporating the complexity and biological realism from neuroscience research. In the next section, we specify several component processes involved in moral cognition and begin to articulate the complex neural, social, and temporal dynamics that underlie moral judgment.

**Toward a psychological and neural model**

At its core, moral judgment juxtaposes the interests of self and other, such as choosing self-enhancement at another’s expense [21*]. The moral self–comprises moral ‘oughts’ and ideals [22], as well as situational and relational needs [23]: What kind of moral person do you aim to be? Who is a member of your moral community? Which moral codes do you endorse? An individual must continually integrate their own needs with the needs of others, within the broader context of what is socially and morally appropriate (i.e. social norms; [24,25]). Furthermore, representations of other individuals often incorporate mental state attributions (what are the intentions behind this person’s behavior? [26]), social categories (is this person a friend or foe? [27]), and reputational information (has this person harmed others in the past? [28]). Thus, the moral value of a given action — its rightness or wrongness — emerges from the integration of a combination of self and other-related information.

This interaction between self and other serves as the foundation upon which other social, societal, and situational pressures are dynamically exerted (see Figure 1). For example, there are times when the needs of the self-outweigh social norms or the needs of others, such as when people steal and harm others for material gain or power or when the others are members of a despised outgroup [21*,29*]. Moreover, the perception of moral ideals or oughts can change depending on the social context in which a moral dilemma is presented [30]. Context can also shift the weighting given to self and other-oriented concerns when computing the moral value of a particular choice or behavior [31–33].

The hypothetical scenarios used to study moral decision-making typically ignore the influence of socio-emotional factors [34] and contextual tensions endemic to real moral dilemmas [21*,35,36]. As such, an individual’s response to these dilemmas often fails to reflect how they would truly behave in a more enriched context. When a decision is hypothetical, individuals routinely report that they would...
Several brain regions appear critical for processing these complex socio-emotional tensions [40]; see Figure 2. For example, compliance with social rules, such as adhering to fairness norms, engages the dorsolateral PFC [41*,42]. When conflicting tensions arise, such as when individuals want to break a fairness norm in order to benefit monetarily, the ACC indexes these conflicting self-other tensions [38,43]. Visceral feedback from the body — as indexed by skin conductance responses [44] and anterior insula activity [45] — can modulate these responses to unfair treatment by increasing how morally wronged a person feels. The vmPFC appears to integrate competing influences into a moral value signal, for example, when choosing between tangible and abstract moral rewards, such as donating to charities [46,47], punishing social norm violators [48], and responding empathically during altruistic behavior [49].

Moral cognition is also dependent on brain structures directly involved in self and other-related processing. The mPFC is engaged when thinking about the self [50], as well as others’ mental states [51], suggesting that the perception of self and other can be intimately entwined [52,53]. The temporoparietal junction (TPJ) also emerged early on as a key region necessary for decoding social cues [54] and specifically for representing the intentions [55,56] and emotions of others [57]. More recent work has found that these other-oriented intentions and motivations have downstream effects on moral behavior [58].

Inferences regarding moral character dominate our impressions of others [59*,60,61] and impact how moral behavior is judged. Yet when tasked with the same decision in a more realistic context, monetary self-benefit often trumps previously stated intentions to be honest [29*,37] or to help another [21*,38]. This hypothetical bias effect [39] contributes to how pressures between the self and other are weighted to guide moral judgment.

Figure 2

The neural substrates of moral judgment and decision-making constitute a widely-distributed network of interactive brain regions. Here we outline several key regions that have been repeatedly implicated in moral cognition. The color-coding in the figure above loosely reflects key brain regions that contribute to specific roles within moral decision-making: purple, self-related processing; blue, other-related processing; red, motivations, green, processing related to conflict and updating. Finally, the vmPFC (shaded in black) integrates these distributed inputs toward an ultimate calculation of the moral value of a given judgment or behavior. The brain regions depicted in this figure are not intended to be an exhaustive list of every structure ever associated with moral judgment and decision-making. (For example, the posterior cingulate cortex and anterior temporal lobe could also be highlighted as playing a role in processing relevant to morality; see [17].) Due to space constraints, we have merely provided an overview of key regions.
The conclusion 

The nascent field of moral neuroscience has already offered numerous insights about the nature of moral judgment and decision-making. Foremost among them, moral cognition appears to be far less simplistic than some psychological models suggest, involving complex neural dynamics (interactions between brain regions), social dynamics (values shift between contexts) and temporal dynamics (rapid and continuous updating). Although these insights are based on a growing body of research, they mesh closely with the neuroscience of evaluation, emotion-regulation, and social cognition, and have the potential to reshape the dominant paradigm for understanding moral judgment and decision-making.

Conflict of interest statement

Nothing declared.

Acknowledgments

The authors would like to thank members of the New York University Social Perception and Evaluation Lab (@vanbavellab) for helpful comments on this manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest


The authors recorded high-density event-related potentials while participants watched brief morally-relevant scenarios involving either intentional or accidental harm. Results indicated that responses in right pSTS distinguished between intentional and accidental harm as quickly as 62 ms after presentation, followed by harm-related processing in amygdala/temporal pole (122 ms) and vmPFC (182 ms). The authors argue that these results suggest that intentionality and affect are critical, early inputs to moral cognition.


Demonstrates that real moral decisions can dramatically contradict moral choices made in hypothetical scenarios involving personal gain and harm to another. Reducing the opportunity for mental simulation by enhancing contextual information in a hypothetical moral problem can align hypothetical responses with real behavior.


29. Baumgartner T, Fischbacher U, Feierabend A, Lutz K, Fehr E: The neural circuitry of a broken promise. Neuron 2009, 64:756-770. The authors investigated the neural bases of keeping and breaking promises, observing that promise-breaking was associated with increased activation in DLPCF, ACC, and amygdala, and further, that the decision to break a promise could be predicted by pre-decision activity in anterior insula, ACC, and IFG. These results suggest that dishonesty in this context is associated with activity in both structures associated with conflict and cognitive control, as well as affective processing.


The author establishes a model of third-party punishment, outlining the psychological and neural processes contributing to the human capacity to learn, follow, and enforce social norms.


47. Hare TA, Camerer CF, Knoepfle DT, Rangel A: Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J Neurosci 2010, 30:583-590.


The authors establish a model of third-party punishment, outlining the psychological and neural processes contributing to the human capacity to learn, follow, and enforce social norms.


The authors suggest a shift towards a more person-focused account of moral judgment and decision-making, wherein more attention is paid to the moral character of actors, rather than the moral content of specific acts. Moreover, the authors demonstrate that such an approach can clarify findings that most prevailing theories of morality fail to explain, or characterize merely as biases in moral judgment.


The authors observed that statistical perceptions of behavioral frequency drive impression updating on neural and behavioral levels, across the domains of both moral character and ability. Specifically, behaviors that are seen as diagnostic in our evaluations of others (e.g., highly immoral or highly competent behaviors) are perceived to occur less frequently, and as a result, trigger larger impression updates. Moreover, activity in regions associated with updating based on diagnostic information (e.g., vIPFC and IFG) also co-varies parametrically with perceptions of behavioral frequency.


The authors observed that judgments of moral wrongness regarding transgressions associated with physical harm, dishonesty, and sexual disgust are subserved by distinct neural systems. While the dmPFC was found to be active across all three conditions of moral wrong, this region was also active during scenarios judged not to be morally wrong. Taken together, these results do not support a unified perspective and indicate that the neural dynamics of moral judgment vary as a function of content and context.
