Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Deconstructing intent to reconstruct morality

Fiery Cushman

Mental state representations are a crucial input to human moral judgment. This fact is often summarized by saying that we restrict moral condemnation to 'intentional' harms. This simple description is the beginning of a theory, however, not the end of one. There is rich internal structure to the folk concept of intentional action, which comprises a series of causal relations between mental states, actions and states of affairs in the world. Moral judgment shows nuanced patterns of sensitivity to all three of these elements: mental states (like beliefs and desires), the actions that a person performs, and the consequences of those actions. Deconstructing intentional action into its elemental fragments will enable future theories to reconstruct our understanding of moral judgment.

Address
Department of Psychology, Harvard University, United States

Corresponding author: Cushman, Fiery (cushman@fas.harvard.edu)

Current Opinion in Psychology 2015, **6**:97–103

This review comes from a themed issue on **Morality and ethics**

Edited by **Francesca Gino** and **Shaul Shalvi**

doi:10.1016/j.copsyc.2015.06.003

'Even a dog', opined Oliver Wendall Holmes, Jr., 'knows the difference between being kicked and being stumbled over.' Holmes was a gentle man, and his assertion about canine cognition was surely grounded in conjecture, not research. For humans, however, hard data abounds. One of the oldest, best-documented and most reliable findings in the field of moral psychology is that people consider intentional harm to be worse than accidental harm.

The workhorse of this literature is a 2×2 design that pits negative versus neutral intentions against negative versus neutral outcomes (Figure 1). A person who intends harm (negative intention) but fails to cause it (neutral outcome) has committed an *attempted* transgression, whereas a person who intends no harm and yet causes it has committed an *accidental* transgression. This basic method has played a pivotal role in studies of moral reasoning among adults [1–3] and children [4–7] and, more recently, has been used to investigate the neural substrates of moral judgment [8–10] and its disruption in clinical populations [11,12,13•,14].

The great virtue of this 2×2 design is elegant simplicity, yet the same simplicity is its gravest flaw. It obscures the underlying complexity of the concept of intentional action and, consequently, diverts our attention from the subtly and diversity of moral evaluations. In recent years, more precise models of intentional action have launched several new and productive directions in moral judgment research. Further advances, in years to come, will surely do the same.

Deconstructing intent

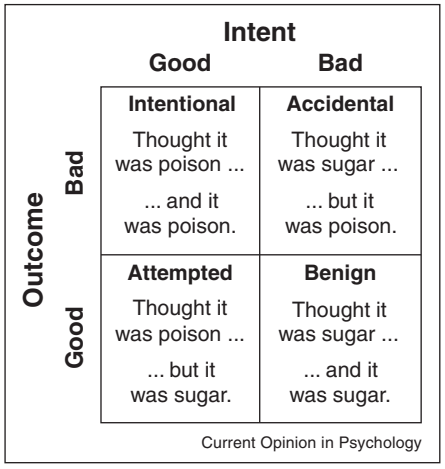
We often attribute other's behaviors to intentional action [15–17]. In other words, the folk concept of 'intentional action' is a causal theory: It explains how mental states cause physical events (Figure 2). And, as several generations of careful research attest, it plays a foundational role in processes of moral judgment [16,18,19••,20,21•,22].

The folk theory of intentional action centers on the concept of a plan. A plan might be very complex, like taking a vacation in China, or very simple, like asking a friend to pass the salt. Either way, its essential function is to link actions to outcomes. In other words, a plan is a mental state representation of an action (or several) that will be performed in order to achieve a goal. According to the folk theory, plans are constructed in response to desires (which establish goals) and beliefs (about action/outcome relations). Planning is also constrained by any undesirable side effects that are foreseen.

In order for a plan to cause behavior, a person must act on it; this is the next step in the canonical causal model of intentional action. When we say that a person acts 'willfully', 'volitionally', 'purposefully', among others, we typically mean that are enacting a plan. Moreover, we assume that plans are enacted with conscious awareness [23]. Finally, voluntary actions may cause certain events in the world — reaching China, shaking salt or, perhaps, killing a person. These events may be intended, foreseen, or wholly accidental.

In summary, when we say that a person has *intentionally* traveled to China, we imply a multi-part causal sequence: (1) They formed mental states including a plan to travel to China, (2) they carried out the actions entailed in that plan, and (3) those actions caused them to be in China. This is why the traditional 2×2 design that purports to cast 'intent' against 'outcome' invites some confusion. The key inputs to moral judgment are rarely isolated from action — a pure mental state ('John intended harm [without acting on it]') or a pure outcome ('harm happened [whatever may have caused it]'). Rather, the inputs are

Figure 1



A factorial combination of intent and outcome yields four basic categories of conduct.
Source: Adapted from Young *et al*., 2007 and Martin & Cushman 2014.

causal relation to action, such as 'John's intent caused him to act', or 'John's action caused harm.'

Of course, intentional action is not the only causal theory of behavior. Ordinary people also attribute behavior to other mental processes that do not involve planning, such as reflex and habit. Curiously, however, people tend not to attribute habitual or reflexive actions to the person who caused them [24]. For instance, if a person intentionally shakes a bee off of their hand, we say that they 'caused their hand to move', but if, instead, they reflexively withdraw their hand from a bee, we do not find it appropriate to say that they 'caused their hand to move'. Rather, we say that it was caused by the bee, or the reflex. Thus, while there is some sense in which we recognize multiple mental-state based causes of human action, there is also an important priority given to intentional actions: We are especially likely to view intentional actions as authored by ourselves.

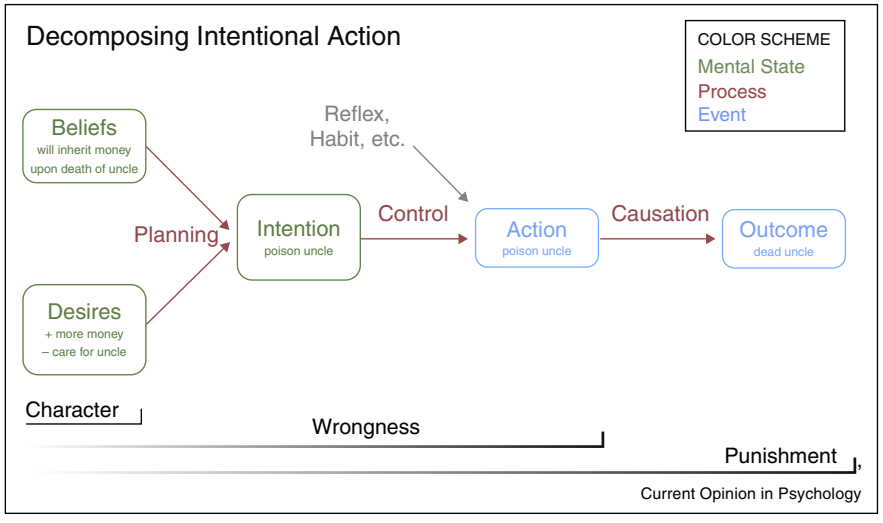
Reconstructing morality

Kinds of moral evaluation

Current theories largely agree that morality is not a single, unified process, but rather a hodge-podge of different cognitive mechanisms [25–27]. For instance, researchers have identified dissociations among kinds of moral evaluation: the character of persons, the wrongness of their actions, the punishment that they deserve, and so forth. In large part, these categories of moral evaluation are distinguished by their sensitivities to different parts of the folk theory of intentional action.

Several studies of this kind expose the dissociation between moral judgments of a person's character and those of their actions. For instance, people judge a wealthy individual to have an impoverished moral character if he chooses to engrave his face on a marble tabletop; yet, they do not consider this action to be morally impermissible [28]. Similarly, they condemn the moral character of a person who would kill one individual in order to save several others, and yet consider the action to be morally praiseworthy [29•]. What underlies these dissociations? Most accounts focus on the idea that attributions of moral character depend on inferences about a person's desires [30•,31]. Imagine, for instance, a person who has wished for years that his elderly neighbor would die of a heart attack. This person has done nothing wrong, and certainly

Figure 2



A decomposition of intentional action into elemental features. As an example, the abstract categories (e.g., 'intention') are illustrated in the context of a harmful moral violation ('poison uncle' in order to gain an inheritance).

our laws make no provision for his punishment, yet the mere presence of a malicious desire speaks strongly against his moral character. This sufficiency of mere desire as a basis for character attribution sets it apart from other categories of moral judgment such as wrongness and punishment.

Another line of research illustrates dissociation between these latter two judgments: wrongness and punishment. In particular, judgments of punishment are sensitive to unintended variation outcomes caused, whereas wrongness and character show less sensitivity. (Meanwhile, wrongness and punishments are alike in that both show a strong sensitivity to a person's culpable mental states.) The unique outcome sensitivity of punishment comprises two parts. First, a person is held to be punishable and blameworthy largely to the extent that they are causally responsible for a harm [32–34]; second, the magnitude of punishment and blame typically scales with the degree of harm caused [35,36]. Both of these dimensions are largely absent from judgments of moral wrongness, which instead track whether a person acts under the belief they may be causally responsible for harm, and the degree of harm they believe will occur as a result of their action [2,37,38••].

To summarize, then, judgments of moral character depend mostly upon a person's *desires*, prosocial or antisocial. Judgments of moral wrongness depend upon a person's *actions* and the beliefs and desires that cause those actions. Finally, judgments of punishment blame depend not only on actions and mental states, but also the *outcomes* caused by their actions (Figure 2).

What are the unique functions that underlie these mechanistic dissociations? An obvious function for judgments of moral character is to help us decide who to trust — or to avoid — in future social interactions. In the literature on the evolution of prosocial behavior, this is referred to as 'partner choice', and evidence suggests that it relies on the assessment of whether a person desired to cause help or harm [39,40]. Meanwhile, an obvious function for punishment is to modify the behavior of others, converting them from harm-doers to do-gooders. This alternative function is sometimes termed 'partner control' [41]. The characteristic reliance of reward and punishment decisions on outcomes (and not just intention and actions) may reflect the pedagogical demands of partner control [42].

This leaves 'moral wrongness' as the odd man out. What functional perspective explains a category of moral judgment that focuses on the *actions* that person performs? One intriguing possibility is that the primitive concept of 'wrongful action' does not reflect the functional demand of regulating others' behavior, but rather the demand of regulating one's own behavior. In other words, the primitive concept of 'morally wrong' may be 'wrong for me' [43,44].

Kinds of moral violation

Just as the moral domain comprises several types of judgment, it also comprises several types of violation. The term 'moral foundations' has been famously applied to the distinctions among diverse criteria for identifying moral violations [45]. For instance, in the 'harm' foundation, moral violations are identified by assessing whether an individual has imposed unjustified suffering on another individual. By contrast, in the 'purity' foundation, moral violations are identified by assessing whether an object of sacred value has been corrupted by the influence of the profane.

Recent research shows that harm and purity violations are identified based on different features of intentional action [46,47]. Specifically, harm violations are identified more by harmful intent than purity violations are by impure intent. For example, people tend to judge that you have not acted wrongly if you accidentally serve a person a dish with an ingredient that they are severely allergic to. However, they tend to judge that you have acted quite wrongly if you accidentally sleep with a person you did not know was your long-lost sister. Harm somebody accidentally and you are mostly off the hook; defile yourself accidentally, and you are very much on it.

A dominant interpretation of these findings posits that harm relies more on intent, and purity more on an outcome [44]. According to one account, this occurs because purity violations are designed principally to regulate one's own behavior, and people's guilt following their own transgressions is driven more by outcome than by intent [48]. As we have seen, however, moral judgments of harmful acts rarely follow from the representation of 'isolated' intent, or outcome. Rather, they more often depend on the causal connections intent and action, or between action and outcome. Is the same true of purity?

In fact, there is reason to suppose that neither of these casual relationships is as important for this domain of transgression. On the one hand, an impure act can be judged wrong even in the absence of intent to defile [46,47]. On the other hand, purity violations clearly do not depend on a causal connection between the transgressive act and some further outcome. On the contrary, it is wrong to have sexual contact with a sibling, or to eat the family dog, or to spit on a religious icon, not because of the consequences of these actions, but simply as a property of the actions themselves [49,50]. This indicates a curious category of moral judgment that focuses on properties of an action alone — divorced both from its causes (e.g., mental states) and from its consequences (e.g., harm or defilement). One recent pair of proposals accounts for

such act-based moral evaluations in terms of contemporary neurobiological and computational models of learning and decision-making [51,52]. Whether these can be successfully applied to the domain of purity violation, however, remains an open question.

Blueprints

As we have seen, several of the most exciting recent developments in moral judgment research are grounded in a detailed decomposition of the theory of intentional action into constituent elements. This same strategy holds the promise to catalyze new insights across a range of additional topics in the field.

Moral development

Piaget inaugurated the study of moral psychology when, in 1930, he interviewed children in Geneva about misbehavior, the rules of marbles, and moral authority [53]. Most famously, Piaget showed that young children assigned moral condemnation mostly based on outcomes, while older children are relatively more influenced by intent. For instance, he found that young children judged it worse to accidentally spill a large pool of ink on a desk than to deliberately make a pinpoint ink stain with a pen, while older children make the reverse judgment. Since then, great quantities of ink have been spilled in a sprawling and controversial literature on the development of intent-based judgment in childhood [5,6,54–60].

This literature centers on debate over *when* children begin to make intent-based judgments, with proposals as old as 8 years [53] and as young as 8 months [61]. Relatively less studied, however, is *what* elements of the theory of intentional action underlie children' judgments across development. As we have seen, there are multiple and conceptually distinct candidates, but there is little consensus about their development. For instance, while some past research indicates that children attain adult-like performance in the judgment of intentional harms earlier than the judgment of negligent harms [55], other research indicates that children actually *over-apply* the concept of negligent action at the youngest ages [62]. Moreover, multiple studies have now shown a dissociation in the developmental acquisition of adult-like judgments of attempted harm (which tend to emerge early) versus accidental harm (which tend to emerge later) [6,63]. It remains for future research to resolve discrepant findings, presumably through a more fine-grained analysis of the constituent mental state concepts that together comprise the adult theory of intentional action.

Planning and negligence

Although the literature is full of studies of accidental harm and intentional harm, far fewer target negligence — a state of affairs intermediate between accidental and intentional harms. As noted above, a negligent harm is not intended, but could have been prevented with further care. In essence, negligence involves a failure of planning. An individual who acts negligently forms and executes a plan of action that they ought to have rejected, had she considered more fully the likely consequences of her behavior. The concept of negligence suffuses our moral language. When we say that a person is 'thoughtless', 'careless' or 'inconsiderate', we are not accusing them of harboring malicious desires; neither do we completely exculpate their actions as mere accidents. Their moral failing consists in planning that is insufficient or misdirected: an inattention to the likely consequences of their action.

One recent theory of moral judgment that does an admirable job of acknowledging negligence is the 'path model' of blame due to Malle and colleagues [19••]. In the event that a person causes harm unintentionally, the model posits that we ask whether they had both the 'obligation' and 'capacity' to have prevented their harmful conduct. The latter criterion, capacity, is a concept as loaded as it is pivotal. Linguists and philosophers call it a 'modal', meaning that it specifies counterfactual possibility — a person *could* have prevented their harmful conduct, even though they did not. In essence, negligence constitutes a comparison between the planning a person actually engaged in, and the planning that constitutes a 'reasonable' and superior counterfactual alternative. As future research clarifies how we make judgments of negligent conduct, an investigation of the relationship between moral judgment and modal judgment is likely to be essential [64].

Control and causation

We typically forgive harmful actions that are uncontrollable; when, for instance, a person sneezes on your cake, hits you during a seizure or runs into your car when her brakes fail. On what basis are these harms forgiven? One possibility is that we forgive such actions because the agent's lack of control implies a lack of intent; that is, that the spoiled cake, bruised cheek and wrecked car were neither desired nor planned outcomes. This explanation is intuitive, and it has some empirical support [65]. Yet, it is easy to devise cases where this provides a poor explanation for the forgiveness of uncontrollable action. Consider, for instance, an individual who suffers from a large tumor that produces an overwhelming urge to commit acts of violence. In this person the violent acts may well be desired and planned, yet the individual still seems less morally responsible for their actions than they would be in the absence of the tumor. In order to explain such cases, it is helpful to remember that theory of intentional action is, at its heart, a theory of causation. The key feature of the tumor may not be that it changes our assessment of the agent's mental states, but rather that it changes our assessment of the agent's causal power. Several studies provide indirect support for this view [66•,67], which deserves further investigation.

Habit and reflex

The distinction between habitual and planned action is among psychology's oldest and most venerable. Habits are formed as rewards 'stamp in' associations between stimuli and responses. Habits are quick and computationally cheap to deploy, but they persist in selecting globally adaptive responses even when local circumstances change (e.g., you might habitually take the bus north to your home even though today you needed to pick up dry cleaning to the south). Planned action instead derives from search over an internal causal model — a person simulates potential sequences of behaviors and then predicts their likely consequences. Planning is slower and more computationally demanding, but can more flexibly adapt to knowledge of changing circumstances.

Given the pervasive role of both mechanisms in determining human behavior, it is remarkable that nearly all research on theory of mind — easily hundreds, if not thousands of studies — assesses how we reason about mental states relevant to planned action: beliefs, desires, goals and plans. By contrast, at best a handful of studies assess how we reason about automatic action [68–70], and few if any focus specifically on the notion of habitual action. Of course, ordinary people talk about habits and apparently have little difficulty identifying and reasoning about them, suggesting that theory of mind encompasses a concept of habitual action. This raises several promising directions for future research. First, how accurately does the folk concept of habitual action reflect known properties of the natural kind? Second, do people accurately infer when others' actions are produced by planning versus habits, or is there a bias to attribute behavior to one or the other cause? Finally, how do people make moral judgments of habitual actions?

Neural basis of mental state attribution

Over the past fifteen years researchers have made a tremendous effort to understand the neural basis of mental state representation and inference. Moral judgment has proved itself an especially useful testbed for research into mental state inference because people spontaneously incorporate mental state inferences into their moral judgments, but with substantial variability both across individuals and items. Researchers have been able to make particularly impressive progress in characterizing one node in a network of brain regions associated with mental state representation: the right temporoparietal junction (rTPJ). For instance, it is now known that disruption of the rTPJ by transcranial magnetic stimulation impairs mental state representation [8]; that the multivariate pattern within rTPJ represents features relevant to moral judgment [13•]; and that populations showing impaired mental state inference also show abnormal patterns within rTPJ [11].

Despite considerable agreement about *where* in the brain mental state representations are computed and deployed, however, we remain remarkably ignorant of *how*. Progress on this question is likely to come through the convergence of low-level theories of neural computation with high-level theories of psychological organization [71]. In other words, deconstructing the psychology of intentional action may help us to reconstruct the neurobiology of mental state attribution, including in the moral domain. This effort will surely be aided by the development of sophisticated and computationally precise formal models of mental state inference [17], including in the context of moral judgment [72•].

Conclusion

Perhaps dogs know the difference between being kicked and being stumbled over; certainly humans do. But humans know much more than this. They know the significance of being loved, respected, neglected and despised; the difference between these attitudes and actions; the distinction between mere action and material consequence. These elements of the folk theory of intentional action each contribute in unique and subtle ways to our capacity for the moral evaluation of others' behavior, and the moral regulation of our own behavior. By deconstructing the theory of intentional action, we can reconstruct a more complete model of human morality.

Conflict of interest statement

The author has no conflicts of interest to disclose in regards to this manuscript this submission.

Acknowledgements

Thanks to Justin Martin and Ryan Miller for valuable feedback. This research was supported by grant N00014-14-1-0800 from the Office of Naval Research, as well as through the support of a grant through the Prospective Psychology project from the John Templeton Foundation. The opinions expressed in this publication are those of the author's and do not necessarily reflect the views of the Office of Naval Research or of the John Templeton Foundation.

References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. Cushman FA, Dreber A, Wang Y, Costa J: **Accidental outcomes guide punishment in a 'trembling hand' game**. *PLOS One* 2009, **4**:e6699 doi:6610.1371/journal.pone.0006699.

2. Cushman F: **Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment**. *Cognition* 2008, **108**:353-380 http://dx.doi.org/10.1016/J.Cognition.2008.03.006.

3. Young L, Saxe R: **The role of intent for distinct moral domains**. *Cognition* 2011, **120**:202-214.

4. Cushman FA, Sheketoff R, Wharton S, Carey S: **The development of intent-based moral judgment**. *Cognition* 2013, **127**:6-21 http://dx.doi.org/10.1016/J. Cognition.2012.11.008.

5. Killen M, Mulvey KL, Richardson C, Jampol N, Woodward A: **The accidental transgressor: morally-relevant theory of mind**. *Cognition* 2011, **119**:197-215.

6. Zelazo PD, Helwig CC, Lau A: **Intention, act, and outcome in behavioral prediction and moral judgment**. *Child Dev* 1996, **67**:2478-2492.

7. Vaish A, Carpenter M, Tomasello M: **Young children selectively avoid helping people with harmful intentions**. *Child Dev* 2010, **81**:1661-1669.

8. Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R: **Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments**. *Proc Natl Acad Sci* 2010, **107**:6753.

9. Kliemann D, Young L, Scholz J, Saxe R: **The influence of prior record on moral judgment**. *Neuropsychologia* 2008, **46**: 2949-2957.

10. Young L, Cushman F, Hauser M, Saxe R: **The neural basis of the interaction between theory of mind and moral judgment**. *Proc Natl Acad Sci U S A* 2007, **104**:8235-8240 http://dx.doi.org/10.1073/Pnas.0701408104.

11. Moran JM *et al.*: **Impaired theory of mind for moral judgment in high-functioning autism**. *Proc Natl Acad Sci* 2011, **108**:2688.

12. Young L *et al.*: **Damage to ventromedial prefrontal cortex impairs judgment of harmful intent**. *Neuron* 2010, **65**:845-851.

13. • Koster-Hale J, Saxe R, Dungan J, Young LL: **Decoding moral judgments from neural representations of intentions**. *Proc Natl Acad Sci* 2013, **110**:5648-5653.
A pioneering application of multi-voxel pattern analysis to the rTPJ during moral judgement, showing that this brain region encodes dimensions of the stimulus not apparent in the univariate response. Additionally, encoding of these dimensions is significantly weaker among high-functioning individuals on the autism spectrum.

14. Young L, Koenigs M, Kruepke M, Newman JP: **Psychopathy increases perceived moral permissibility of accidents**. *J Abnormal Psychol* 2012, **121**:659.

15. Malle BF: **How people explain behaivor: a new theoretical framework**. *Personal Soc Psychol Rev* 1999, **3**:23-48.

16. Heider F: *The Psychology of Interpersonal Relations*. Wiley; 1958.

17. Baker CL, Saxe R, Tenenbaum JB: **Action understanding as inverse planning**. *Cognition* 2009, **113**:329-349.

18. Weiner B: *Judgments of Responsibility: A Foundation For a Theory of Social Conduct*. Guilford Press; 1995.

19. •• Malle BF, Guglielmo S, Monroe AE: **A theory of blame**. *Psychol Inquiry* 2014, **25**:147-186.
A comprehensive effort to link models of moral judgment with a sophisticated understanding of constituent elements of the folk theory of moral judgment.

20. Alicke M: **Culpable control and the psychology of blame**. *Psychol Bull* 2000, **126**:556-574.

21. • Kiley Hamlin J, Ullman T, Tenenbaum J, Goodman N, Baker C: **The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model**. *Dev Sci* 2013, **16**:209-226.
Unusual in the literature on social cognition, this paper combines traditional methods in infant cognitive development with a formal Bayesian model of mental state inference to show how children rely not just on direct perceptual features to make moral judgments, but also on inferred mental states.

22. Mikhail J: *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press; 2011.

23. Malle BF, Knobe J: **The folk concept of intentionality**. *J Exp Soc Psychol* 1997, **33**:101-121.

24. Knobe J, Nichols S: **Free will and the bounds of the self**. *Oxford Handbook of Free Will*. 2nd edn. New York: Oxford University Press; 2011, 530-554.

25. Greene J: *Moral tribes. Emotion, Reason and the Gap between Us and Them*. New York: The Penguin Press; 2013.

26. Haidt J: *The Righteous Mind*. Pantheon; 2012.

27. Young L, Dungan J: **Where in the brain is morality? Everywhere and maybe nowhere**. *Soc Neurosci* 2012, **7**:1-10.

28. Tannenbaum D, Uhlmann EL, Diermeier D: **Moral signals, public outrage, and immaterial harms**. *J Exp Soc Psychol* 2011, **47**:1249-1254.

29. • Uhlmann EL, Zhu LL, Tannenbaum D: **When it takes a bad person to do the right thing**. *Cognition* 2013, **126**:326-334.
Demonstrates dissociations between the evaluation of a person's actions and the evaluation of their moral character as informed by their actions.

30. • Chakroff A, Young L: **Harmful situations, impure people: an attribution asymmetry across moral domains**. *Cognition* 2015, **136**:30-37.
Argues that people are disposed to situational attributions of violations in the harm domain, but personal attributions of violations in the purity domain.

31. Inbar Y, Pizarro DA, Cushman F: **Benefiting From misfortune: when harmless actions are judged to be morally blameworthy**. *Personal Soc Psychol Bull* 2012, **38**:52-62 http://dx.doi.org/10.1177/0146167211430232.

32. Cushman F, Dreber A, Wang Y, Costa J: **Accidental outcomes guide punishment in a 'Trembling Hand' game**. *PLOS ONE* 2009, **4** http://dx.doi.org/10.1371/journal.pone.0006699 doi:ARTN e6699.

33. Mazzocco P, Alicke M, Davis T: **On the robustness of outcome bias: no constraint by prior culpability**. *Basic Appl Soc Psychol* 2004, **26**:131-146.

34. Gino F, Moore D, Bazerman M: *No Harm, No Foul: The Outcome Bias in Ethical Judgments*. 2008.

35. Robinson PH, Darley JM: *Justice, Liability and Blame*. Westview PRess; 1995.

36. Carlsmith K, Darley J, Robinson P: **Why do we punish? Deterrence and just deserts as motives for punishment**. *J Personal Soc Psychol* 2002, **83**:284-299.

37. Young L, Nichols S, Saxe R: **Investigating the neural and cognitive basis of moral luck: it is not what you do but what you know**. *Rev Philos Psychol* 2010:1-17.

38. •• Mikhail J: *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press; 2011.
An ambitious project that aims to provide a complete and formal descriptive model of human moral judgment and the mental state inferences that support it.

39. Bull J, Rice W: **Distinguishing mechanisms for the evolution of co-operation**. *J Theor Biol* 1991, **149**:63-74.

40. Noe R: **A veto game played by baboons: a challenge to the use of the Prisoner's Dilemma as a paradigm of reciprocity and cooperation**. *Anim Behav* 1990, **39**:78-90.

41. Baumard N, André J-B, Sperber D: **A mutualistic approach to morality: the evolution of fairness by partner choice**. *Behav Brain Sci* 2013, **36**:59-78.

42. Martin JW, Cushman F: *The Adaptive Logic of Moral Luck. The Blackwell Companion to Experimental Philosophy*. Wiley-Blackwell; 2015.

43. Miller R, Cushman F: **Aversive for me, wrong for you: first-person behavioral aversions underlie the moral condemnation of harm**. *Soc Personal Psychol Compass* 2013, **7**:707-718.

44. Young L, Tsoi L: **When mental states matter, when they don't, and what that means for morality**. *Social and Personal Psychol Compass* 2013, **7**:585-604.

45. Graham J *et al.*: **Mapping the moral domain**. *J Personal Social Psychol* 2011, **101**:366.

46. Young L, Saxe R: **When ignorance is no excuse: different roles for intent across moral domains**. *Cognition* 2011, **120**:202-214.

47. Russell PS, Giner-Sorolla R: **Moral anger, but not moral disgust, responds to intentionality**. *Emotion* 2011, **11**:233.

48. McGraw KM: **Guilt following transgression: an attribution of responsibility approach**. *J Personal Soc Psychol* 1987, **53**:247.

49. Graham J, Haidt J, Nosek B: **Liberals and conservatives use different sets of moral foundations**. *J Personal Soc Psychol* 2009, **96**.

50. Haidt J, Koller S, Dias M: **Affect, culture and morality, or, is it wrong to eat your dog?** *J Personal Social Psychol* 1993, **65**: 613-628.

51. Crockett MJ: **Models of morality**. *Trends Cogn Sci* 2013, **17**: 363-366.

52. Cushman F: **Action, outcome, and value: a dual-system framework for morality**. *Personal Soc Psychol Rev* 2013, **17**:273-292 http://dx.doi.org/10.1177/1088868313495594.

53. Piaget J: *The Moral Judgment of the Child*. Free Press; 1965/1932.

54. Hebble PW: **Development of elementary school childrens judgment of intent**. *Child Dev* 1971, **42**:583-588.

55. Yuill N, Perner J: **Intentionality and knowledge in childrens' judgments of actors responsibility and recipients emotional reaction**. *Dev Psychol* 1988, **24**:358-365.

56. Shultz TR, Wright K, Schleifer M: **Assignment of moral responsibility and punishment**. *Child Dev* 1986, **57**:177-184.

57. Imamoglu EO: **Children's awareness and usage of intention cues**. *Child Dev* 1975, **46**.

58. Baird JA, Astington JW: **The role of mental state understanding in the development of moral cognition and moral action**. *New Direct Child Adolescent Dev* 2004, **103**:37-49.

59. Costanzo P, Coie J, Grumet J, Farnill D: **A reexamination of the effects of intent and consequence on children's moral judgments**. *Child Dev* 1973, **44**:154-161.

60. Armsby RE: **A reexamination of the development of moral judgments in children**. *Child Dev* 1971:1241-1248.

61. Hamlin JK: **Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations**. *Cognition* 2013, **128**:451-474.

62. Nobes G, Panagiotaki G, Pawson C: **The influence of negligence, intention and outcome on children's moral judgments**. *J Exp Child Psychol* 2009, **104**:382-397.

63. Cushman F, Sheketoff R, Wharton S, Carey S: **The development of intent-based moral judgment**. *Cognition* 2013, **127**:6-21 http://dx.doi.org/10.1016/J.Cognition.2012.11.008.

64. Knobe J, Szabó ZG: **Modals with a taste of the deontic**. *Semantics Pragmatics* 2013, **6**:1-42.

65. Fincham F, ROBERTS C: **Intervening causation and the mitigation of responsibility for harm doing. II: The role of limited ....** *J Exp Soc Psychol* 1985, **21**:178-194.

66. • Phillips J, Shaw A: **Manipulating morality: third-party intentions alter moral judgments by changing causal reasoning**. *Cogn Sci* 2014 http://dx.doi.org/10.1111/cogs.12194. [Epub ahead of print].
Illustrates the importance of intental action as a causal explanation of behavior, and of causal representation as a key input to moral judgment.

67. Darley J, Carlsmith K, Robinson P: **Incapacitation and just deserts as motives for punishment**. *Law Hum Behav* 2000, **24**:659-683.

68. Uhlmann EL, Nosek BA: **My culture made me do it**. *Soc Psychol* 2015.

69. Cameron CD, Payne BK, Knobe J: **Do theories of implicit race bias change moral judgments?** *Soc Justice Res* 2010, **23**: 272-289.

70. Pizarro DA, Uhlmann E, Salovey P: **Asymmetry in judgments of moral blame and praise: the role of perceived metadesires**. *Psychol Sci* 2003, **14**:267-272.

71. Koster-Hale J, Saxe R: **Theory of mind: a neural prediction problem**. *Neuron* 2013, **79**:836-848.

72. • Kleiman-Weiner M, Gerstenberg T, Levine S, Tenenbaum JB: **Inference of intention and permissibility in moral decision making**. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. 2015.