Cognition 167 (2017) 107–123

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Learning a commonsense moral theory

Max Kleiman-Weiner [*], Rebecca Saxe, Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 1 April 2016
Revised 26 February 2017
Accepted 9 March 2017
Available online 27 March 2017

Keywords:
Moral learning
Hierarchical Bayesian models
Social cognition
Moral change
Value alignment

ABSTRACT

We introduce a computational framework for understanding the structure and dynamics of moral learning, with a focus on how people learn to trade off the interests and welfare of different individuals in their social groups and the larger society. We posit a minimal set of cognitive capacities that together can solve this learning problem: (1) an abstract and recursive utility calculus to quantitatively represent welfare trade-offs; (2) hierarchical Bayesian inference to understand the actions and judgments of others; and (3) meta-values for learning by value alignment both externally to the values of others and internally to make moral theories consistent with one's own attachments and feelings. Our model explains how children can build from sparse noisy observations of how a small set of individuals make moral decisions to a broad moral competence, able to support an infinite range of judgments and decisions that generalizes even to people they have never met and situations they have not been in or observed. It also provides insight into the causes and dynamics of moral change across time, including cases when moral change can be rapidly progressive, changing values significantly in just a few generations, and cases when it is likely to move more slowly.

1. Introduction

> *Common sense suggests that each of us should live his own life (autonomy), give special consideration to certain others (obligation), have some significant concern for the general good (neutral values), and treat the people he deals with decently (deontology). It also suggests that these aims may produce serious inner conflict. Common sense doesn't have the last word in ethics or anywhere else, but it has, as J. L. Austin said about ordinary language, the first word: it should be examined before it is discarded.* – Thomas Nagel (1989), The View From Nowhere

Basic to any commonsense notion of human morality is a system of values for trading off the interests and welfare of different people. The complexities of social living confront us with the need to make these trade-offs every day: between our own interests and those of others, between our friends, family or group members versus the larger society, people we know who have been good to us or good to others, and people we have never met before or never will meet. Morality demands some consideration for the welfare of people we dislike, and even in some cases for our sworn enemies. Complex moral concepts such as altruism, fairness, loyalty, justice, virtue and obligation have their roots in these trade-offs, and children are sensitive to them in some form from an early age. Our goal in this paper is to provide a computational framework for understanding how people might learn to make these trade-offs in their decisions and judgments, and the implications of possible learning mechanisms for the dynamics of how a society's collective morality might change over time.

Although some aspects of morality may be innate, and all learning depends in some form on innate structures and mechanisms, there must be a substantial role for learning from experience in how human beings come to see trade-offs among agents' potentially conflicting interests (Mikhail, 2007, 2011). Societies in different places and eras have differed significantly in how they judge these trade-offs should be made (Blake et al., 2015; Henrich et al., 2001; House et al., 2013). For example, while some societies view preferential treatment of kin as a kind of corruption (nepotism), others view it as a moral obligation (what kind of monster hires a stranger instead of his own brother?). Similarly, some cultures emphasize equal obligations to all human beings, while others focus on special obligations to one's own group e.g. nation, ethnic group, etc. Even within societies, different groups, different families, and different individuals may have different standards (Graham, Haidt, & Nosek, 2009). Such large differences both between and within cultures pose a key learning challenge: how to infer and acquire appropriate values, for moral trade-offs of this kind. How do we learn what we owe to each other?

[*] Corresponding author at: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, 46-4053, Cambridge, MA 02139, United States.

E-mail address: maxkw@mit.edu (M. Kleiman-Weiner).

http://dx.doi.org/10.1016/j.cognition.2017.03.005
0010-0277/© 2017 Elsevier B.V. All rights reserved.

Children cannot simply learn case by case from experience how to trade off the interests of specific sets of agents in specific situations. Our moral sense must invoke abstract principles for judging trade-offs among the interests of individuals we have not previously interacted with or who have not interacted with each other. These principles must be general enough to apply to situations that neither we nor anyone we know has experienced. They may also be weighted, such that some principles loom larger or take precedence over others. We will refer to a weighted set of principles for how to value others as a "moral theory," although we recognize this is just one aspect of people's intuitive theories in the moral domain.

The primary data that young children observe are rarely explicit instructions about these abstract principles or their weights (Wright & Bartsch, 2008). More often children observe a combination of reward and punishment tied to the moral status of their own actions, and examples of adults making analogous decisions and judgments about what they (the adults) consider morally appropriate trade-offs. The decisions and judgments children observe typically reflect adults' own moral theories only indirectly and noisily. How do we generalize from sparse, noisy, underdetermined observations of specific instances of moral behavior and judgment to abstract theories of how to value other agents that we can then apply everywhere?

Our main contribution in this paper is to posit and formalize a minimal set of cognitive capacities that people might use to solve this learning problem. Our proposal has three components:

- **An abstract and recursive utility calculus.** Moral theories (for the purposes of trading off different agents' interests) can be formalized as values or weights that an agent attaches to a set of abstract principles for how to factor any other agents' utility functions into their own utility-based decision-making and judgment.
- **Hierarchical Bayesian inference.** Learners can rapidly and reliably infer the weights that other agents attach to these principles from observing their behavior through mechanisms of hierarchical Bayesian inference; enabling moral learning at the level of values on abstract moral principles rather than behavioral imitation.
- **Learning by value alignment.** Learners set their own values guided by meta-values, or principles for what kinds of values they value holding. These meta-values can seek to align learners' moral theories externally with those of others ("We value the values of those we value"), as well as internally, to be consistent with their own attachments and feelings.

Although our focus is on the problems of moral learning and learnability, we will also explore the implications of our learning framework for the dynamics of how moral systems might change within and across generations in a society. Here the challenges are to explain how the same mechanisms that allow for the robust and stable acquisition of a moral theory can under the right circumstances support change into a rather different theory of how others interests are to be valued. Sometimes change can proceed very quickly within the span of one or a few generations; sometimes it is much slower. Often change appears to be progressive in a consistent direction towards more universal, less parochial systems – an "expanding circle" of others whose interests are to be taken into account, in addition to our own and those of the people closest to us (Pinker, 2011; Singer, 1981). What determines when moral change will proceed quickly or slowly? What factors contribute to an expanding circle, and when is that dynamic stable? These questions are much bigger than any answers we can give here, but we will illustrate a few ways in which our learning framework might begin to address them.

The remainder of this introduction presents in more detail our motivation for this framework and the phenomena we seek to explain. The body of the paper then presents one specific way of instantiating these ideas in a mathematical model, and explores its properties through simulation. As first attempts, the models we describe here, though oversimplified in some respects, still capture some interesting features of the problems of moral learning, and potential solutions. We hope these features will be sufficient to point the way forward for future work. We conclude by discussing what is left out of our framework, and ways it could be enriched or extended going forward.

The first key component of our model is the expression of moral values in terms of utility functions, and specifically recursively defined utilities that let one agent take others' utilities as direct contributors to their own utility function. By grounding moral principles in these recursive utilities, we have gained a straightforward method for capturing aspects of moral decision-making in which agents take into account the effects of their actions on the well-being of others, in addition to (or indeed as a fundamental contributor to) their own well-being. The specifics of this welfare are relatively abstract. It could refer to pleasure and harm, but could also include other outcomes with intrinsic value such as "base goods" e.g., achievement and knowledge (Hurka, 2003) or "primary goods" e.g., liberties, opportunities, income (Rawls, 1971; Scanlon, 1975; Sen & Hawthorn, 1988) or even purity and other "moral foundations" (Haidt, 2007). This proposal thus formalizes an intuitive idea of morality as the obligation to treat others as they would wish to be treated (the 'Golden Rule', Popper, 2012; Wattles, 1997); but also as posing a challenge to balance one's own values with those of others (captured in the Jewish sage Hillel's maxim, "If I am not for myself, who will be for me? But if I am only for myself, who am I?"). Different moral principles (as suggested in the opening quote from Nagel) can come into conflict. For instance one might be forced to choose between helping the lives of many anonymous strangers versus helping a single loved one. Quantitative weighting of the various principles is a natural way to resolve these conflicts while capturing ambiguity.

On this view, moral learning is the process of learning how to value (or "weight") the utilities of different groups of people. Young children and even infants make inferences about socially positive actions and people that are consistent with inference over recursive utility functions: being helpful can be understood as one agent taking another agent's utility function into account in their own decision (Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Ullman et al., 2009). Young children also show evidence of weighting the utilities of different individuals, depending on their group membership and social behaviors, in ways that strongly suggest they are guided by abstract moral principles or an intuitive moral theory (Barragan & Dweck, 2014; Hamlin, 2013; Hamlin, Mahajan, Liberman, & Wynn, 2013; Kohlberg, 1981; Powell & Spelke, 2013; Rhodes, 2012; Rhodes & Chalik, 2013; Rhodes & Wellman, 2017; Shaw & Olson, 2012; Smetana, 2006). On the other hand, children do not weight and compose those principles together in a way consistent with their culture until later in development (Hook & Cook, 1979; House et al., 2013; Sigelman & Waitzman, 1991). Different cultures or subcultures might weight these principles in different ways, generating different moral theories (Graham, Meindl, Beall, Johnson, & Zhang, 2016; Schäfer, Haun, & Tomasello, 2015) and posing an inferential challenge for learners who cannot be pre-programmed with a single set of weights. But under this view, it would be part of the human universal core of morality – and not something that needs to be inferred – to have the capacity and inclination to assign non-zero weight to the welfare of others.