Cognition 167 (2017) 172–190

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Moral Learning: Conceptual foundations and normative relevance

Peter Railton

Department of Philosophy, University of Michigan, 2215 Angell Hall, 435 South State Street, Ann Arbor, MI 48109-1003, United States

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 6 August 2016
Accepted 25 August 2016
Available online 3 September 2016

Keywords:
Moral judgment
Moral development
Causal model
Evaluation
Simulation
Empathy
Bayesian
Reinforcement learning
Dual-process
Model-free and model-based learning and control
Trolley problem

ABSTRACT

What is distinctive about a bringing a *learning* perspective to moral psychology? Part of the answer lies in the remarkable transformations that have taken place in learning theory over the past two decades, which have revealed how powerful experience-based learning can be in the acquisition of abstract causal and evaluative representations, including generative *models* capable of attuning perception, cognition, affect, and action to the physical and social environment. When conjoined with developments in neuroscience, these advances in learning theory permit a rethinking of fundamental questions about the acquisition of moral understanding and its role in the guidance of behavior. For example, recent research indicates that spatial learning and navigation involve the formation of non-perspectival as well as egocentric models of the physical environment, and that spatial representations are combined with learned information about risk and reward to guide choice and potentiate further learning. Research on infants provides evidence that they form non-perspectival expected-value representations of agents and actions as well, which help them to navigate the human environment. Such representations can be formed by highly-general mental processes such as causal and empathic simulation, and thus afford a foundation for spontaneous moral learning and action that requires no innate moral faculty and can exhibit substantial autonomy with respect to community norms. If moral learning is indeed integral with the acquisition and updating of casual and evaluative models, this affords a new way of understanding well-known but seemingly puzzling patterns in intuitive moral judgment—including the notorious "trolley problems."

1. Introduction

A query to Google Books requesting an Ngram from 1950 onwards for the phrase *moral development* reveals that this expression underwent a dramatic growth in frequency from 1960 to 1980, before declining gradually to 2008 (the latest year for which results are given). Adding an Ngram for *social learning* shows that this expression followed essentially the same trajectory, climbing yet more dramatically to its 1980 peak before drifting downward in recent years. But request an Ngram for *moral learning* during the same period, and the Ngram Viewer draws a blank. Which leads to the question: If there already are well-established research literatures in moral development and social learning, what might a moral learning perspective add?

The existing literatures in moral development and social learning are far too varied and extensive, and the field of moral learning far too undeveloped, to permit more than a preliminary comparison and contrast. Certainly there is much by way of overlap. A moral learning approach shares with social learning theory the idea that much of our learning takes place by observing others, rather than through direct external reward or punishment. And it shares with moral development theory the idea that our capacity for moral thought emerges over time, drawing upon the development of capacities in other domains.

However, a moral learning approach sees the acquisition of moral understanding as the result of domain-general learning processes, and thus as an integral part of our modeling of the physical and social world. Such modeling generates expectations continuously that guide perception, thought, and action, and permit learning from discrepancies with expectation throughout life. Moral learning therefore can go beyond the acquisition of known moral concepts or internalization of prevailing social norms, and can extend to the formation of novel moral concepts and evaluations, resulting in dramatic personal and social change even within one lifetime.

In this paper, I will examine a series of issues, consideration of which makes it possible to give more substance to a moral learning perspective. Section 2 will present criteria for distinctively *moral* learning. Section 3 will look at causal and evaluative learning as exemplars of the kind of *learning* moral learning might be, and ask why *now* is a particularly apt moment for asking about the power of learning. Section 4 will then apply the model-based

E-mail address: prailton@umich.edu

http://dx.doi.org/10.1016/j.cognition.2016.08.015
0010-0277/

picture of learning developed in Section 3 to the moral case, presenting evidence for the acquisition of non-perspectival evaluative representations that satisfy the criteria presented in Section 2. Section 5 will look into the phenomenon of "intuitive judgment," and use informal student polling data to ask how a "deep" moral learning perspective might account for the puzzling patterns of intuitive moral judgment found in "trolley problems." And Section 6 will conclude by briefly considering how explicit and implicit processes interact in moral learning.

2. Identifying the subject matter

Learning is a success term, and if moral learning is to be an integral part of the knowledge we gain in representing ourselves and the world, then it must be subject to some notion of representational success. Does this require a theory of moral learning to take a stand on which moral theory is correct—seemingly in violation of David Hume's celebrated distinction between *is* and *ought* (1738/1978)?

It is possible, however, for a theory of moral learning to bracket many controversial moral questions by focusing instead on *criteria* of moral evaluation that are shared across a wide range of normative moral theories. Just as we can speak of criteria characteristic of a *scientific point of view* that are implicitly or explicitly followed by those pursuing competing theories, we can speak of criteria characteristic of a *moral point of view*. It is thanks to such shared criteria that there can be a scientific or moral "community," with shared methods and questions, and meaningful disagreement over answers.

Scientific and moral inquiry both aspire to a kind of objectivity that overcomes the limitations of subjective or sectarian perspectives or interests by following methods, and seeking understanding and justification, that are (i) *impartial*, (ii) *general*, (iii) *consistent* (or, more broadly, *coherent*), and (iv) *independent of appeals to special authority*. For example, both require that *like cases be treated alike*, and that the evidence or grounds given in defense of particular positions be in principle *shareable*. Moreover, competing parties to moral and scientific disputes agree that their disputes are not merely speculative. That is, they see themselves as seeking to answer questions about what to believe and how to apply this in practice—whether this is a matter of accepting a scientific hypothesis, following a methodological norm, or deciding upon an ethical course of action. Let us call this the criterion of (v) *thought- and action-guidingness*. One could hardly make sense of the intensity of scientific and moral disputes if one thought that making up one's mind in scientific or moral disputes were a merely notional matter, with no relevance to how we should think and act.

Of course, moral disputes also differ from scientific disputes in a number of respects. For example, morality has a proprietary, non-instrumental concern with questions of (vi) the *harm or benefit of those actually or potentially affected*. Scientists of course are not indifferent to such questions, but they are not treated as an essential part of the evidence or grounds for scientific judgment. Criterion (vi) does not say that impartial concern with harm or benefit is the entire basis of morality, as some utilitarians maintain, but rather that harm and benefit have direct relevance to moral judgment across the full array of major ethical traditions—including deontologies (which typically include duties not to harm and to render assistance to those in need) and virtue theories (which typically connect virtue with human flourishing, and identify beneficence and generosity as central virtues).

To study moral learning or scientific learning, then, it is not necessary to embrace a particular substantive theory or to provide a definition of *morality* or *science*—it is enough to study how individuals or groups develop, and treat as normatively important, forms of inquiry or ways of regulating thought and action that exhibit such features as (i)–(v) or (i)–(vi).

From an evolutionary standpoint, it can appear quite extraordinary that people would impose upon themselves the limitations of forms of inquiry and practice that would meet criteria (i)–(v) or (i)–(vi). Why would natural selection favor the development of mental processes or social dispositions that can be so independent of the reproductive interests of individuals and their kith and kin? Answering this question is one of the key challenges faced by accounts of scientific or moral learning—and we will have something to say about it, below.

3. Causal and evaluative learning

3.1. Philosophical background

Hume framed one of the foundational texts of modern philosophy, *A Treatise of Human Nature* (1738/1978), in terms of the joint problem of understanding *how* we arrive at the attitudes we do on the basis of experience, and whether these attitudes are *warranted*. Hume focused especially on causal and moral beliefs, and perhaps surprisingly, the author of the *is*/*ought* distinction emphasized the fundamental similarities of these two forms of domains of thought. Hume saw that there is a general problem of bridging the gap between sensory impressions, which are particular, concrete, actual, and transient, and what we come to believe on their basis, which is general, abstract, modal, and temporally-extensive. How, he asked, do we come to form causal and moral beliefs which *logically* outstrip all our evidence, and what does this tell us about how or why they might nonetheless be justified?

Hume concluded that *we* must add something to sensation to bridge this gap. Earlier philosophers had often invoked innate ideas, yet these could not really solve the problem he had identified. After all, innateness is not validity, and even if we were endowed with valid general, abstract ideas or rules, we would still have to figure out how to apply these to particular, transient, unruly experiences, or to decisions or actions in concrete contexts. Abstract concepts and rules do not apply themselves, and to appeal to yet other innate concepts or rules to tell us how and when to apply them would be to launch a regress—and "it is impossible for us to carry on our inferences *in infinitum*" (1738/1978; sect. I. iii.4).

Hume's answer is that *imaginative projection* effects the bridge that strictly logical inference cannot. He posited general, default psychological dispositions to respond to certain regularities in sensory experience by mentally extending these patterns to novel experiences and abstract relations of similarity and difference. Forming expectations on the basis of such default projective dispositions might seem to be epistemically reckless, but Hume argued that, by "spreading itself over the world" in this way, belief could make experience into trial-and-error experimentation. Belief for Hume is an active *sentiment* rather than a mere idea, and its projective "initial impulse" will be "broke into pieces" in response to the proportion of success or failure in expectation (1738/1978; sect. I. ii.12). Although Hume is often considered an outright skeptic, on a more plausible interpretation he combined skepticism about the powers of pure reason with realism about the ways sentiments such as belief can ground us in reality and attune our thought and action to the world. Indeed, he claimed, logical reasoning itself can avoid regress only because belief projects spontaneously along the network of the "association of ideas" via relations of similarity and analogy—if such default mental operations cannot be trusted, then reasoning cannot be trusted either (1738/1978; conclusion of Book 1).