A Computational Account of Complex Moral Judgement

Lillian Liu, Pat Langley, and Ben Meadows

Department of Computer Science, University of Auckland
Private Bag 92019, Auckland 1142, New Zealand
{lliu113@aucklanduni.ac.nz, patrick.w.langley@gmail.com, bmea011@aucklanduni.ac.nz}

Abstract

In this paper, we analyze the task of complex moral judgement from a computational perspective. We present a theoretical framework that posits this process often involves the construction of a coherent explanation for observed behavior in terms of the mental states of the agents involved. We extend the framework to incorporate moral values as numeric annotations on cognitive structures and mitigating factors as influences on these weights that modulate overall moral judgement. In closing, we discuss other work related to moral cognition and behavior understanding, along with directions for additional research.

1 Introduction

Of all the differences between man and the lower animals, the moral sense or conscience is by far the most important.
– Charles Darwin, *The Descent of Man*

We may view morality as a set of principles by which an agent evaluates intentions, decisions, and actions as 'right' or 'wrong', or as 'better' or 'worse'. Morality defines standards of conduct and influences behavior in social situations. Although the criteria on which actions are judged vary across individuals and cultures, moral judgement itself remains a constant throughout humanity. Whereas other animals have demonstrated rudimentary problem-solving abilities and even limited culture, moral reasoning is widely regarded as an inherently human trait. [4]

Consider a simple situation. Kelly sits in a tavern sipping her drink. She is oblivious to another patron, John, until he approaches her and shoves her violently. Kelly topples from her bar stool and hits the ground, yelling out in distress. On the face of it, this scenario seems straightforward. Most would agree that John's actions fell outside the bounds of propriety, and would deem his attack on Kelly as 'immoral'. However, context can make moral judgement far more nuanced. Would John be equally culpable if Kelly had reacted with gaiety instead of distress? What if John's behavior was a retaliation for some previous provocation?

Interaction in a social world often involves complex evaluations. Moral judgement of behavior depends on a complex set of inferences and calculations that, together, produce cognitive structures with associated evaluations. Any theory intended to address the complexities and richness of moral reasoning must characterize this challenging process.

In this paper, we present a computational framework that aims to account for moral judgement. We begin by defining the task in terms of inputs and outputs, and we clarify the class of moral settings that interests us. We proceed to describe our theoretical framework, using John and Kelly's interaction as a running example. We expand on the basic features of our model using variations of the original scenario. Finally, we discuss other work on moral cognition and consider some directions for future research.

2 The Task of Complex Moral Judgement

As moral reasoning is a broad field, it is important to constrain the phenomena we plan to address. To this end, we will distinguish between the task of *moral decision making*, which involves making some choice from a set of possible actions, and *moral judgement*, which involves interpreting and evaluating observed actions in moral terms. The former has been studied widely in both philosophy and psychology (e.g., [8],[14]), and some have even developed computational models to match experimental findings ([1], [17], [18]). However, while philosophy and psychology have examined moral judgement ([16], [19], [5], [11]), computational modeling of the phenomenon has received far less attention.

We will also constrain the moral settings on which we will focus. Some moral tenets are highly specific and simply forbid or require certain behaviors, such as not eating meat on Friday. We will not address such moral rules, both because they are linked to domain content and because they require little analysis by the observer. Although they involve moral cognition, they do not require *complex* moral cognition, which is our primary interest. We maintain that there is a hierarchy of moral strictures ranging from very simple to very complex, and that the latter require some form of multi-step reasoning.

Now that we have clarified the nature of our moral reasoning task, we can state it more formally.

- *Given*: A sequence S of observed actions, including the agent(s) A who performed them;
- *Given*: Knowledge about these and related actions, including their relation to moral concepts;
- *Infer*: An explanation, E, that accounts for S in terms of this knowledge and that posits beliefs, goals, and intentions for A; and
- *Infer*: A moral evaluation of S that takes into account the explanation E.

We will refer to the explanation-evaluation pair as the *moral judgement* of the observed activities. The decomposition of judgement into these two elements is a key aspect of our framework.

We should clarify that, in this paper, we concentrate on the *representation* of moral knowledge and inferences, rather than on the mechanisms that operate over them. However, we believe that moral judgement relies centrally on a process of plan understanding, which involves generation of an explanation for observed actions. Thus, a computational mechanism like UMBRA [13] seems highly relevant to the moral judgement task. This system utilizes knowledge about activities to produce a cohesive explanation of observed events that includes not only inferences about physical relations but also about the mental states of participants. Mechanisms of this sort appear to be necessary, although not sufficient, for complex moral judgement.

3 Basic Theoretical Framework

We can now present a computational framework intended to address the task of complex moral judgement. To help convey the key concepts, we return to our original scenario. We will organize our initial analysis in terms of three claims about the character of moral judgements.

Reconsider the key events in the scenario. First, Kelly is settled on a chair. Second, without warning, John approaches and pushes her off her seat. Third, Kelly reacts with obvious distress. This presentation suggests that John is a malicious agent who has accosted a stranger without provocation. Most people would readily pass judgement on John's actions, as they have violated a basic moral tenet.

An intuitive analysis reveals that one can draw a variety of inferences from the scenario. John acted with knowledge that pushing Kelly would shock her; he intended to carry out the action anyway, he understood the shock resulting from the act would cause Kelly distress, and

Table 1: Representation of the *intentionally-cause-distress* rule, with submethods given in bold.

intentionally-cause-distress(A, B, Action)
← **acts-deliberately(A, Action, Effect)**
← **intends-emotional-effect(A, distressed-about(B, Effect))**
← cause(A, Effect, distressed-about(B, Effect))

he desired this distress to occur. This initial analysis suggests that reasoning about John's transgression focuses less on physical actions than on the agents' mental states. This leads to our first theoretical claim:

- *Complex moral judgement focuses on the mental states of the agents who participate in a scenario.*

More specifically, we posit that the judging entity encodes mental states in terms of each agent's beliefs and goals in the situation. The notation *belief(A, X)* represents A's belief that X is true. In our earlier interpretation, John believes that the physical consequences of his actions would cause Kelly distress: *belief(John, cause(John, feel-shock(Kelly), distressed-about(Kelly, feel-shock(Kelly))))*. Similarly, *goal(A, Y)* denotes A's aim that Y should become true. Such goals capture the notion of *intention*, which plays an important role in moral judgement as it allows attribution of autonomy and culpability. In our informal analysis, John possessed several relevant goals, such as *goal(John, feel-shock(Kelly))*. Together, beliefs and goals (including ones about others' mental states) are inferred during the process of moral judgement.

Further analysis of the example scenario leads naturally to a second theoretical claim:

- *Complex moral judgement depends on rules that abstract away from domain-specific details and that focus on relations among mental states.*

In our example, John would remain just as morally culpable if he had instead pulled Kelly out of her chair rather than pushed her, or even if he had intentionally caused her distress by some completely different means, such as shouting at her across the room. This means that the relevant moral tenet will hold across a wide variety of situations, which suggests that the rules involved abstract away from details of the domain.

For instance, we can characterize John's behavior as fitting some "intentionally cause distress" rule. This rule would apply whenever the consequences of one's actions, combined with the mental states underlying those actions, match its antecedents. Because this rule refers to agents' mental states but not to domain-level predicates, the exact nature of John's act is not relevant, since the relations among his and Kelly's mental states remain unchanged.

Table 1 presents a statement of such a rule, *intentionally-cause-distress*. This includes two domain-independent antecedents: an "acts deliberately" predicate that denotes the agent is carrying out his actions autonomously and an "intends emotional effect" predicate that ensures John's intentions match the action's usual consequence. The rule also incorporates an antecedent requiring that distress actually occurs. The use of these meta-level predicates suggests that multiple layers of abstraction may arise between the inferred mental states and high-level moral concepts.

Note that not all relevant rules must necessarily involve morally-laden predicates. The rule "intends emotional effect", for example, only requires that the acting agent believes that an outcome of some action will have a particular emotional effect and that the outcome is a goal of the acting agent. The predicate *intends-emotional-effect* could appear as an antecedent in

Table 2: Representation of one decomposition of *intends-emotional-effect* (involving distress).

intends-emotional-effect(A, distressed-about(B, X))
← goal(A, distressed-about(B, X))
← belief(A, cause(A, X, distressed-about(B, X)))

non-moral scenarios, such as when an agent intends to experience an emotional effect by having a long bath. Table 2 presents a formalization of one decomposition of this rule that is useful for our scenario. The generality of these moral structures give our proposed framework high contextual sensitivity and enables moral judgement in diverse situations without the need for highly specific rules. This is consistent with the intuitive analysis of how humans engage in complex reasoning about moral situations that we gave earlier.

Recall that our initial analysis assumed John had violated a moral tenet, which we then formalized as a rule with the predicate *intentionally-cause-distress* in its consequent. This rule referred to other abstract predicates in its antecedents, which leads to our third claim:

- *Complex moral judgements involve the linking of rule instances into a connected explanation of observed behavior.*

Together, these chained rule instances constitute a coherent account that takes the form of a tree, where the root incorporates a morally-relevant conceptual predicate. We may view this tree as a proof that the agent's behavior is an instance of this concept. Table 3 gives a formalization of the analysis that we presented earlier.[1]

Note that the explanation incorporates both observations and inferred elements, and includes instances of actions, moral rules, and mental states. The rule instances converge at *intentionally-cause-distress*, which provides an overarching schema consistent with the observed actions: John pushed Kelly because he knew and intended that it would shock her, and he intended to shock her because he wished her to cause her distress. We maintain that explanatory structures of this sort are a primary product of moral judgement.

Table 3: Explanation for John deliberately causing Kelly distress. Bold elements are non-terminal nodes.

intentionally-cause-distress(John, Kelly, shove(John, Kelly))
← **acts-deliberately(John, shove(John, Kelly), feel-shock(Kelly))**
← goal(John, shove(John, Kelly))
← goal(John, feel-shock(Kelly))
← belief(John, cause(John, shove(John, Kelly), feel-shock(Kelly)))
← occurs(shove(John, Kelly))
← cause(John, shove(John, Kelly), feel-shock(Kelly))
← **intends-emotional-effect(John, distressed-about(Kelly, feel-shock(Kelly)))**
← goal(John, distressed-about(Kelly, feel-shock(Kelly)))
← belief(John, cause(John, feel-shock(Kelly), distressed-about(Kelly, feel-shock(Kelly))))
← cause(John, feel-shock(Kelly), distressed-about(Kelly, feel-shock(Kelly)))

[1] The arguments of 'cause' specify the actor, the cause, and the effect, whereas the 'occurs' predicate indicates that an event has taken place.

Table 4: Explanation for John *accidentally* causing Kelly distress. Elements that differ from the deliberate version presented in Table 3 are given in italics.

accidentally-cause-distress(John, Kelly, shove(John, Kelly))
← **acts-deliberately(John, shove(John, Kelly), feel-shock(Kelly))**
← goal(John, shove(John, Kelly))
← goal(John, feel-shock(Kelly))
← belief(John, cause(John, shove(John, Kelly), feel-shock(Kelly)))
← occurs(shove(John, Kelly))
← cause(John, shove(John, Kelly), feel-shock(Kelly))
← ***intends-emotional-effect(John, surprised-about(Kelly, feel-shock(Kelly)))***
← *goal(John, surprised-about(Kelly, feel-shock(Kelly)))*
← *belief(John, cause(John, feel-shock(Kelly), surprised-about(Kelly, feel-shock(Kelly))))*
← cause(John, feel-shock(Kelly), distressed-about(Kelly, feel-shock(Kelly)))

4 Extending the Framework to Moral Values

The framework just described handles some key aspects of moral judgement but not all. Consider an alternative scenario in which John *accidentally* causes Kelly distress. In this version, John still deliberately performs the act of unseating Kelly, but he believes and intends that she will merely be surprised by the painful fall. The immediate effect, and Kelly's emotional response, are the same as in our initial example, but John's intent is different.

Table 4 shows the explanation structure for this case, as captured by the rule *accidentally-cause-distress*. There is only *one* point of difference, in either structure or content, between this and the deliberate version shown in Table 3: the *intends-emotional-effect* predicate is instantiated with John's desire to cause surprise rather than distress. All other world states and mental states remain the same.

Despite this correspondence, the two explanations clearly have different moral status. We would expect someone to judge John much less harshly in the accidental case, even though the effect is the same and the explanation structure is congruous. In both cases, a person evaluating John's behavior makes a number of assumptions concerning the mental states that underpin his actions. We have already noted that it is John's *intent* (or lack thereof) that is relevant to whether he is morally transgressing, but we must extend our framework to handle different degrees of moral response.

Our new example makes it clear that inferring an explanation for observed actions is not sufficient. We also require some moral evaluation to be associated with this cognitive structure. Furthermore, we need more than an evaluation of right or wrong. John's behavior was undesirable in both cases, but worse in the former. This means we require some way to express *degree* of moral evaluation. A person may still pass judgement on John in the accidental scenario – concluding, perhaps, that he should have taken more care. Alternatively, if someone infers that John pushed Kelly off her chair in an attempt to engineer her death, we would expect an even more negative moral evaluation. Conversely, where John accidentally pushes a chair over while sweeping the floor, the negative response would be weaker.

To capture this idea, we extend our framework to distinguish between moral *structures* – the combination of instantiated rules and inferred literals that appear in our explanations – and moral *values* that function as annotations on these explanations, rather than being reified as a direct outcome or relational structures within an explanation. We further distinguish between static values associated with elements of rules and dynamic values associated with elements or literals in a moral explanation. The former include a *default weight* associated with each

conceptual predicate and *upward and downward factors* associated with each rule antecedent. The default weight is a numeric constant that represents, *ceteris paribus*, the basic moral value of a conceptual literal. A negative sign on a moral value corresponds to moral deficit, the extent of which increases with the magnitude of the numeric value. A value with a positive sign corresponds analogously to moral merit. The upward and downward factors are each real numbers in the range -1 to 1 inclusive. They specify, for each element of an instantiated rule, what proportion of the accrued dynamic value should be passed up or down through an explanatory tree.

The 'dynamic value' of an element is initially assigned its default weight when the rule is applied.[2] However, this may change in the course of the judgement process as it is modulated by the moral values of the element's ancestors and children in the tree. The process begins with an upward pass that propagates dynamic values up the explanation tree from the leaf nodes. For each rule in turn, the dynamic value of every rule condition is multiplied by its upward factor and the resulting total added to the dynamic value associated with the instantiated rule head.[3] This new value is passed on in the same manner to any rules in which this literal is involved, until a root is reached. Next comes an analogous downward pass that combines the top-level dynamic value with the downward factors to modulate the scores generated on the upward pass.[4] The result is an explanation in which the moral worth of each action may be informed, to varying degrees, both by the details of subevents below it and the broader context provided by rule instances above it. These value calculations occur only after the explanation has stabilized.

Note that this process describes the judgement of an agent's *activity*. In order to render judgement of the responsible actor, we include a condition in each moral rule of the form *accumulator(X)*, where X is an agent involved in the activity. The purpose of this element is to ascribe a moral value for an agent based on the entire explanation. We assume that the predicate *accumulator* has a default weight of zero, whereas antecedents that refer to it have upward and downward factors of zero and one, respectively. The effect is that the accumulator accretes the sum of the dynamic values for all activities in which the agent is responsible, but does not pass this total to any other part of the tree. At the end of value calculation, the dynamic value for each instance of *accumulator(X)* reflects the accrued moral worth of X's actions, which corresponds to a moral judgement about the agent itself.

5 Extension to Mitigating Factors

We have explained how alternative morally-laden concepts can produce judgements with different values, but this is not enough to handle all variations. Consider another interpretation of our scenario in which John's actions in causing Kelly distress were motivated by Kelly's having insulted John at work earlier. This leads naturally to the notion of *mitigating factors*, which are extenuating circumstances that lessen the severity of a moral transgression (or reduce import of meritorious actions). We will also include aggravating factors, which have the opposite effect, within the term 'mitigation'.

[2]Note that not all rules are necessarily morally laden. For example, *shove(X,Y)* may be neither particularly despicable nor nor especially meritorious when an inanimate object is being shoved. Furthermore, different individuals have different value systems; they may differ in which rules they consider to carry moral substance, so the default values on predicates may vary. However, we claim that the same moral rules and explanatory structures are typically available to all individuals, regardless of the values they associate with them.

[3]One can imagine different schemes for combining the values; we give the simplest for ease of explanation.

[4]Whereas the upward process passes on a fraction of each element's summed value, the downward mechanism passes on a fraction of the top-level value that reflects the contextual influence of the entire explanation.

In our framework, a mitigating factor is another event that, along with its associated structures, is connected to the judged event through some higher-level relation. Thus, reasoning about mitigation takes into account moral structures *other* than those under scrutiny. We judge an action $A1$ taken in revenge for another action $A2$ differently from $A1$ done in $A2$'s absence. We view John's misdeeds as less negative when he is avenging Kelly's earlier actions than when he commits misdeeds for their own sake. Our judgement also takes into account the scale of Kelly's transgressions, as the person he is exacting vengeance against.

Because mitigating factors take the same form as the event being judged, they require no extensions to our basic framework. However, representing their connection to the judged event depends on knowledge about high-level relations like revenge. We will not attempt to specify the rules associated with such predicates here, but we will note that they also typically refer to mental states of the participating agents. We should also note that they may be even more abstract than predicates for the events they relate. One can enact revenge in response to many different activities, and the response may take many different forms. Multiple mitigating factors introduce one complication. Rather than taking the form of a tree with a single root, a moral explanation with multiple mitigators takes the form of a directed acylic graph, with a different root for each high-level relation.

This clarifies the cognitive structures that underlie mitigating factors, but we must also consider the calculation of values that annotate them. Fortunately, the weights associated with rule heads and antecedents described earlier, along with the mechanisms for computing them, appear sufficient for cases that involve mitigation. As before, values are propagated upward upward through the explanation, taking into account the default values and upward scores on rule antecedents. These produce initial values not only for the judged and mitigating events, but also for the high-level relation that connects them.

Downward propagation then modulates these numbers, with the presence of a high-level relation like revenge altering the moral value associated with the judged event, as well as that attached to the responsible agent. This assumes that the downward weight associated with the antecedent for the judged event is positive; for aggravating factors they would be negative and lead to harsher judgements. Note that the combination of upward and downward propagation also explains how the moral value assigned to the mitigating event influences that for the judged event. If John had pushed Kelly for getting him fired, this would be viewed less severely than if she had merely taken his stapler. The framework also supports cases that involve multiple mitigating factors, which it assumes have additive effects on the value judgement.

6 Related Research

A number of distinct literatures address issues related to moral cognition. There is a substantial body of work, in both philosophy and psychology, on moral decision making [8]. Many analyses take a consequentialist rather than a deontological view of moral cognition, and thus focus on the outcome of choices, rather than on the mental state of the acting agent. Some studies [14] have examined the influence of affect and mood in moral decision making, directing attention away from the cognitive processes that also clearly play a key role.

In addition, both social psychologists and philosophers have studied moral judgement, with special interest attached to the influence of intentionality, causality, and emotion ([16], [5], [19]). Classic paradigms involve the assignment of morality scores to variations on moral vignettes [11]. There have also been some research reports on moral judgement in the literature on cognitive development. For example, Piaget [15] presents evidence for a shift in focus on outcomes to one on intentions. He links this change to the acquisition of capacities for abstract thought, which

is consistent with our framework. Kohlberg [12] has expanded on this account, emphasizing the role of conscious reasoning in later stages of moral cognition.

As noted earlier, most computational efforts have focused on moral choice rather than judgement. This includes both applied work that aims to assist humans in making decisions [1] and models of moral cognition that attempt to match psychological findings ([6], [18]). The latter share our concern with the joint roles of cognitive structures and values, but otherwise differ substantially in emphasis. Most work in this paradigm focuses on single decisions rather than sequential plans, leading to simpler mental structures. In contrast, Sun [17] presents a computational model of moral judgement that is motivated by psychological results in the area, but his approach does not incorporate the structured representation of mental states that is central to our account. Iba and Langley [10] discuss both moral judgement and decision making, but they do not present an implemented model of either cognitive activity.

We should also mention research on plan recognition ([7], [2]) and abductive inference of explanations ([9], [3]), each of which have influenced our approach to moral judgement. These paradigms have addressed the problem of inferring agents' mental states from their behaviors, and thus are relevant to generation of the cognitive structures that play the central part in our treatment of moral cognition.

7 Concluding Remarks

In this paper, we contrasted the task of moral decision making with the less-studied problem of moral judgement. We defined the latter task in terms of inputs and outputs, and presented a theoretical framework that, we maintain, covers some central aspects of this cognitive activity. Our basic theoretical claims were that complex moral judgement revolves around inference of the participating agents' mental states, that the rules responsible for these inferences are often highly abstract and domain independent, and that the resulting explanation takes the form of a proof tree with a morally-laden concept at its apex. We also proposed extensions to the framework that associated moral values with elements of explanations and took into account mitigating factors that modulate these values.

Although our framework offers a promising initial account of complex moral judgement, considerable work remains. The next step is to implement a computational model that incorporates the framework's assumptions. To this end, we plan to extend UMBRA, an abductive system for plan understanding that already generates explanatory structures but requires modification to incorporate moral values. We should test the augmented system on sample scenarios that involve various moral concepts, including not only negative judgements, such as those which arise in legal and religious settings, but also positive ones that are associated with desirable behaviors. In addition, we should explore different schemes for calculating moral values and compare them to human ratings on these scenarios. Taken together, these extensions will offer a more complete account of complex moral judgement.

8 Acknowledgements

This research was supported in part by Grant No. N00014-10-1-0487 from the Office of Naval Research. We thank Will Bridewell, Miranda Emery, Alfredo Gabaldon, and Wayne Iba for discussions that influenced the framework we have reported here.

References

[1] Michael Anderson, Susan Leigh Anderson, and Chris Armen. An approach to computing ethics. *Intelligent Systems, IEEE*, 21(4):56–63, 2006.

[2] Douglas Appelt and Martha Pollack. Weighted abduction for plan ascription. *User modeling and user-adapted interaction*, 2(1-2):1–25, 1992.

[3] Will Bridewell and Pat Langley. A computational account of everyday abductive inference. In *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society*, 2011.

[4] Josep Call and Michael Tomasello. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science*, 12(5):187–192, 2008.

[5] Fiery Cushman. Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgement. *Cognition*, 108:353–380, 2008.

[6] Morteza Dehghani, Emmett Tomai, Ken Forbus, and Matthew Klenk. An integrated reasoning approach to moral decision-making. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1280–1286, 2008.

[7] Robert Goldman, Christopher Geib, and Christopher Miller. A new model of plan recognition. In *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*, pages 245–254. Morgan Kaufmann, 1999.

[8] Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814, 2001.

[9] Jerry Hobbs, Mark Stickel, Paul Martin, and Douglas Edwards. Interpretation as abduction. In *Proceedings of the Twenty-Sixth Annual Meeting on Association for Computational Linguistics*, pages 95–103. Association for Computational Linguistics, 1988.

[10] Wayne Iba and Pat Langley. Exploring moral reasoning in a cognitive architecture. In *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society*, Boston, 2011.

[11] Kristine Knutson, Frank Krueger, Michael Koenigs, Angelina Hawley, Jessica Escobedo, Viren Vasudeva, Ralph Adolphs, and Jordan Grafman. Behavioral norms for condensed moral vignettes. *Social Cognition and Affective Neuroscience*, 5(4):378–384, 2010.

[12] Lawrence Kohlberg. *The philosophy of moral development: Moral stages and the idea of justice*. Harper & Row, 1981.

[13] Ben Meadows, Pat Langley, and Miranda Emery. Seeing beyond shadows: Incremental abductive explanation for plan understanding. In *Proceedings of the AAAI Workshop on Plan, Activity, and Intent Recognition*, 2013.

[14] Bernhard Pastötter, Sabine Gleixner, Theresa Neuhauser, and Karl-Heinz T Bäuml. To push or not to push? Affective influences on moral judgment depend on decision frame. *Cognition*, 126(3):373–377, 2013.

[15] Jean Piaget. *The moral judgment of the child*. Routledge & Kegan Paul, London, 1932.

[16] Jared Piazza, Pascale Russell, and Paulo Sousa. Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cognition and Emotion*, 27(4):707–722, 2013.

[17] Ron Sun. Moral judgment, human motivation, and neural networks. *Cognitive Computation*, in press.

[18] Wendell Wallach, Stan Franklin, and Colin Allen. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, 2(3):454–485, 2010.

[19] Robert Woolfork, John Doris, and John Darley. Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100:283–301, 2006.