CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Moralization of preferences and conventions and the dynamics of tribal formation
Authors	Ross, Don
Publication date	2018
Original Citation	Ross, D. (2018) 'Moralization of preferences and conventions and the dynamics of tribal formation', Behavioral and Brain Sciences, 41, e111. doi: 10.1017/S0140525X1800016X
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1017/S0140525X1800016X
Rights	© Cambridge University Press 2018. This article has been published in a revised form in Behavioral and Brain Sciences, http://dx.doi.org/10.1017/S0140525X1800016X This version is free to view and download for private research and study only. Not for re-distribution, re-sale or use in derivative works.
Download date	2026-11-16 11:02:12
Item downloaded from	https://hdl.handle.net/10468/8581



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Commentary on P. Kyle Stanford

Word counts: Abstract: 57; Main text: 999; References: 7; Entire text: 1163

Title: Moralization of preferences and conventions and the dynamics of tribal formation

Don Ross

University College Cork

School of Sociology and Philosophy
4 Elderwood
College Road
Cork
Ireland
T12 AW89

+353 85 7508200

University of Cape Town

School of Economics
University of Cape Town
Private bag
Rondebosch 7701
South Africa

+27 83 701-3757

Georgia State University

Center for Economic Analysis of Risk
J. Mack Robinson College of Business
35 Broad Street NW
Atlanta GA
USA 30303

+1 205 396-9071

don.ross931@gmail.com

http://uct.academia.edu/DonRoss

Abstract: Stanford casts original light on the question of why humans moralize some preferences. However, his account leaves some ambiguity around the relationship between the evolutionary function of moralization and the dynamics of tribal formation. Does the model govern these dynamics, or only explain why there

are moralizing dispositions that more conventional modeling of the dynamics can exploit?

Stanford's problem can be succinctly expressed: Why do humans moralize some preferences? This means: most humans treat some of their preferences are expressing objectively grounded, universally binding norms. Further content consistent with Stanford's evidence and reflections can be suggested. A moralized preference is one that the moralizer cannot propose to trade off against other preferences without expecting to experience shame, and without expecting others to legitimately regard her as shamed, except in cases where two or more moralized preferences unavoidably and clearly enjoin opposed actions.

Stanford's solution to the problem is innovative and broadly convincing. It allows him to explain various psychological and social features of moralization, and yields a neat explanation of why moral judgments have been so philosophically perplexing. How well supported by available evidence is this solution? It depends on evidence about moralization *and* on evidence about the pressures on cultural adaptation that Stanford invokes to explain moralization. I will focus on the latter.

The crucial driver of Stanford's model is what he characterizes as humans' unique plasticity. He is not as explicit about this as he might be. Plasticity typically refers most directly to learning capacity. But what mainly does the work in Stanford's model is a *consequence* of learning capacity, namely, the observed fact that people have colonized a remarkable range of niches. This has in turn given rise to, and required cultural adaptation to, a diversity of lifeways. On Stanford's account, this continuous dynamism disrupts stabilization of coordination and control of free riding by mere conventions entrenched in motivational drives. Moralization has allowed people to repeatedly segregate themselves into tribes which are, according to Stanford, endogenously equilibrated but still potentially unstable because of their continuing dispositions to construct or find new niches. The potential instability preserves the functional value of moralization.

Stanford says little about the dynamics of tribal formation. One might naturally think of our ancestors radiating from warm grasslands and scrubland into climates with cold winters or dense forests. But human tribes manifestly bifurcate *within* shared physical environments. On the face of it this seems to be just what Stanford's hypothesis predicts: a subset of a founder population in a niche moralizes some of its new conventions in order to achieve and maintain correlated equilibrium and successfully exclude those most disposed to free riding. Then, presumably – Stanford is not explicit on this point – the excluded villains interact with one another for lack of an alternative, and form and then moralize different conventions.

Does the tribe of cast-offs moralize for the same reason as the original moralizers? If so, they should be expected to spin off yet another tribe, and we predict a recursive pattern that perhaps terminates in the creation of marginal 'sick societies' (Edgerton 1992), where free riding is impossible because benefits from cooperation have shrunken to their biologically minimal core. One might speculatively imagine a

sick society that, forced into geographical isolation, endures for long enough that natural section catches up to cultural selection and we end up with normative psychology resembling that of chimpanzees.

An alternative account, also consistent with Stanford's dynamics but making it less general as a model, might go as follows. We begin with a first stage of social evolution in which the 'Stanford process' gives rise to the natural disposition to moralize suggested by the developmental evidence that Stanford cites. Once this biological adaptation has occurred, we enter stage two, and tribal formation with rival moral codes is supported by more familiar strategic dynamics of cultural group selection: to compete successfully, groups need effective solidarity; therefore they need costly entry barriers and membership fees that reliably signal commitment; moralization that limits individual freedom of action serves this function, along with the function of making members inadmissible by rival groups with different moral codes, and aligning their self-interest with militant patriotism. Tribal formation itself might then be mainly governed by exogenously varying resource constraints that constitute parameters on stable tribe sizes, with new tribes forming whenever, on the margin, some people are better off forming a new tribe than receiving a diminishing share of the pie generated by the existing tribe. Generation of new niches (i.e., 'plasticity') on this second interpretation might mainly, in stage two, help to make new moral codes relatively economically adaptive, thereby counterbalancing the initial disadvantages typical of a smaller start-up.

On the first interpretation above, Stanford's model governs the dynamics of tribal formation. On the second interpretation, it explains why these dynamics find moralizing dispositions to exploit, but the dynamics themselves are modeled by a fusion of anthropological group selection and the economics of dynamic industrial organization.

One would be forced to disambiguate these interpretations if the model were formalized. This leads to a methodological observation about evolutionary psychology. Economists prefer formal models partly because these generate relatively precise discriminating empirical predictions that might not be evident to the theorist in advance of the formal specification. Such predictions are important not because prediction is the primary goal of science (it is not), but because specification of predictions is a crucial tool for identifying a model's empirical scope.

We see scope ambiguity in Stanford's informal model if we compare humans not only with chimpanzees, but with more genetically distant animals that are more similar to humans along some social dimensions. Whereas chimpanzees form rival, warring groups but *not* morally differentiated tribes, orcas resemble humans in forming geographically overlapping communities with strikingly different core behaviors, communication codes, and social organization. Individuals drawn from different groups housed together in captivity don't seem to get along well. Orcas inhabit all oceans, and we have no independent metric for determining how profoundly or shallowly their inhabited functional niches vary *with respect to what matters to them*. Is Stanford's model intended to be sufficiently general to be used in

deciding whether we should predict morality in (e.g.) killer whales? Or is it intended mainly to explain a dimension of divergence in the ape / hominid evolutionary line? Formalization of the model might usefully force the distinction.

Reference

Edgerton, R. (1992). *Sick Societies*. Free Press.