Computational Models of Moral Perception, Conflict and Elevation

Morteza Dehghani[1], Mary Helen Immordino-Yang[1], Jesse Graham[1], Stacy Marsella[1], Kenneth Forbus[2], Jeremy Ginges[3], Milind Tambe[1] & Rajiv Maheswaran[1]

[1] University of Southern California, CA, USA
[2] Northwestern University, IL, USA
[3] The New School, NY, USA

mdehghan@usc.edu, immordin@usc.edu, jesse.graham@usc.edu marsella@ict.usc.edu, forbus@northwestern.edu, gingesj@newschool.edu, tambe@usc.edu, maheswar@usc.edu

Abstract

Computational models of moral cognition will be critical to the creation of agents and robots that operate autonomously in morally sensitive and complex domains. We propose a framework for developing computational models of moral cognition based on behavioral and neurobiological experimental results and field observations. Specifically, we discuss the following critical issues in building such models: 1. Managing conflicts between different moral concerns; 2. The role of moral perceptions in moral judgments; 3. Mechanisms and consequences of moral emotions; 4. Learning and adjusting moral behavior. Moreover, we discuss computational architectures for building and exploring models of moral cognition at different levels of analysis: individual, small groups and large groups.

1. Introduction

As philosophers have recognized for over 2500 years, the study of human decision-making in social contexts is amazingly complex, involving considerations such as managing conflicts between one's own and one's group's concerns, moral perceptions, moral emotions and learning. Here, we propose to apply computational modeling, in conjunction with behavioral and neurobiological studies, to shed new light on parts of this problem space. Specifically, the goal of this paper is to provide a framework for developing computational models capable of autonomous, human-like decision-making in morally complex situations, where simple cost-benefit analyses are not appropriate or feasible. Such models are critical to the creation of agents and robots that operate in morally sensitive domains (e.g. military operations, medicine). Progress in agent research is bringing us closer to the reality of humans working together with robots and software agents to carry out critical missions in which agents' actions may lead to life/death outcomes for humans or significant loss/savings of property [29]. Such missions will include battlefield deployment and search and rescue missions where computational models of moral cognition would accommodate human moral concerns, and may be needed in situations where obtaining human input is either infeasible (e.g. communication

barriers) or undesirable (e.g., cases with intense time constraints that compromise human decision-making).

Models of moral decision-making can also provide decision support for policy-makers in analyzing cultural and political conflicts revolving around sacred and moral values. Recent work in social and cognitive psychology suggests that moral and sacred values may be critically involved in sustaining seemingly intractable cultural and political conflicts [e.g. 1, 9].

Finally, we expect that these computational models will provide new tools for creating parsimonious yet accurate theories of moral decision-making. Such computational models will force specific commitments about how moral decisions are represented, how tradeoffs between decisions are assessed, and how factors like psychological distance, meta-perception, saliency and emotion may influence decisions. We expect these commitments will raise issues that the abstract specifications of traditional theories cannot accommodate. Additionally, once computationally realized, simulation will allow the model to be systematically explored and manipulated, thereby generating predictions that can be tested in subsequent behavioral and neurobiological research.

Our framework focuses on four critical yet underexplored research questions:

1. Managing conflicts between different moral concerns. How can agents make decisions in situations where they have conflicting moral concerns, such as saving the life of an in-group versus an out-group member, or obeying an authority versus being fair?
2. The role of moral perceptions in moral judgments. How can agents recognize and comprehend other agents' moral concerns and emotions, and how does this recognition process influence their own moral decision-making?
3. Mechanisms and consequences of moral emotions. In humans, social emotions play an important part in understanding and responding to moral situations. How can agents model such emotions and incorporate them into moral decision-making?
4. Learning and adjusting moral behavior. In humans, complex moral reasoning and emotions are shaped within social settings, and dynamically shift to accommodate relevant cultural and social concerns. How can agents be imbued with abilities to learn from the morally significant actions they perceive in others, as well as from their perceptions of others' emotional reactions to their own decisions and actions?

We believe that these questions cover four dimensions of functioning of moral agents, and that each offers targeted opportunities for bridging across behavioral, neurobiological and computational domains. We discuss how our framework can be operationalized into computational models at three levels of granularity: individual moral cognition, small group interactions, and large group interactions. By developing models at three levels of granularity, we seek to support different levels of analysis: intrapsychic, interpsychic, inter-group and intra-group. Below, we discuss the four components of our framework, followed by a brief discussion on the computational modeling architectures researchers can use to realize this framework.

2. Specific Research Thrusts

1. Managing Conflicts Between Different Moral Concerns

Historically, the science of morality has mostly focus on harm and unfairness. Extending earlier work on conflicts between potential harms, Moral Foundations Theory [11] posits multiple intuitive moral sensitivities that cultures build upon to differing degrees. Empirical work on this theory has focused on five kinds of moral concerns (reflecting both virtues and vices), evolved to serve distinct but related social functions. Below, we describe each moral concern:

1. Care/harm. Whatever functional systems made it easy and automatic to connect perceptions of suffering with motivations to care, nurture, and protect are what we call the Care/harm foundation.

2. Fairness/cheating. The Fairness/cheating foundation is triggered by acts of cheating or cooperation by one's own direct interaction partners, or interactions among third parties that one learns about through gossip.

3. Loyalty/betrayal. Among humans, intergroup competition can be decisive for survival. Sports fandom and brand loyalty are examples of current triggers built on this foundation.

4. Authority/subversion. The Authority/subversion foundation is often at work when people interact with and grant legitimacy to modern institutions such as law courts and police departments, and to bosses and leaders of many kinds.

5. Sanctity/degradation. Disgust and the "behavioral immune system" [26] have come to undergird a variety of moral reactions, e.g., to immigrants and sexual deviants [22].

Trolley-type dilemmas [e.g. 21] have been very useful for testing and manipulating conflicts between moral concerns about harmful actions and moral concerns about harmful consequences. However, many of the moral situations confronting artificial agents operating in morally sensitive domains, such as on the battlefield in the heat of the moment, involve tensions between concerns not related to harm, such as caring for an outgroup member vs. remaining loyal to your unit/group/nation. This also applies in cases of intergroup and intercultural conflicts [e.g. 9]. The first component of our framework requires investigating and modeling how people make moral decisions in the face of conflicting moral concerns. This would provide a better understanding of how humans weigh moral concerns in high-pressure situations, and would also be used to improve alignment between humans' and agents' moral decision-making.

2. Exploring the Role of Moral Perceptions in Moral Judgments

The ability to recognize and share another's thoughts and feelings [5, 24] is a basic psychological and neurological entry point into social emotions and moral decision-making [e.g. 17]. One fundamental task when encountering another person is to assess their moral worldview. Deciding, for example, that another entity shares one's moral code will assist in making decisions about how to act. In cases of strong similarity, an agent is able to predict how the other entity will behave, and can confidently coordinate its actions to achieve mutually desired outcomes. Perceptions of similarity in moral world-view predict individuals' desires to cooperate with each other, and influence their action coordinates. While there is a growing interest in moral decision-making [e.g. 13], and a history of work examining how meta-perceptions shape intergroup relations (e.g. perceived similarity is associated with greater cooperation) [e.g. 3], little work has investigated how perceived moral similarities across different domains influence moral decision-making. Moral Foundations Theory [11] suggests that perceptions of moral intentions, emotions, and character in target agents may depend on the relative weighting of different moral concerns in the perceiver.

We suggest that moral decision-making is influenced by the extent to which each agent believes that the other shares relevant moral codes and concerns. By sharing a moral code we mean a belief that certain behaviors, like fairness or hospitality to strangers, are "relevant to right and wrong." Furthermore, while most cultures moralize the abstract idea of fairness, there are considerable cultural differences in what fairness means in particular contexts [14]. Thus, moral decision-making is influenced by the ability of agents to make culturally competent signals that one shares, for example, the other's moral code of fairness. Therefore, computational models of moral cognition need to provide abilities for agents to recognize and comprehend other agents' moral concerns and emotions, and to incorporate the influence of such recognition in their moral decision-making processes. This requires investigating the extent to which specific moral choices (e.g., to reciprocate in an economic game or to care for others in need) may be tied to perceived domain-specific similarities in moral worldview. Related to this component, research is also needed to understand whether "signals" of

similarity/differences can be used in specific moral domains to manipulate moral decision-making in intergroup contexts

3. Mechanisms and Consequences of Moral Emotions

When people engage with the social world, they emotionally react to others' actions as well as to their perceptions of others' desires, values, motivations and beliefs. Psychologically and neurobiologically, moral responses to others' physical and mental situations involve either oppositional, antagonistic emotions like disgust, contempt or anger, or else sympathetic/prosocial emotions like admiration and compassion. The third component of our framework relies on neurobiological and behavioral studies to explore the processes of social reasoning and emotion that set the social context for moral decisions. Their results will help constrain cognitive architectures by clarifying how people retrieve personal memories in social situations with moral content, how cognitive and affective empathy shape their emotions, and the role of self-awareness in responding to a situation.

Many studies of moral judgment focus on reactions to moral violations. In this framework, though, we encourage researchers also to focus on the psychological and neurobiological mechanisms underlying positive moral reactions. Positive social emotions are powerful motivators of behavior; people emulate those whose behavior they admire, and empathize with those whose behavior they find virtuous [e.g. 18]. These positive emotional reactions can precipitate potentially extreme reactions that can be either harmful or beneficial in a security context. No agent capable of moral judgment and decision-making would be complete without them. Understanding what is admirable or appalling to its teammates will be important for making robots and software agents that can be trusted to operate in accordance with military codes and human values.

Basic emotions, such as anger, fear, happiness and sadness, are complex physiological processes that involve an interplay of body and mind [e.g. 4], and depend importantly on brain systems for body regulation (e.g. cardiorespiratory, metabolic, endocrine, somatosensation of physical pain and pleasure). However, the social emotions and judgments involved in morality are even more complex because they involve dynamic interactions between neural systems for body regulation, sensation and systems that support memory retrieval, reasoning, and perspective-taking relative to self processes [59]. In brief, feeling emotions about other people in moral contexts linked to judgments of virtue, purity, fairness and reciprocity, involve brain systems responsible for representing body states, brainstem nuclei, systems related to construction of one's own self, the inferior/posterior sector of the mesial parietal lobe and posterior cingulate (PMC) and dorsomedial prefrontal cortex (dmPFC); [e.g. 2], and systems related to episodic memory (e.g. medial temporal regions including anterior hippocampus) [17] and cognitive empathy (perspective-taking) [25]. The biological reactions to social situations are grounded in the body, as for example, brain systems involved in the direct sensation of physical pain are also involved in the feeling of one's own social or psychological pain [e.g. 6].

Thus, we believe that computational models of moral cognition need to take in to account (a) how moral emotions are induced and felt; (b) how constructs that are central to social emotion and moral judgment individually and collectively contribute to moral processing; and (c) how cultural differences can produce predictable individual differences in emotions, feelings and moral behavior. We believe each of these issues are important for informing design of moral agents and to incorporate results from agent modeling into more comprehensive tests.

4. Learning and Adjusting Moral Behavior

When people interact, they form beliefs and attitudes about others, but that is not all. They actively interpret the situation in light of previous experiences and belief systems in order to build personally relevant meanings that can inform future decisions. Making sense of social situations, therefore, involves dynamic interactions among multiple processes, which can also mutually shift one another [16]. Relevant psychological processes include reasoning about the motivations and intentions of others (cognitive empathy), feeling what they are feeling (affective empathy), and applying semantic and cultural knowledge to interpret the relevance of the situation to their own lives (retrieving and forming memories) [4]. Together, engaging these processes often results in social emotions and moral judgments that lead to the strengthening or revising of personal convictions for future action, in accordance with personal and cultural norms and values [17].

Thus for artificial agents to operate appropriately and efficiently in social settings, they need the ability to change their preferences, concerns, and reasoning strategies to accommodate cultural norms and values. No computational model of moral cognition would be complete without the ability to learn from others' morally significant actions, as well as perceptions of others' emotional reactions to agents' decisions and actions. This is contingent on agents being able to interpret moral emotions, but also it requires mechanisms for learning and revising beliefs and moral preferences.

3. Moral Cognition in Cognitive Architectures

We see the development of computational models as needing to meet the above four requirements, in order to ensure that the autonomous agents we build are capable moral agents. These four requirements of our framework can be modeled at the level of the individual, the small group, and the large group. Next we discuss, for each level, the platform that can be used for modeling.

1. Modeling Individual Morality

We propose that the Companions cognitive architecture [10] can be used as one of the experimental test-beds for our framework. The goal of the Companions architecture is to create software social organisms, i.e. systems that one can treat as collaborators, rather than tools. The Companion architecture focuses on conceptual reasoning and learning, making it a natural for tackling moral reasoning, where the ability to explain why one thinks a choice is admirable or appalling is important. Dehghani's MoralDM [8] model was originally implemented in FIRE, the reasoning engine underlying Companions. MoralDM is a cognitively motivated model of recognition-based moral decision-making. It uses analogical matching to provide human-like robustness and the ability to model different cultures by changing the background stories available to the system. Our recent research shows [7] that human moral decision-making uses analogical reasoning, compatible with reasoning in Companions.

2. Modeling Small Group Interactions

The test-bed for modeling small group interactions in our framework is Marsella's multi-agent social simulation architecture PsychSim [19]. PsychSim provides a bounded set of possible behaviors in the form of best-, worst-, and expected-case analyses that encompass the unpredictability of human behavior, while also identifying the specific conditions that would lead to (for example) undesirable side effects. PsychSim implements theory of mind as a recursive system of beliefs, providing a platform to capture each entity's divergent perspectives, as well as its perspectives on others' perspectives. In addition, our previous work on computational models of emotions [20] demonstrates

how emotional processes can be integrated within PsychSim's theory of mind reasoning. This architecture also combines theory of mind representation with a decision-theoretic capacity that translates the internal mental states of the relative entities into behaviors. Within the framework of this project PsychSim will be extended in several ways to model moral cognition. First, modeling moral perceptions and social reasoning more generally requires a theory of mind capacity that supports agents having beliefs about others as well as goals with respect to others' beliefs, such as wanting their in-group to perceive them as fair. PsychSim's theory of mind representations can be extended to capture moral perceptions. Second, emotional processes integrated in PsychSim can also be extended to model complex moral emotions. Third, although decision-theory is very common in agent architectures, it is likely insufficient to capture the complexities involved in reasoning about moral conflicts, and therefore, the reasoning mechanism of the system will need to be extended to capture the complexities involved in moral tradeoffs.

3. Modeling Large Group Interactions

To model moral behavior in larger groups, we propose using various multi-agent decision-theoretic techniques as test-beds for our framework. These techniques includes bounded-parameter multi-objective Markov decision processes, which model how humans value multiple independent features in situations where many connected decisions are made over time and people may not know how effects will cascade over time [e.g. 27]. For human-agent interaction, an autonomous system may not fully understand the decision context, but may be in a time-critical situation. Thus it is important to have a framework where agents can (1) ask for information yet understand that questions have cost, (2) pass the decision on to others with the understanding that it may not be accepted, or (3) make a decision if previous options are unfeasible. Consider a team of human and non-human agents making decisions together in a distributed environment, i.e., agents may not see or know what other agents know and may not be able to communicate with each other extensively. This leads to different agents having subjective views of the world even if their goals are the same. The significance of this type of modeling to moral cognition is that, particularly in groups, differences in subjective views between agents may give rise to different moral concerns. Also, an agent after a period of absence of interaction with other agents might focus on other priorities that lead to different moral concerns. This platform has the promise to help us develop insight into how groups of agents converge or diverge over differing moral concerns, the bifurcation points that determine the final states and the critical factors that determine how the evolution proceeds. Additionally, the capacity for adjustable autonomy will allow us to explore decision tradeoffs between the agents making morally significant decisions autonomously vs. seeking human input.

4. Discussion

In this paper we proposed a framework for developing computational models of moral cognition based on behavioral and neurobiological experimental results and field observations. The proposed framework focuses on four interrelated open issues in moral decision-making that are critically important for making robots and autonomous agents capable of making human-like decisions in morally complex situations. First, such agents should be able to make decisions in situations where they have conflicting moral concerns, such as caring for an outgroup member vs. remaining loyal to one's group. Second, agents need to recognize and comprehend other agents' moral concerns and emotions. Third, moral emotions play an important part in understanding and responding to moral situations, and agents need to be able to model such emotion and also to incorporate others' moral emotions in their decision-making. Fourth, no agent capable of moral cognition would be complete

without the ability to learn moral behavior. To capture the breadth, complexity and flexibility of human moral thought and behavior, agents need to have various mechanisms for learning moral codes of conduct. We encourage researchers to use this framework via a combination of behavioral, neurobiological, and computational work, leading to new models of commonsense moral reasoning embedded in cognitive architectures. Furthermore, such computational models will allow ethicists' to explore the application of moral issues to the design of our evolving relation with machines.

References

1. Atran, S. and Ginges, J. (2012). Religious and sacred imperatives in human conflict. Science, 336, 855-857.
2. Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. Trends in Cognitive Sciences, 11(2), 49-57.
3. Byrne, D., & Clore, G. L. (1970). A reinforcement model of evaluative responses. Personality: An International Journal, 1, 103-128.
4. Damasio, A. R. (1994/2005). Descartes' error: Emotion, reason and the human brain. London: Penguin Books.
5. Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., et al. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci, 3(10), 1049-1056.
6. Decety, J., & Chaminade, T. (2003). Neural correlates of feeling sympathy. Neuropsychologia, 41(2), 127-138.
7. Dehghani, M., Gentner, D., Forbus, K., Ekhtiari, H. & Sachdeva, S. (2009). Analogy and Moral Decision-Making. In Proceedings of the 2nd International Analogy Conference.
8. Dehghani, M., Forbus, K., Tomai, E. & Klenk, M. (2011). An Integrated Reasoning Approach to Moral Decision-Making. In Anderson, M., Anderson, S (Ed.) Machine Ethics. Cambridge University Press.
9. Dehghani, M., Iliev, R., Sachdeva, S., Atran, S., Ginges, J. & Medin, D. (2009). Emerging sacred values: Iran's nuclear program. Judgment and Decision Making, 4, 7, pp. 930-933.
10. Forbus, K, Klenk, M. and Hinrichs, T. (2009, July/August). Companion Cognitive Systems: Design Goals and Lessons Learned So Far. IEEE Intelligent Systems, 24(4), 36-46.
11. Graham, J., Haidt, J., Koleva, S., et al. (in press). Moral Foundations Theory: The pragmatic validity of moral pluralism. Advances in Experimental Social Psychology.
12. Gray, J. R. (1999). A bias toward short-term thinking in threat-related negative emotional states. Personality and Social Psychology Bulletin, 25(1), 65-75.
13. Greene, J.D., Sommerville, R.B., Nystrom, L.E., et al. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105-2108.
14. Henrich, J., Ensminger, J., McElreath, et al. (2010). Markets, religion, community size, and the evolution of fairnessand punishment. Science, 327, 1480-1484
15. Immordino-Yang, M. H. (2010). Toward a microdevelopmental, interdisciplinary approach to social emotion. Emotion Review, 2(3), 217-220.
16. Immordino-Yang, M. H. (2011). Me, myself and you: Neuropsychological relations between social emotion, self awareness, and morality. Emotion Review, 3(3), 313-315.
17. Immordino-Yang, M. H., Christodoulou, J., & Singh, V. (2012). Rest is not idleness: Implications of the brain's default mode for human development and education. Perspectives on Psychological Science, 7(4), 352 - 364.

18. Immordino-Yang, M. H., McColl, A., Damasio, H., & Damasio, A. (2009). Neural correlates of admiration and compassion. Proceedings of the National Academy of Sciences, 106(19), 8021-8026.
19. Marsella, S. C., Pynadath, D. V., & Read, S. J. (2004). PsychSim: Agent- based modeling of social interactions and influence. Paper in the Proceedings of the 6th International Conference on Cognitive Modeling, Carnegie Mellon University, Pittsburgh, PA.
20. Marsella, S. & Gratch, J. (2006). EMA: A computational model of appraisal dynamics. in Agent Construction and Emotions, Vienna, Austria.
21. Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. Trends in Cognitive Sciences, 11, 143–152.
22. Navarrete, C. D., & Fessler, D. M. T. (2006). Disease avoidance and ethnocentrism: the effects of disease vulnerability and disgust sensitivity on intergroup attitudes. Evolution and Human Behavior, 27, 270-282.
23. Powers, T. M. (2006). Prospects for a Kantian machine. IEEE Intelligent Systems, 21 (4), 46-51.
24. Riggio, R. E., Tucker, J., & Coffaro, D. (1989). Social skills and empathy. Personality and Individual Differences, 10(1), 93-99.
25. Saxe, R., and Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia 43, 1391–1399.
26. Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). Current Directions in Psychological Science, 20, 99-103.
27. Schurr, N., Marecki, J., and Tambe, M. (2009). Improving adjustable autonomy strategies for time-critical domains. In the Eighth International Conference on Autonomous Agents and Multiagent Systems. Budapest, Hungary.
28. Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? Journal of Personality and Social Psychology, 88(6), 895-917.
29. Tambe, M. (2011). Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned. Cambridge University Press