Bending toward Justice

P. Kyle Stanford*

Michael Tomasello, ***A Natural History of Human Morality.*** Cambridge, MA: Harvard University Press (2016), x+194 pp., $35.00 (cloth).

Although Michael Tomasello's *A Natural History of Human Morality* is in some ways a companion volume to his 2014 *A Natural History of Human Thinking*, no familiarity with that earlier work is required to thoroughly understand, enjoy, and benefit from this impressive book. In accessible and lively prose, it first provides a wide-ranging (if necessarily selective) tour of recent experimental literature (much of it conducted by Tomasello himself and his collaborators) comparing various respects in which the capacities and motivations for prosocial or cooperative behavior in humans (including very young children) systematically exceed those of our closest phylogenetic relatives (i.e., chimpanzees and bonobos). It then goes on to propose a plausible (if necessarily speculative) evolutionary trajectory concerning the phylogenetic emergence of these differences and the origins of human moral psychology more generally. The book makes a fascinating case for a novel thesis that will be of interest to scholars in a wide variety of academic fields, even if (as I argue below) it fails to deliver on all of its grandest ambitions.

Tomasello divides the challenge of explaining human moral psychology into that of explaining how we came to have both a 'morality of sympathy' and a 'morality of fairness'. He argues that the former emerged fairly straightforwardly from a process of 'self-domestication' in which the sort of genuinely other-directed sympathetic concern exhibited by chimpanzees and bonobos (and presumably the most recent common ancestor we share with them) for kin and 'friends' was extended to encompass unrelated group members and even strangers. But he regards the emergence of a morality of fairness as a far more difficult explanatory challenge, requiring a large number of cog-

Received December 2016.

*To contact the author, please write to: 5100 Social Science Plaza, University of California, Irvine; e-mail: stanford@uci.edu.

Philosophy of Science, 84 (April 2017) pp. 369–376. 0031-8248/2017/8402-0010$10.00

nitive, emotional, and regulative capacities not shared by even our closest phylogenetic relatives, and most of the book goes on to sketch a complex, two-stage process by which such a distinctively human morality of fairness might have emerged over the course of our evolutionary history.

Perhaps unsurprisingly, Tomasello's version of this tale prominently features a number of capacities (like shared intentionality and joint commitment) whose importance for understanding human ultrasociality and cooperation he has long advocated, but he does not simply revisit earlier arguments for these claims. Tomasello's broadest thesis is that evolutionary thinking about cooperation and prosociality has unfairly and unfortunately privileged processes and contexts of *reciprocity*, in which the sacrifices agents make to their own welfare to benefit others must be consistently repaid in order for the behavior to remain evolutionarily stable (often demanding in turn mechanisms like careful accounting, reputation management, or punishment), over processes and contexts of *mutualism* or *interdependence*. In such mutualistic contexts, an agent can recover a significant fraction of her investment in the welfare of collaborative partners, potential mates, fellow group members, and others with whom she is interdependent, because her contributions to the fitnesses of those with whom she is truly interdependent also make an immediate contribution to the fitness of the agent herself: I might, for instance, lend my collaborative partner a tool so that she may perform her role in our joint activity more effectively (increasing our total yield) or protect her in a fight so that she remains available for our next collaborative enterprise. Creatures who stand in such interdependent relationships have a vested (evolutionary) interest in one another's survival and flourishing, ensuring that they are far more likely to find themselves in circumstances in which it pays to be altruistic, cooperative, or helpful, whether or not this behavior is eventually repaid in kind. (This makes the formal structure of such mutualistic interactions much more like kin selection than reciprocal altruism; in the language of evolutionary game theory, it pushes games of interaction that would otherwise be Prisoners' Dilemmas toward becoming Stag Hunts, and Stag Hunts toward Prisoners' Delights.) And Tomasello's "interdependence hypothesis" argues that distinctively human moral psychology was engendered by the emergence of new cognitive, social-motivational, and self-regulatory psychological capacities for generating, managing, and sustaining such interdependent relationships. Thus, insofar as the emergence and development of distinctively human psychological capacities like shared intentionality and joint commitment were central in establishing new and more complex forms of interdependent relationships and collaborative interactions among humans, these same capacities played a crucial role in both generating and shaping the distinctive features of our moral psychology.

More specifically, Tomasello argues that the first step in our complex, two-part phylogenetic journey to a morality of fairness occurred roughly 2 million

years ago, when ecological changes forced humans to become (unlike nonhuman primates) *obligate* cooperative foragers. This transition was mediated in part by the emergence of a new and unique set of proximate psychological mechanisms fostering such new forms of interdependence, including most importantly the capacity to form plural agents ('we') with others and to form agent-neutral conceptions of collaborative activities in which either partner could play either role and that prescribed ideal standards for the performance in the role no matter who occupied it. Tomasello argues that these socially shared normative standards for the performance of roles generated in turn a recognition of "self-other equivalence," leading collaborating agents to exclude free riders but to regard collaborative partners as equally deserving and entitled to mutual respect, including fair treatment. And he argues that the cognitive capacities required to construct such plural agents (along with those needed to choose among and control partners, to establish and maintain cooperative identities, and to explicitly engage in such joint commitments with others on the basis of those identities) collectively fostered a form of 'second-personal morality', not simply based on strategic reciprocity or evading punishment or protecting one's reputation as a cooperative partner but instead reflecting a sincere effort to live up to such shared role ideals in collaborative activities.

In the second step of this transition, modern human groups became larger, eventually splitting into smaller bands that were nonetheless unified by shared culture at the tribal level. Such tribal groups competed with one another and operated as much larger collectives with which all group members identified and in which special forms of loyalty and sympathy were owed to fellow group members but not to outsiders. This development required further novel psychological capacities, including processes of collective (rather than merely joint) intentionality; processes of cultural agency with respect to conventions, norms, and institutions; and processes of self-regulation or self-governance based on a commitment to membership in a community. Collectively, these processes made possible the introduction and persistence of such cultural conventions, norms, and institutions in a common ground shared by all members of that group—conventional cultural practices thus came to include role ideals that members of the group saw as regulating how anyone who would be part of the group must play those roles successfully, that is, the right and wrong way to do things. In this way, the second-personal morality of early humans was 'scaled up' to include collective commitments authored by our group and for our group, with fully objective normative standards governing all group members as well as moral identities within the group that require justification and defense. This second step took early humans from the joint, second-personal morality of specific responsibilities to particular collaborative partners to a more impersonal collective morality of cultural norms and institutions specifying the obligations of all group members to one another.

Tomasello sometimes overstates how much shifting to an interdependence perspective will single-handedly achieve. For example, he seems to suggest (18–20) that this shift helps explain the need for mechanisms of partner choice and control, whereas in fact mutualistic interdependence instead reduces the need for such mechanisms by allowing the altruist or cooperator to share in the benefit she confers on the recipient and thus to benefit from helping even partners who need not be carefully chosen or controlled to ensure reciprocation.[1] And inevitably, of course, there are many specific points at which it would be perfectly appropriate to worry about the empirical credentials and evidence in favor of Tomasello's unavoidably speculative account, as he himself candidly acknowledges. But I will now suggest that a far larger problem looms for Tomasello, in that even if we simply accept his account it does not actually explain as much of human moral psychology as he contends.

In particular, Tomasello is admirably clear and forthright in recognizing that a satisfying account of human moral psychology will have to explain why we attribute a salient sort of objectivity to moral directives, judgments, and norms: we do not simply enjoy or prefer to act in ways that satisfy the demands of morality, we see ourselves as obligated to do so no matter what our subjective preferences or desires may be, and we regard such demands as imposing obligations not only on ourselves but on any and all agents whatsoever, regardless of their preferences or motivations. As Tomasello notes, the morality of fairness produces "judgments [that] typically carry with them some sense of responsibility or obligation: it is not just that I want to be fair to all concerned, but that one *ought* to be fair to all concerned" (2). Tomasello devotes considerable effort to seeking to account for this aspect of our moral psychology within the evolutionary framework he adopts, and he argues that the phylogenetic trajectory he proposes can indeed provide a convincing explanation for this salient and distinctive feature of human moral cognition. But I suggest that the proposed explanation fails even if we assume that Tomasello's evolutionary genealogy is entirely correct in its finest details as well as its broadest strokes.

Tomasello argues that the sort of objectivity we attribute to the moral domain is itself a consequence of several transitions in the evolution of our moral psychology. The first and most important of these transitions arose as part of the establishment of 'second-personal morality', in which collaborating agents achieved agent-neutral representations of their own collaborative activities, including norms governing ideal performance of roles no matter who occupied them. Such standards were impartial and normative,

1. I am grateful to Aydin Mohseni for making this point to me explicitly and for further useful suggestions incorporated into this review. I am also grateful to Michael Tomasello for his helpful comments.

specifying how either partner (or anyone at all) must perform the role in order to achieve success in the joint activity, but they were not yet moral in character. Nonetheless, Tomasello argues, recognizing that the relevant standard applied impartially to whomever occupied the role provoked the further recognition that "self and other were of equivalent status and importance in the collaborative enterprise" (4), ensuring in turn that "partners came to consider one another with mutual respect, as equally deserving" (4) even as they "evolv[ed] the tendency to deter . . . free riders by denying them a share of the spoils" (60). As he later puts it, collaborating partners "came to understand that particular collaborative activities had role ideals—socially normative standards—that applied to either of them indifferently," and the recognition of this "self-other equivalence" in turn generated "a mutual respect between partners, and a sense of the mutual deservingness of partners" (41), especially when agents were free to choose among partners in pursuing such collaborative activities and to express dissatisfaction with a partner's suboptimal performance. The emergence of such a "second-personal morality," Tomasello argues, was itself "the decisive moral step that bequeathed to modern human morality all of its most essential and distinctive elements" (78).

The central problem with this proposal lies in its slide between two quite distinct forms of "self-other equivalence." We can freely grant that early human collaborators came to recognize a kind of instrumental or functional equivalence between themselves and potential partners who might fill the same roles and be subject to the same normative standards of ideal performance. But the further inference that such potential partners are therefore entitled to 'mutual respect' or are 'mutually deserving' itself depends on a distinct (and distinctively moral) sense of 'self-other equivalence'. That is, we can only move from the recognition that multiple partners could equally well occupy a given role and be subject to the associated standards of instrumental performance to the conclusion that such partners are therefore entitled to mutual respect or are equally deserving (unlike free riders) by means of substantive moral commitments concerning (at a minimum) what entitles an agent to respect or what makes her equally deserving of the spoils of a common enterprise. Thus, the recognition of instrumental or functional self-other equivalence will not produce or generate such a further recognition of moral self-other equivalence unless we help ourselves to distinctively moral commitments snuck in through the back door.

Of course, Tomasello might instead lean more heavily here on the importance of partner choice and control, arguing that potential collaborators who saw others as equally deserving and entitled to mutual respect would be selectively favored as more desirable collaborative partners. But this does not help explain the emergence of moral objectivity, because it requires only that we have desires or preferences for interacting with appealing partners (who would themselves need only to desire or prefer to treat others as de-

serving and entitled to respect), rather than requiring that the demand for mutual respect or equal treatment be experienced by either potential partners or those who select among them as objective or externally imposed on us in any way. Nor will it help to appeal to the second step in Tomasello's evolutionary trajectory, in which such normative attitudes are 'scaled up' into the full-blown norms and ideals of an entire group, for one of the most salient and puzzling features of human moral psychology is that across human cultures children between 2.5 and 3 years of age reliably begin to distinguish genuinely moral norms from those that are merely conventional (such as norms of etiquette and fashion), treating the former as unconditional, universal demands that all agents must satisfy and the latter as merely local, contingent, authority-dependent rules that we ourselves have created to guide our own conduct (see Turiel 1983; Turiel, Killen, and Helwig 1987; Smetana 2006). Indeed, Tomasello's repeated description of the "objective" norms and institutions of a group as "commitments . . . made by 'us' for 'us'" (e.g., 86) neatly captures this conception of merely conventional norms, but one of the most notable features of distinctively moral norms is that we do not experience or regard them in this same way. Tomasello sometimes recognizes the need to explain why some group norms become moralized and others do not, but he seeks to finesse the problem (e.g., 86, 99–100) by connecting the process of objectification back to violations of the sense of self-other equivalence to which he appeals in the first step in his evolutionary trajectory, and we noted above that this appeal already presupposes rather than explains any distinctively moral form of commitment.

This inability to explain why we attribute a distinctive form of objectivity to moral norms and judgments seems like an especially important shortcoming of Tomasello's account even by his own lights, but it should perhaps also be unsurprising since this feature of our moral psychology has proved extremely difficult to explain for evolutionary approaches to understanding human moral psychology quite generally (see Joyce 2006, 92–93 and passim). I will conclude, however, by describing what seems to me to be a widespread and systematically misguided way of thinking about the evolution of human moral psychology that regularly recurs throughout the book. I am not at all sure that Tomasello would explicitly endorse the mistake I will describe, but perhaps the fact that an evolutionary thinker of his evident sophistication seems to have repeatedly fallen into it should serve as a cautionary tale for the rest of us.

Writers on the evolution of human moral psychology have sometimes expressed wonder and admiration at our good fortune in evolving the various psychological capacities and motivations needed for us to be genuinely moral beings or to actually behave morally at least some of the time. Tomasello seems to invite or encourage this same sort of admiration or wonder at a number of points, such as when he writes that we "should simply marvel

at the fact that behaving morally is somehow right for the human species" (7), that recognizing self-other equivalence "had nothing to do with strategy, only with reality" (81), that humans see cooperative interaction as "the right thing to do because . . . well . . . it just is the right thing to do" (126), and that agents who behave morally "are acting with a kind of cooperative rationality based on an accurate recognition of social reality" that need only be "at least viable on the evolutionary level" (153). But this is surely not the right way to regard our situation from an evolutionary point of view. We are not lucky that evolution has equipped us to act in ways that are in fact moral; instead, we use moral language to describe the norms of behavior that we have actually evolved (biologically or culturally) to endorse and favor; that is, we use moralized language to describe what has evolved for humans in the way of motivations and norms governing altruism, cooperation, and other forms of social interaction. This point has been made by evolutionary game theorists (e.g., Skyrms 1996; Binmore 2005; Alexander 2010) regarding specific norms like fair division or reciprocity that emerge as stable equilibria in particular games of interaction: we are not lucky that the genuinely moral or fair norms of behavior happen to be those that emerge as evolutionarily stable; instead, we regard them as fair or appropriate or morally admirable because they are the norms that emerged. Had our evolutionary trajectory been different, quite different motivations, norms, and mechanisms governing cooperation, altruism, and other aspects of our social interaction might have evolved as well, but we would marvel no less in that case at our good fortune in having evolved the constituent psychological elements required for such 'truly' moral behavior. This makes such marveling seem as misplaced as admiration for the accuracy of the famous Chinese archer of legend, who shot his arrows into a fence and then carefully painted a target around each one with the arrow located at the exact center of the bull's-eye. Nonetheless, Tomasello concludes the book as follows: "it is a miracle that we are moral, and it did not have to be this way. It just so happens that, on the whole, those of us who made mostly moral decisions most of the time had more babies. . . . We should simply marvel and celebrate the fact that . . . morality appears to be somehow good for our species, our cultures, and ourselves—at least so far."

I have suggested that from an evolutionary point of view we should resist any conception of this convergence as a miraculous coincidence or marvelous good fortune. And earlier I argued that Tomasello's account does not actually explain as much about human moral psychology as he supposes, more specifically that it fails to explain what might well be the most salient and perplexing feature of that psychology: the distinctive sort of objectivity we attribute to moral (as opposed to merely conventional) norms and judgments. Fortunately, neither these challenges nor any other shortcomings of the book undermine the interest or significance of its many quite genuine

explanatory achievements and illuminating discussions, and I am happy to report that it is a pleasure to teach to graduate students. *A Natural History of Human Morality* represents the state of the art in evolutionary theorizing on this subject, and it fully deserves the close attention I expect it to receive from scholars in a wide variety of academic fields who seek an evolutionary understanding of human moral psychology itself.

REFERENCES

Alexander, J. 2010. *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.
Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.
Joyce, R. 2006. *The Evolution of Morality*. Cambridge, MA: MIT Press.
Skyrms, B. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Smetana, J. 2006. "Social-Cognitive Domain Theory: Consistencies and Variations in Children's Moral and Social Judgments." In *Handbook of Moral Development*, ed. M. Killen and J. Smetana. Mahwah, NJ: Erlbaum.
Turiel, E. 1983. *The Development of Social Knowledge*. Cambridge: Cambridge University Press.
Turiel, E., M. Killen, and C. Helwig. 1987. "Morality: Its Structure, Functions, and Vagaries." In *The Emergence of Morality in Young Children*, ed. J. Kagan and S. Lamb. Chicago: University of Chicago Press.