PAPERS on Economics & Evolution

0616

Moral fictionalism, preference moralization and anti-conservatism: why metaethical error theory doesn't imply policy quietism

by

Don Ross

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

Moral fictionalism, preference moralization and anti-conservatism: why metaethical error theory doesn't imply policy quietism

Don Ross

Department of Philosophy and Department of Finance, Economics and Quantitative Methods
University of Alabama at Birmingham

School of Economics
University of Cape Town

dross@commerce.uct.ac.za

abstract

The evolutionary explanation of human dispositions to prosocial behaviour and to moralization of such behaviour undermines the moral realist's belief in objective moral facts that hold independently of people's contingent desires. At the same time, advocacy of preferences for significant departures from hallowed policies (that is, for 'loud policies') is generally sure to be ineffective unless it is moralized. It may seem that this requires the economist who would advocate loud policies, but is also committed to a naturalistic account of human social and cognitive behaviour, to engage in wilful manipulation, morally hectoring people even when she knows that her doing so ought rationally to carry no persuasive force. Furthermore, it might be wondered on what basis just for herself an error theorist about morality advocates loud policies. I argue that understanding the role of moralized preferences in the maintenance of the self, and in turn understanding the economic rationale of such self-maintenance, allows us to see how and why preferences can be moralized by a believer in error theory without this implying hypocrisy or manipulation of others.

1. Introduction

As an economist who works on African trade and development policy, there are various policies I favour with respect to which I would feel significantly ashamed were I to succumb to bribery, or to institutional pressure short of physical threat, to publicly support their opposites. Here are a few of these policies:

(1) Rich countries should not impose trade barriers, including subsidies for their own producers, against imports from poor countries.
(2) Notwithstanding the importance of the above problem, governments of poor countries should be denied aid or debt relief unless they acknowledge that their principal economic problems are not caused by outsiders, but by their own institutions.
(3) Much freer movement of people across national borders for purposes of labour should be allowed everywhere.
(4) Destitution is an unduly severe penalty for laziness, lack of talent or intelligence, or bad luck. More productive people should provide the lazy, the unskilled and the unlucky with as large s basic income grant as is compatible with maximizing net social productivity (which no present society comes very close to doing, partly due to confused social morality and partly due to incentive-distorting tax systems).

None of these are 'quiet' policy preferences. By this I mean: none of them are policies that are likely to be implemented as a result of purely incremental reform within existing institutional frameworks. When I publicly express the economic arguments on which these preferences are based, which I do regularly, I therefore usually accompany them with a certain amount of moralizing. Such moralizing might sometimes or often just represent self-indulgence, but I think (and hope) that it isn't always or necessarily *just* that. Inductively, one can see that a prerequisite for the kind of institutional change that makes loud policies possible is getting more people to share the advocate's disgust with the status quo; and use of moralized discourse is the basic means of trying to do this.

As well as being an economist with these and other loud policy opinions that I sometimes press in a moralistic way, I am an evolutionary naturalist about the sources of human belief and behavior, and on the basis of metaphysical and epistemological arguments I believe that there are no moral facts. That is, I do not believe there are true propositions about what is morally right and what is not.

The subject of the present essay is whether it is necessarily embarrassing from the point of view of rationality to espouse moralized opinions while denying that there are moral facts.

There are various non-economic general policies I also support. For example, I think that the state ought not to be allowed to kill born people except in circumstances of war or when they freely ask to be killed. However, I will concentrate here on economic policies because I judge at least prima facie that I have clearer and surer justifications for them than I do for those on which it is harder to bring rigorous empirical measurement and clear theory to bear. I have difficulty advancing my opposition to the death penalty with full confidence because I believe that, sincere

though it is, it amounts in the end of the day to emotional reactions mixed with borrowed reasoning. Of course, I suspect the same thing about the other side on this issue. Part of the problem is that I don't know, and I don't think anyone knows, to what extent if any the threat of the death penalty deters murder. But even if I thought I knew this fact, it might not settle the matter, since an opinion on how far we ought to go in punishment to deter murder seems like the sort of position that is *essentially* a moral judgment and, for privileged people unlikely to be either murderers or murder victims, a moral judgment mainly about the motives and dispositions of people I don't know. Where judgments like these are concerned, I'm convinced that everyone is on shaky ground even *before* the persuasiveness of error theory is put onto the metaethical table.

By contrast, I know explicitly why I favour policies (1)-(4) above: if politics allowed their implementation, then the resulting gains in aggregate wealth would be sufficient to compensate those made worse off by them, with the possible exceptions of people whose present positions derive entirely from capture of rents. Like most economists, I regret *waste* as something close to a primitive response.

To this statement it might be objected that regardless of how economists might often *feel* about the matter, abhorrence of waste can't really be a *primitive* normative stance unless it is irrational. The waste in question, after all, is waste of welfare, and it would make little sense to resent it unless one cared in the first place about people's well being. This much must surely be granted in some sense. But the idea of 'caring about people's well being' is ambiguous in this context. On the one hand, one might care about people's well being in a *benevolent* sense, thinking that one ought to expend a significant proportion of one's human and other capital trying to improve people's welfare (in general) for their sake. Alternatively, one might only care about people's welfare in the weaker sense of having *non-vicious* preferences with respect to it: we should not pursue policies that are Kaldor-inefficient, that make people worse off than they could be given an alternative that would allow us to redistribute gains in such a way that some enjoyed improved welfare and losers could be fully compensated. This really does look like it is close to being a primitive dislike of waste *per se*. However, it can be added that when our subject is the social policy order, a person who has parochial preferences about some particular people's welfare (including, typically, her own), should care about the welfare of a wider set of people *instrumentally*, because if general welfare is allowed to be degraded too far below a certain (multi-dimensional) set of thresholds this tends to upset the social stability upon which the well being of the parochially favoured people partly depends.

In the present essay I will summarize and discuss a metaethical argument due to Richard Joyce (2001, 2006) to the effect that, in light of Darwinian naturalism, there is no argument that should, as a matter of rationality, turn a non-benevolent person into a benevolent one. This is important because I take it as obvious that most people are not naturally benevolent in their policy preferences in advance of such argument. This serves to isolate the importance of non-vicious preference in normative debates about policy. Does Joycean skepticism also suggest that there is no argument that ought to turn a rational person with vicious preferences into a person with non-vicious ones? Ken Binmore (1994, 1998, 2005) has argued that this doesn't matter, because there are economic reasons why people with merely parochial concern for some people's welfare will be driven, in bargaining over policy, to the outcomes that would

be realized by people with benevolent preferences. I will contend that while this is an importantly reassuring insight for those of us who, being economists, are conditioned to express non-vicious preferences because we are trained to dislike waste, it doesn't get us quite as far as some of us wish to get. In particular, it is not likely to give us a standing justification for advocacy of loud policies. (Binmore himself advocates policies that are not quiet, but only because, he says, this is in his nature;[1] this gives no one with a different nature encouragement not to be quiet.) At this point, however, I will argue that some well-supported empirical moral psychology can provide those of us who are determined to try to be rational, consistent and Darwinian and avoid indulging metaphysical fantasies about the foundations of morality a basis for going on lecturing people about the rightness of loud policies, such as (1)-(4), without having to blush.

Let me be clear, however, that my conclusion is only the relatively weak one that, *given* the role played by preference moralization in human political psychology, people who are error theorists about morality are not necessarily hypocritical or irrational when they express their preferences about how people should behave in moral terms. I will *not* argue for the stronger claim that the universal human practice of preference moralization is a net boon all things considered. Preference moralization has been a necessary contributor to the worst behaviour in which humans indulge – mass murder, torture, and enslavement of others. I have no idea, and neither does anyone else, whether a hypothetical species just like human beings except for their lack of dispositions to moralize preferences would be better or worse off on average than actual humans.

2. There are no moral facts

Joyce (2001) argues for an error theory of moral discourse, which he uses to defend a positive thesis he calls 'moral fictionalism.' There is room for confusion here, because this same label is used by Kalderon (2005) to defend a specific version of what philosophers call 'non-cognitivism'. This is the view that moral claims do not aim to state propositions, but are simply expressions of utterers' attitudes and/or emotional states and dispositions. Joyce's error theory, by contrast, is grounded in cognitivism. According to his view, when people make moral claims they typically intend to say something they think is true, and when they interpret moral claims they understand claimants to be similarly intending. Nevertheless, Joyce argues, no one who tries to state a moral truth ever succeeds in doing so, because all claims about what is moral are false, for the simple reason that the world lacks the sorts of features which, in general, it would have to have for any positive moral claims to be true. I will not here aim to critically defend Joyce's argument for error theory, but will simply summarize it. I have no doubt that there are philosophers out there who will identify ways in which the argument can and should be improved. But if we observed a rule saying we could never discuss the policy implications of a philosophical conclusion until we had obtained a perfect argument for it, then philosophy truly would be practically irrelevant.

[1] Binmore claims that he writes on social justice because his mind has been colonized by reformist memes.

Joyce argues that the practice of moral discourse presupposes as an essential condition that moral reasons are categorical, universally normatively applicable and binding. That is, a body of morality, as a condition on its *being* morality, does not excuse some (cognitively competent) people from being answerable to it just because they have deviant desires. It can sometimes excuse them because of their environmental circumstances, but these grounds for excuse can in principle be endogenized within the rules, leaving them at least logically categorical.[2] In this respect, people interpret moral principles differently from the way in which they interpret mere conventional norms, such as the rule that one should not start eating until everyone in a dinner group has joined the table. Following Nichols (2004), Joyce (2006) cites evidence that the observation of this distinction is not itself a mere social convention. Children as young as three judge that conventions may vary with place and circumstance, and lose their force if whatever authority maintains them decides they no longer matter. (Young children probably do not imagine or understand self-enforcing conventions.) By contrast, children deem that moral principles (e.g., 'unprovoked hitting of people is wrong') apply everywhere and always and do not depend on any authority's practice or opinion. Solid majorities of children from religiously conservative households judge that hitting would still be wrong even if God became indifferent to it; whereas the same children think that if God relaxed rules against working on Sundays, then working on Sundays would cease to be a bad thing to do (Nucci 1986). In general, Joyce's claim about the commitments intended by users of moral propositions is based not on conceptual analysis but on empirical psychology.

Despite these generally held convictions about what morality is, Joyce (2001) argues, no moral reasons can motivate a person's behavior (including 'private' behavior consisting in acknowledging reasons to oneself) independently of that person's actual, contingent, desires. He does not mean by this claim that every person must have desires that connect explicitly with (putative) moral reasons. Rather, the (carefully defended) contention is that responsiveness to practical reasoning *in general*, but not responsiveness to all actual, particular reasons, is a necessary condition for agency independently of any institutional (including cultural) criteria. Then a defender of the claim that there are true propositions identifying the existence of categorical, universally normatively applicable and binding requirements on a rational person's behaviour would have to defend the claim that there are universal desires such that these desires in the presence of sound practical reasoning in general make moral reasons categorical, universally normatively applicable and binding.[3] Joyce then argues that, in fact, there are no such desires. Thus there is no argument that could convince every amoral person that her beliefs about what she ought and ought not (morally) to do are mistaken. There are no general facts about what is desirable on which people who are and aren't motivated by moral claims disagree, such that if that disagreement were resolved practical reasoning could get them the rest of the way.

[2] As emphasized in Dennett (1988) and reiterated in Dennett (1995), such logically categorical rules might fail to be *practically* categorical because the excusing circumstances can't all be anticipated, so that the escape clauses are ceteris paribus. This is an important point, but not directly relevant to the issues I will consider here.

[3] Joyce (private correspondence) points out that he doesn't think even this would work, since it would allow for the possibility of moral obligations that are merely obligations to oneself; and Joyce doubts that people use the concept of morality in a way compatible with this implication.

Joyce does not deny that people are often well-disposed toward one another. Evolutionary theory explains, through its identification of conditions for fitness-promoting kin selection, mutualism, reciprocity and cultural group selection, why behavioral tendencies compatible with typical moral codes will often be favored in social animals (Joyce 2006, Chapter 1). In humans, cultural evolution (including cultural group selection) has built dispositions even more superficially impressive to the moral realist than what natural selection bestows on intelligent social animals generally. Pareto-superior outcomes can, in evolutionary games that plausibly were played by our ancestors, be promoted by selection of tendencies to believe that behavior identified as 'moral' is commanded by something like Kant's categorical imperatives or by gods. Joyce's (2001) leading example for making this point is a standard Prisoner's Dilemma (PD). Suppose that two hunters jointly maximize their survival chances in the face of tiger attacks by standing together and fighting, but that each hunter's dominant strategy is to flee. They might avoid getting into this PD if they emotionally care about one another. However, if they are *somewhat* selfish – as natural selection will almost certainly have built them to be unless they are genetically haplodiploid, like bees and ants – or if they have competing altruistic concerns, such as their potentially orphaned children back at camp, then the PD will resurface whenever the danger is great *enough*. The probability of the PD will be truly minimized to the extent that they have *moralized preferences*[4]: that is, if they believe that 'the universe' grounds a categorical imperative to the effect that one must never abandon one's partner in the face of danger. Moralized preferences achieve their effectiveness by being coordinated targets of belief in communities, and this means they can do their work even for some community members who fail to fully imbibe them. Thus, even if the hunter Grog, through astute philosophical reflection, has become a sophisticated skeptic about categorical imperatives, these may still motivate him away from the PD if he knows that his reputation in his more credulous community will be damaged if his compatriots infer, when he rushes back to camp alone, that he has violated the moral law. If moralized preferences can thereby be supported in the equilibria of (cultural) evolutionary games, as indeed they can, then we can explain why history seems to have produced a species composed of a majority of explicit moral fundamentalists and a minority of implicit moral Kantians, while we moral skeptics are vanishingly thin on the ground.

Majority convictions notwithstanding, however, the skeptics have science on their side. The universe isn't the sort of thing that can command anyone to do anything relevant to morality. Neither biological nor cultural evolution are sensitive to intrinsic values, because there aren't any. Selection *does* respond to instrumental values, in the sense that it consistently builds certain desires – e.g., for sex in sexually reproducing animals – that are clearly as ubiquitous as they are because sex is (statistically) good for fitness. But instrumental values are precisely *not* what interests the moral realist. Note also that even if natural selection built some desire or other into absolutely everyone, this would have to be because this desire was overwhelmingly reliably fitness-enhancing, at least among some subset of our ancestors. But something's being overwhelmingly reliably fitness-enhancing implies nothing about its being morally right; as Dennett has repeated in several places, there is no justification, within the language game of morality, for a moral principle to the effect that people should side

[4] This is my phrase, deriving from work in evolutionary game-theoretic modeling independent of Joyce's philosophical contribution. But the phrase is exactly consonant with what Joyce does explicitly say (2001, p. 177).

with their genes. Similarly, if cultural group selection built some universal desire (which it probably never does), there is no justified general principle to the effect that people should always side with their communities. The kind of universal desire that would be needed to underwrite the truth of moral claims would have to be one that was (a) impossible to lack while still being a person, or perhaps even an agent at all, and (b) was similar enough to what we take to be fundamental to morality that we could regard it as the natural property on which morality supervenes. We all know, however, that although most people may be loosely Kantian in their moral *theory*, most in practice break what they take to be the moral law from time to time – usually because self-interest successfully urges another course – and some people break it whenever they believe doing so will promote their selfish interests. Thus even widespread desires relevant to morality are not universal. Desires that might be defended as universal – e.g., the desire for adequate oxygen – do not meet criterion (b). Lacking desires relevant to morality is sufficient to make amorality rational for a person. And if amorality can be rational then there cannot be moral facts.

The opponent doctrine to this skepticism that Joyce takes most seriously is moral naturalism, as defended by (inter alia) Richards (1986), Campbell (1996), Dennett (1995), and Casebeer (2003). This is the view that enough moral properties can be identified with or supervene on natural properties that moral realism can be defended without appeal to intuitionism, theology, or other metaphysical magic. Joyce's essential strategy for answering moral naturalism involves showing that none of the natural properties offered by naturalists as supervenience-base properties for moral properties satisfy criteria (a) and (b) above. Moral naturalists have successfully defended the idea that natural and cultural selection can build *Humean* moralists, that is, organisms that experience desires to do what is typically deemed the right thing by liberal moralists. This is the point conceded above when we grant that most people are generally well-disposed toward one another, at least when competitive stakes aren't made too high by either enormous prizes or dire threats. But Joyce in effect shows that conservatives (whether Kantian, fundamentalist, or intuitionist) are right to think that having nice Humean dispositions falls short of being genuinely *moral*. In so judging, the conservatives reveal their evolutionary heritage. Liberals, as conservatives will hasten to point out, are prone to waffle by comparison with other kinds of moralists when the going gets tough. Consider Grog and his hunting partner again. The latter is more secure when a tiger charges if Grog is a Kantian or a religious fundamentalist (though, in the latter case, perhaps only if the partner shares Grog's faith) because Grog may believe that his rational duty or his deity allow no loopholes and so might stand his ground despite what terror and good sense combine to tell him.

As Joyce (2006) notes several times, this rejection of moral naturalism is based on *accepting* naturalism *in general*. Citing the value to group fitness of moral fundamentalists in cultural group selection furnishes part of a naturalistic explanation for the prevalence of fundamentalist and Kantian metaethics. Such naturalistic explanation of morality can readily be mistaken for a version of moral naturalism. But for Humeans, taking the naturalist explanation of morality as a *justification* for morality would lead to proving too much. Humeans such as Dennett and Casebeer, after all, are skeptics about Kantian morality, not to mention fundamentalist morality. There are no such things as true categorical imperatives, just as there are no gods. The Kantian and fundamentalist moralists have not been pressured by natural selection

into believing moral *truths*; they believe instrumentally useful falsehoods. Once this is conceded, the point recoils against the idea that nice Humean dispositions should be identified with morality. The Humean doesn't act nicely *because* she correctly believes that doing so minimizes (e.g.) her probability of getting enmeshed in repeated PDs with people disposed to defect. If this were the basis of her nice dispositions then these very dispositions would be undermined, since she'd then recognize that she should defect whenever the expected cost of doing so is lower than the expected cost of punishment. The Humean moralist just happens exogenously to have nice dispositions as a result of natural and cultural selection, and then might rationalize them, thus avoiding cognitive dissonance, by spinning fallacious arguments like that of Gauthier (1986), which purports to show that even narrowly self-interested agents can find it rational to cooperate in one-shot PDs.[5]

We may summarize as follows: it is not enough for the moral naturalist to show us that and how natural selection can produce behavior or sentiments endorsed by prevailing moral intuitions. He must show instead that, given commitment to practical rationality, natural selection can yield moral *reasons*. Gauthier, like Hume before him, at least understands that this is the naturalist's challenge, even if his game-theoretic efforts fail to meet it, for Gauthier acknowledges that his argument must be able to convince even the bearer of the ring of Gyges, who lacks the typical naturally produced desires and has escaped from the force of naturally normal punishments, to behave morally. Since I prefer staying metaphysically at home, when possible, to visiting contrived possible worlds, let us substitute the case of Stalin in power for the ring-bearer. He was contemptuous of conventional ('bourgeois') morality, indeed thought himself a Bolshevik hero for being able to sternly ignore his education in it. That this really was a sincere part of his psychology is suggested by the fact that he enjoyed boasting about his imperviousness to moral sentiment even in the presence of people, such as Churchill, whom he couldn't have expected to share his metaethic. He prodigiously demonstrated his amoral heroism by (among other stunts) setting a then-world record (until Mao) for homicide. Since Stalin correctly (as it happened) supposed that he would not be punished by anyone for all this mayhem, he represents a real ring-of-Gyges case. The moral naturalist can succeed in his project only if he can produce an argument that would have given Stalin pause.

Let us avoid setting the bar *too* high. The moral naturalist need not find an argument that would have actually *altered* Stalin's murderous ways. Stalin's sadistic passions and need to sate his inferiority complex by humiliating and killing more able people, after all, were almost certainly stronger motivators than his rationality. As Joyce validly sets the challenge, the naturalist must only show that natural facts yield moral reasons that motivate behavior *insofar as practical rationality motivates behavior.* But these have to be *categorical* reasons *given* practical rationality, not merely hypothetical reasons that motivate only if one wishes to be nice and liberal, as Stalin didn't. Stalin will be given pause only if we can naturalistically show that his behavior was bound to harm him given his actual desires. The Kantian, as Joyce argues at length, thinks she can accomplish something by ascribing to Stalin an objective interest in being 'fully human,' and then loading 'full humanity' with morally significant properties. This route is not open to the naturalist, for whom it begs the question. Joyce's argument derives extra power from giving the naturalistic moral

[5] For a definitive explanation of why Gauthier's argument is fallacious, see Binmore (1994).

rationalist his best-case analytic scenario (in allowing practically rational but hypothetical desires, and not just actual ones, to connect motivations and reasons; call this 'relaxed Humeanism'), but over-turning his view nevertheless.

Moral fictionalism (MF), Joyce's positive thesis, is his basis for claiming that moral propositions are not useless despite being false. According to Joyce's version of MF, moral discourse is especially causally significant to (some) real behavior because it is built around a distinctive and culturally universal myth, and this special causal significance can and (from an instrumental point of view ratifiable in a game-theoretic equilibrium) should survive widespread recognition that the myth *is* a myth. I will provide a summary of MF in section 4 below, when I consider it, in conjunction with other possible moral psychologies, in pursuit of the rationality of advocating loud policies. First, however, we will consider how far we can get in that direction by means of economic analysis alone.

3. Natural justice

Binmore (1994, 1998, 2005) has given us a naturalistic theory of political-economic bargaining dynamics and norm stabilization that he calls 'natural justice'. Ross (2006a) argues that Binmore's theory has been largely misunderstood by its major critics to date among economists, Gintis (2006) and Seabright (2006), because they have failed to grasp its normative thrust, mistaking it for a mainly descriptive account of the evolution of human sociality in competition with those they have offered themselves (Gintis 2004, Seabright 2004). Binmore is himself partly to blame for this, since he slides into defending an empirical thesis opposed by Gintis, to the effect that no hard-wired preference for fairness evolved among hominids, when all he needs for his purposes is the modal claim that we *need not* posit such an evolutionary discontinuity in order to account for the behavior of contemporary humans in social games. I have defended Binmore's argument for his modal thesis elsewhere (Ross 2006a). Gintis's argument for the empirical thesis takes as a premise his contrary verdict on the modal claim, so because that premise is false his argument should be rejected. On the other hand, the developmental literature on moral judgment, discussed briefly above and surveyed by Nichols (2004), makes a stronger case for Gintis's position on the empirical thesis. This case is explicitly defended by Joyce (2006) (though not in specific defense of Gintis).

This tangle of argument can unfortunately obscure Binmore's main result and point. This is that in extensive-form two-person bargaining games among self-interested economic agents, provided only that they are equipped to take parochial interest in the welfare of *some* other agents (e.g., kin), the egalitarian equilibrium defended on Kantian grounds by Rawls (1971) exists and is reachable. I referred above to the 'normative thrust' of Binmore's theory. This alludes to the fact that Binmore takes himself to be providing a philosophical theory of justice with economic foundations, in the proud tradition of classical welfare theory, because he shows how well-off people in the currently prevailing social equilibrium exploit bad arguments for conservatism to block progress toward superior (more just) equilibria. Binmore's theory is therefore, in its fundamental motivation, advocacy of what I am calling loud policy.

Binmore's model of people and the role of moral norms in their behavior can be summarized as follows. A typical individual person biologically inherits some dispositions to act in accordance with the standard utility function of a sexually reproducing, diploid organism. That is to say, she will be wired so as to behave in approximate concordance with Hamilton's rule where the welfare of others is concerned. Suppose that two such people play a culturally familiar repeated game to which the folk theorem applies. Their problem then consists in coordinating on one of the equilibria within the Nash bargaining set. But how do they identify this set? First, in order to be able to predict one another's strategic and other behavior, they will exploit the species-wide capacity to form empathetic preferences. These are models of others' relative well being that are indexed, in each individual case, to that (other) individual's conception of their own good. As long as these models are not excessively divergent,[6] the bargainers can use them to locate mutual gains from trade (including trade in promises not to plunder one another so long as a certain constitution is observed). The Nash bargaining equilibria at which they arrive will reflect differences in power. However, because we are imagining a repeated game with multiple equilibria, the balance of power is not decisive by itself. To achieve equilibrium selection, the bargainers will additionally rely on a second device: accidents of history in their culture that have generated conventional focal points. The conventions in question tell each individual what constitutes equilibrium behavior in her society. In particular, they tell her when she can expect favors to others to be reciprocated and to what degree, and when, if a favor is not reciprocated, she can call on others to punish the free rider. Since complete conventions must tell everyone what to do in every social role they'll typically occupy at some time, they thus also tell the individual when she is obligated to reward and punish others herself, and when she should expect punishment if she yields to what the community defines as selfish temptation. The set of such conventions constitutes a society's conception of fairness, its moral norms. These can evolve over time if less well-off bargainers demand payments, over and above what the fairness norms dictate, to make up for their lower incentive to maintain the social stability on which the property rights of the better off depend. If this is done, the new distribution sets the fall-back point that feeds into the next round of Nash bargaining, and so there is incremental convergence over successive rounds towards egalitarian distribution. As this convergence occurs, people revise the moral narratives that encode their evolving norms. They might have started off, for example, agreeing with Aristotle that slaves owe almost everything to their masters and are owed almost nothing in return, but end up as Rawlsians who think they must make the former slaves as well off as possible when determining what share of profits earned by their greater resources they should draw for themselves.

If this is read as a purely *descriptive* account of political-economic bargaining, as Gintis and Seabright seem to read it, then anyone who is concerned to improve the welfare of the less well off, or who merely values peace and stability, is likely to be struck by its optimism. This optimism seems at first glance to enjoin quietism: if hard-headed Nash bargaining among the rational will carry us smoothly toward justice, then why rock the boat to aid the less fortunate? Indeed, since Binmore is clear that falling off the equilibrium path is typically catastrophic for a society's general

[6] In Binmore's model, divergences are themselves bargained away in signaling games, by which people converge to 'empathy equilibria' – vectors of empathetic preference assignment in which no one has any incentive not to signal honestly.

welfare, might there not even be, given his account, a rational *requirement* to be quiet?

Putting the question this way leads straight to an answer. In light of his recognition that most societies have at one time or another undergone violent revolutions that represent breakdowns in moral norms and are usually enormously destructive of lives and assets, Binmore clearly doesn't think that societies inevitably or naturally move along the equilibrium path to egalitarianism. Furthermore, he appreciates that convention disruption, whether caused by technological change, cross-cultural migration, or drift represented by subcultural speciation within a population (as may accompany, for example, class polarization), greatly complicates outcomes. The quietism concern that might arise for anyone who misses Binmore's reformist purposes thus should not be the crude one just canvassed. However, it might seem that on his model *as long as* social stability is maintained, *then* the evolution of egalitarianism should be expected as a scientific prediction. This, if it were what Binmore intended, would license an argument that should greatly please a conservative, especially because it derives conservative advice from the normative premises typically promoted by liberals and socialists.

However, we can best convince ourselves that this is not an accurate reading of Binmore by addressing our attention to what is supposed to be *moral* about the norms that solve distributional coordination problems. Binmore's conception of moral beliefs as culturally evolved bargaining norms is fully compatible with Joyce's denial of moral facts (as Binmore intends it should be; his jeremiads against the Kantian view of morality are both logical and rhetorical tours de force). He does not, however, explain why people generally do not see the conventions for what they are; his account does not explain why some, but not most, cultural norms are *moralized*. Someone might point out in this context that moralization may be of heuristic value. Social conventions are most useful if they apply to classes of logically similar situations that are not reliably marked by stereotypical perceptual salience cues. Institutional moral conventions probably help – perhaps even essentially help – to encode these more abstract relationships. In addition, as Binmore argues, explicit fairness norms can enable people to coordinate on equilibria in novel situations. These will be the kinds of situations in which people notice and explicitly invoke their own norms. By contrast, when they deploy them in familiar scenarios, they'll tend to be unaware of doing so. Thus, as Binmore puts it, medium-run use of a body of fairness norms will tend to leach out the norms' moral content and reduce them to purely habitual judgments about the relative status of different sorts of people. What begin as moral norms become, in effect, thoughtless prejudices.

As argued in Ross (2006a), here lie the roots of the kind of social corruption that Binmore's work aims, as a normative project, to upset: people can fail to notice when changing circumstances make better equilibria available to them, because they go on playing habitually rather than re-packaging their cultural salience landscapes in light of new feasibility sets. Indeed, some people who enjoy privileges that others could strip away *while still staying on equilibrium paths* will be incentivized to actively suppress general recognition that new fairness norms are feasible; these people represent each generation's version of conservatism. Binmore's work is intended as a weapon against conservatism that at the same time avoids utopian disdain for the

importance of stability (that is, for the importance of keeping institutional conventions on equilibrium paths).

As said above, Binmore's is primarily a project in political philosophy grounded in behavioral science, rather than a first-order intervention in behavioral science as Gintis and Seabright take it to be. He aims to convince us that even if people are not hard-wired to prefer fairness *in itself* (on some, perhaps culturally relative, conception) we are not stuck with a Hobson's choice between conservative refusal to risk stability for the sake of the less advantaged, on the one hand, and reckless utopian schemes, on the other hand. We can instead examine our current fairness norms, identify their equilibrium conditions, and then define forward equilibrium paths that dynamically promote both welfare efficiency and the relative well-being of the worse off. As pointed out in Ross (2006a), recognition that this is how Binmore should be read explains why what I earlier referred to as his modal disagreement with Gintis and Seabright is more important than their empirical dispute. It is an intended virtue of Binmore's normative theory that it should persuade people to support its proposals for institutional reform along equilibrium paths if they endorse only limited concern for only particular others. The rich and powerful, Binmore seeks to show, can venture forward from the status quo without having to sacrifice expected utility at any point; this is the property that makes Binmore's reform path an *equilibrium* path. Of course, as they move along the path the rich *will* surrender some *wealth*. They are incentivized to do this in exchange for gains in security against loss of their property rights in a Hobbesian meltdown, which is why they lose no expected utility. However, as I noted in the opening section of the paper, this incentivizes them to try to keep the poor from learning about their own bargaining power, if possible, and that in turn explains the sound intuition that conservatism crucially rests on suppression of truth. If Binmore's analysis were common knowledge, then the self-interested rich should sign on to the new social contract. So the self-interested rich should not recommend Binmore's work to the poor or to the more benevolent rich.

Binmore's strategy here reflects two excellent principles to which all political philosophers should swear oaths. The first is Hume's principle. It tells us that wherever we can accomplish our ends using either, on the one hand, institutions that can be exploited by the unscrupulous, or, on the other hand, institutions that are proof against rational knavery – that is, narrow selfishness – then we should choose the latter. The second principle is that useful political critique does not consist in condemnation of the status quo from godlike perspectives that suggest that the critic is not herself a player in the social game. Reformers should instead propose specific bargains to specific agents who are presumed to possibly hold preferences that differ from their own. It accomplishes nothing when Kantians preach to Kantian choirs, Marxists to Marxist ones, and libertarians to libertarian ones. Most people are not Kantians or Marxists or libertarians and are never going to be. The political philosopher, Binmore often stresses, should address people as they actually are, if institutional reform is truly what she cares about.

So far, I submit, so good. Thanks to Binmore, we can see how and why someone can consistently deny that there are moral facts while urging loud reforms, especially economic ones of the sort that represent Kaldor improvements: non-vicious preferences (on the part of the economist) furnish sufficient motivation. But this suggests that if the economist continues to moralize her advocacy of the loud policies

after reading Binmore and Joyce, she is engaging in a kind of deceit, or, at the very least, arch condescension. If she says "It is morally wrong for inhabitants of rich countries to herd poor people within the borders of poor countries under conditions where significant economic improvement appears to be unlikely or impossible," then she is saying something she knows to be false. Perhaps it will be more politically effective than saying "There is an equilibrium path in global Nash bargaining that leads to relaxation of restrictions on the movement of people across borders". But is the former moralized expression of the preference necessarily merely manipulative? Is the only difference between the moralized and the non-moralized versions of the preference rhetorical?

To try to make progress on these questions, let us turn back to Joyce and to his theory of moral fictionalism. I will argue that moral fictionalism does not furnish a fully adequate answer to this question, but that it takes us close to considerations in moral psychology from which such an answer can be derived.

4. Moral fictionalism

Let me now be more explicit about where Joyce and Binmore respectively stand on the main subject of this essay, the rationality of preference moralization (that is, expressing preferences as moral judgments). Joyce thinks that preference moralization is rational. His interpretation of this is that rational moralization consists in deliberately continuing the narration of what should be recognized to be the *myth* that there are objective moral obligations.

We can best understand this idea by considering different ways in which people participate in myths. Imagine first the fans of a local baseball team. All might know that their team isn't actually more worthy of support, for any objective reason, than teams based in any other city. But it's fun to cheer for the home team, and holding in the forefront of general acknowledgment that the home team doesn't really *deserve* loyalty would ruin this fun; so fans collectively pretend, and all are better off for doing so. Similarly but more significantly, histories of moral enculturation are highly valuable, perhaps even irreplaceable (at least in one fell swoop) devices for social coordination on important collective problems. It might be difficult, perhaps even impossible, for people to coordinate on their interpretation of these histories they share unless they pretended the histories in question were, like bodies of genuine legal precedent, interpretations of holy or natural (Kantian) law. Note that if, as per supposition, the community we are imagining is a community of moral fictionalists, they will not believe in the existence of the gods they talk about or the truth of the categorical duties they cite. The gods and duties will serve merely as collective reference points. The people will show their sophistication in their philosophy seminars, where anyone who demonstrates belief will be viewed as embarrassingly naïve. Unfortunately, Stalin is also sophisticated in this way, so if he gets the equivalent of his ring of Gyges – his dictatorial power – the carnage will be on. But fictionalism might show its advantages in light of just this consideration, because remembering that morality is a collectively maintained *pretense* gives us reason not to rely on it for more than it can deliver, and to be extra careful never to let anyone have dictatorial power.

Binmore's attitude to moral norms is compatible with all of this. His idea that morality becomes corrupted into thoughtless habit over time finds an interesting anticipation – with a twist – in Joyce's earlier (2001) discussion. For Binmore, norms are corrupted when people forget their original justification. This is at least *logically* akin to what a moral objectivist is likely to regard as moral corruption, namely, abandonment of the conviction that a moral norm is metaphysically sanctioned. On Joyce's more specific account, what counts as corruption is turned upside down relative to what the objectivist has in mind. Its paradigm case according to Joyce would be a community of fictionalists who forget that their gods or Kantian obligations are myths and who thus become incapable of adapting their norms to changing circumstances. It is not clear that Joyce would share Binmore's judgment that evolution of self-conscious pretenses into mere habits constitutes corruption, since, as will be discussed below, being habitual is one of the most valuable properties of a moral fiction on Joyce's account.

In my view Joyce and Binmore are each right with respect to issues on different time scales. On the one hand, an individual person, as long as she lives in a society where most preferences are non-vicious, is probably better off on short timescales to the extent that prevailing moral norms have become habits for her, since others will likely then pick up on and reward the fact that she will be a reliable partner for game-playing. This point of course applies to all other symmetrically situated people. But Binmore is also right that societies where people have forgotten the rationales for their moral principles are likely to be inefficiently conservative. Such conservatism can be the path to long-run social ruin; consider imperial China.[7] As we have seen, this concern is what motivates Binmore's reformism.

In considering why fictionalism might be recommended to an agent, Joyce recognizes one line of justification that is broadly economic in nature but which Binmore neglects. This involves *picoeconomic* reasoning (Ainslie 1992, 2001). A great deal of evidence shows that people, at least outside of heavily structured institutional contexts such as financial markets, structure preferences over intertemporal rewards according to hyperbolic future discounting functions that generate dynamic preference reversals (see the papers in Herrnstein 1997 for evidence). The main phenomena to which this insight has been applied by behavioral economists are addictions. However, Joyce (2001, pp. 210-218) notices that the idea is also directly relevant to normative coordination. A person may recognize that prudence advises respect for prevailing social norms: if one is *not* in Stalin's situation, then trying to cheat when one thinks one can get away with it is likely to trade off a few higher payoffs for eventual disaster, because most people are led to overestimate every special skill they cultivate, including skill at social crime. However, such recognition will tend to go along with felt temptations to cheat "just this once"; and hyperbolic discounters will be prone to succumbing to such temptations. As Ainslie shows, the principal mechanisms people use to combat hyperbolic discounting in general are *personal rules*. The dyspeptic person, for example, might establish a policy according to which she may only get drunk on Friday nights. If no one but herself will punish her should she break the rule, then to be effective it must be self-enforcing; but how? Ainslie's answer appeals to

[7] Both Joyce and Binmore would agree there is one still worse form of general corruption: the society where people have both forgotten the justifications for their norms *and* have become cynical about the idea that there ever *were* sound justifications for them. Such societies are likely to devolve into mafia states – perhaps like contemporary Russia?

the fact that the rule itself is a present asset to the hyperbolic discounter. Because a lapse against the rule *predicts* further future violations, if it can't be given a convincing principled excuse then it *immediately* devalues the asset and thus gets its motive force working in the present, where it can regulate behavior despite hyperbolic discounting. Joyce notices that moral rules are exactly of this logical character. The non-religious dyspeptic knows that the *universe* doesn't prevent her from going on a Tuesday bender, even though she pretends it's a *fact* that she can't, so she is a fictionalist about her personal rule. Likewise the sophisticated person who internalizes a moral rule to the effect that one must never chisel from the company pension fund might be moral fictionalist.

Notice that at this point what's 'fictional' about a moral rule has become quite subtle. On the picoeconomic account, a personal rule is quite real (though virtual – being virtual is a way of being real). If it weren't, it wouldn't be a valuable asset and the whole story would lose its economic sense. Note also that, in our example, the dyspeptic person need not adopt any pretense about anything: she is best off if she knows that the fate of the asset is in her control. Advocates of so-called "twelve-step" movements such as Alcoholics Anonymous deny this, insisting instead that people suffering from chronic preference-reversal syndromes should entrust maintenance of their personal rules to a "higher power". There is evidence that this often tends to amplify small stumbles away from abstinence into headlong binges, since the person expects in advance of a lapse to find herself powerless if and when the lapse comes (see Fingarette 1989). Where fictionalism enters the picture in the case of *moral* personal rules, on Joyce's account, is that what makes these rules instances of moralization is the rule-adopter's pretense that they apply to her *because* they apply to everyone. That is, she treats them as if they were categorical, and derive their force from this feature. Notice, however, that to the extent that this is so then the mechanism by which moral personal rules work *isn't* Ainslie's picoeconomic mechanism after all; it is more like the mechanism of the "Twelve Steps" follower. I suggest that the relevant negative parallel applies here: the fictionalist is likely to be at risk of binging on immorality whenever she remembers that her moral rules are enforced by pretenses. If she is *not* then this must be because the picoeconomic mechanism is sufficient to hold up the side; but in that case why bother with the fiction in the first place?

I suggest that there are only two general kinds of answer available to Joyce here. The one for which there is most evidence in his own writing is to lean harder on the importance of unthinking habit. "Someone armed with moral beliefs," he says, "thinks of cooperation as categorically required, and the distinctive value of categorical imperatives is that they silence calculation" (2001, p. 213). Of course, the fictionalist is not supposed to have moral *beliefs*; but, as Joyce points out "moral images" – by which he means unreflective but salient thoughts produced by mental habits – "may well *seem* to the subject to be very much the same as beliefs" (p. 219). I have already given some indication of my attitude to this line of reasoning. I agree that habits of this sort can be – indeed are – quite useful to people where all that is under consideration is their own probability of playing by the local normative rules from occasion to occasion. However, to the extent that fictionalism leans on this defense it is poorly equipped to address the sort of long-scale corruption that worries Binmore. We can put this in the context of the preoccupation of the present essay by asking

rhetorically: what could possibly be less conducive to support for loud policies than moralized habits? Are these not precisely what conservatism is all about encouraging?

Thus I concede that Joyce might provide the basis for a positive answer to the question "Can moral fictionalism be prudentially recommended as individual psychological therapy?" But the question presented by reflections on Binmore's theory is: "Is there a way to reconcile an error theory of morality with non-cynical moralized advocacy of loud policies?" The point made just above is that moral fictionalism buttressed by appeal to the value of mental habits is an unpromising – to put it mildly – avenue to this.[8]

I said above that there is an alternative emphasis available to Joyce. It is 'available' only in a Pickwickian sense. Opting for it threatens to – and perhaps more than merely threatens to – drain the semantic appropriateness out of the label 'fictionalism'. I am of course not suggesting that Joyce might fail to follow where logic leads merely out of attachment to the name of his theory. Rather, the suggestion is that Joyce calls his (positive) theory "fictionalism" in part because he thinks the habit defense of moralization is superior to the one I will turn to next. To the extent then this so, then it seems to me that he and Binmore part philosophical company. I will continue traveling along the path cut by Binmore.

5. Moralized preferences

If moralized advocacy of loud policies can possibly involve anything beyond rhetorical / emotional manipulation, then this must imply that we are not, as Joyce believes us to be, completely without recourse in the face of the bearer of the ring of Gyges (e.g., Stalin in power). One moralizes the preference for the loud policy in hopes of giving people who haven't seen a prudential reason for adopting it to change their minds. The loudness of a policy is relative to times and places: advocating abolition of slavery was loud in America in 1820 and opposing slavery is about as quiet as can be in America in 2006.[9] What makes a policy loud in *p* at *t* is that it involves significant opportunity cost to implement and most people in *p* at *t* don't support it. If expression of moralized preference for the policy is to contribute to changing minds, then this must be because encountering the moralized preference confronts people with a new relevant reason for belief they'd previously been missing. But this is exactly the ring-of-Gyges problem. So we can make progress past the point where Binmore's argument left us at the end of section 3 above only if, pace Joyce,

[8] Joyce (personal correspondence) points out that there may be other reasons for being a moral fictionalist besides appeal to its value in shoring up personal rules. This must of course be admitted. However, I think that Joyce is right that moral fictionalism *is* apt to shore up moral personal rules. And then *my* contention is that reliance on *moral* personal rules (in contrast to Ainslie's prudential ones) is favourable to conservatism in social policy.

[9] I recognize that slavery continues to exist in the world and that most Americans aren't presently acting as if they think that abolishing it is a policy priority. Urging that people oppose slavery is nevertheless a way of being quiet in contemporary America. Urging an immediate military invasion of the Sudan to free people held as slaves there, which goes well beyond opposing slavery in general, or slavery in America itself, would be loud. So let me add that I wish the United Nations Security Council would declare the Sudanese government a criminal organization and authorize its urgent military overthrow by a coalition of member countries.

that problem isn't totally stubborn. I'll now offer some considerations suggesting that it isn't.

As discussed above, in considering the value to contemporary people of the disposition to moralize, Joyce emphasizes self-control functions. However, in speculating about the *evolutionary* functions of morality, Joyce (2006) argues that its main contribution lay in its capacity to stabilize intragroup helping behavior. This is also very plausible. Following Binmore's lead, I suggest that we generalize further. What constitutes 'helping' is a function of agents' utilities and sets of feasible projects. These cannot be determined afresh on a case-by-case basis, because if they were this would open room for strategic manipulation of private information that would tend to unravel equilibria. Widespread cooperation in any community of organisms depends on stable expectations about what others want, about what they can and can't do, and about the extent to which such expectations are shared and generally known to be shared. At a very abstract level, therefore, the fundamental games underlying the evolution of sociality are coordination games. By this I don't refer to what might be called 'local' coordination, where all elements of a game are antecedently clear and players merely need focal points to converge on one of the Pareto-efficient equilibria. Rather, I refer to a more abstract and complex (dynamic) sort of coordination, in which what matters to fitness over ranges of interaction across many agents is that individuals' strategies support *some or other* relatively stable equilibria, rather than that they be in *particular* equilibria. The evolution of a signaling system in a population is a solution to a coordination game of this sort; the system will tend to be selected for properties that make it at least a second-best solution across commonly recurring types of games, rather than for properties that might make it a best solution in a more restricted set of game situations.

Evolutionary history has featured two broad classes of mechanisms for stabilizing complex social coordination capacities. One approach directly builds on kin selection, as in hapladiploid creatures who, as a result of their genetics, live in communities of unusually close relatives. The second approach involves equipping animals with brains big enough for computational processing of signal meaning, both with respect to sending and receiving, in non-parametric environments. I have argued elsewhere (Ross forthcoming 2007) that the importance of such resources for non-hapladiploid sociality, rather than the overly specific Machiavellian hypothesis (Byrne and Whiten 1988) explains why all very intelligent creatures we know of are highly social.

The second design strategy solves one problem by creating another: large brains must employ distributed control to avoid being fatally slowed down in their reactions by internal processing bottlenecks. Distributed control makes agency itself, here interpreted, as always in economics, by reference to preference stability, inherently unstable (Ross 2005). How can big-brained creatures avoid being money-pumped by less flexible but more consistent small-brained ones?[10] Furthermore, if the point of the large brain is to facilitate interpretation of complex and adaptive conspecific behavior, does evolution have any way around hitting a low ceiling on the intelligence solution arising from the fact that distribution of control undermines the very feature for which

[10] It has been established for years that rats learn more efficiently than humans in simple operant conditioning environments where abstract decision rules offer no improvement on pure matching.

the large brain is built, viz., arriving at reciprocally stable behavioral predictions in interactions?

Ross (2005, 2006b) argues that *H. sapiens* achieves higher levels of complex coordination than other non-hapladiploid animals by means of a disposition to engage in socially regulated construction of narrative biographies. The basis of a person's distinctiveness and coherence is *narrative*: people are allowed to fully participate in networks of reciprocation just in case they are able to narrate relatively consistent histories of their dispositions, actions, tastes and motivations that other people will reliably interpret and respond to as a *biography*. (For similar ideas see Bruner 1992, Hutto forthcoming 2007). A biography is (roughly) a teleologically structured history of an entity whose behavior and manifest emotional expressions can be non-redundantly predicted and explained using *the intentional stance* (Dennett 1987) – the perspective that organizes behavioral data against a backdrop of attributed beliefs, desires and similar representational states,[11] the so-called 'propositional attitudes'. Psychologists study such processes under the rubric of 'the social construction of the self'.

As Joyce (2006), following many other commentators over the past two decades, recognizes, stakes in reputations are the most basic commitment devices available among non-kin in typical ecological circumstances. In this context, narrative selves fuse two functions: (1) they provide structure that allows specific reputations to be encoded and remembered using mnemonic devices (natural story plots, as it were) biologically natural to humans; and (2) the coherence requirements that govern them embody strategic commitments – agents know that if they re-narrate themselves too freely in order to allow more strategic flexibility, they risk being regarded as unpredictable (in contemporary English, labelled 'flighty' or 'two-faced') and to be excluded from potentially profitable relationships and projects. Self-maintenance equilibrium dynamics are therefore self-enforcing. Supplementing these considerations is the need for coordination in the narrower sense familiar in introductory game theory: many games that are not strictly competitive have multiple equilibria. Relatively fine constraints on stability of selves (e.g., "She's the sort of person who puts more value in an exciting new experience than in security") facilitate joint equilibrium selection in such games. Social constraints on self-narration produce enough homogeneity to allow social norms to stabilize; then the fact that, within these constraints, agents do *better* (by attracting a wider range of interaction partners, and being paid higher social wages for scarce skills) if they aren't boring facilitates fine-grained focal point coordination in specific games, and allows for gains from trade of complementary social talents.

Ross (2006b) sketches a game-theoretic modeling framework for the development of selves that I call 'game determination theory'. In this framework, groups of selves play games that yield, among their dynamic equilibrium outcomes, new games to be played among narratively revised selves, where revision is interpreted as preference change (and therefore, given requirements of economic formalism, as creation of new

[11] I here defer to standard usage in the philosophy of mind. However, as Mark Rowlands pointed out in a recent conference presentation, thinking of beliefs and desires as *states* has probably contributed to much confusion about the nature of mind. They are better thought of as *processes*, since their attribution, either by a person herself or by another, labels and organizes a stream (however short) of behaviors and expected accounts and consequences of those behaviors.

agents). This is apt to look at first glance like constrained maximization, as in Gauthier (1986), Danielson (1992), McClennen (1990), and Nozick (1993), if it is interpreted as suggesting that agents act against current preferences to pre-commit future selves. However, this (formally incoherent[12]) idea is not the intended interpretation. Game determination theory appeals to picoeconomic mechanisms as discussed earlier. Picoeconomic games are strictly non-cooperative and picoeconomic agents that observe personal rules act to maximize only present (not even expected) utility (Herrnstein 1997). Where game determination specifically is concerned, agents are motivated by threats of punishment, especially damage to reputational assets through cheap resorts to gossip, and by Ainslie's personal-rule mechanism then controls in its usual way for hyperbolic discounting of the value of these assets. I will consider the role of preference moralization within this framework.

I begin by stating necessary conditions on a preference's being moralized. I do not claim to derive these from *a priori* conceptual analysis. Rather, I base them on the empirical claim that if a preference lacks any of the conditions I state, then in no human society will people (isolated eccentrics aside) respond to the preference as legitimately moralized. That is, they will not demonstrate the distinctive sorts of behavioural patterns with respect to which moralization is identified as a genuine empirical phenomenon. For agent *i*'s preference *P* to be generally regarded in a society as moralized by *i* just *is* what it means for that preference to be moralized by *i.*

The first necessary condition on a preference being moralized is that it must be socially signifiable. By this I denote the idea that moralization is a social institution, that is, that to be moralized a preference must express a *norm*. Following Bicchieri (2006), but with substantial modifications, I define two kinds of norms, *descriptive* and *social*, as follows: a descriptive norm is instantiated for a class of games *G* in a group *H* if there are at least two agents *i, j* in *H* such that *i* and *j* prefer to φ in game *G* if they expect more than *y*% of their reference group to φ in *G*. Typical φ's might include wearing mini-skirts, getting drunk on Friday nights or using chopsticks at Chinese restaurants. Descriptive norms are not appropriate candidates for moralization, in the sense that if a person moralizes a descriptive norm this will not be strategically effective because others will regard the moralizing behavior as deviant. Candidates for moralization are an individual's preferences with respect to social norms, defined thus: a social norm is instantiated for a class of games *G* in a group *H* if there are at least two agents *i, j* in *H* such that *i* and *j* believe that if more than *y*% of a reference group φ's, then there is an expectation generally regarded as legitimate that everyone in the group (or everyone not exempted from the expectation due to special status) will φ. Typical φ's include honouring promises, respecting others' basic liberties, or refraining from non-harmless lies. (In some non-liberal communities they include praying several times a day, shunning heretics, or mutilating girls by clitorectomy. Many social norms are, by liberal lights, foolish or atrocious or both.) Note that although I thus claim that a person who lived as a hermit from birth could not have moralized preferences, I don't rule out secret moralized preferences, or even moralized preferences that can't be communicated to others, so long as they descend from norms. So, for example, Robinson Crusoe, standing on his lonely beach watching a warship in the distance machine-gun survivors of its sunken foe, could be

[12] See Binmore (1994).

sincerely disgusted and shake his fist and stomp his feet and shout into the wind; but only because he was brought up in a society.

The second necessary condition on moralization of a preference is that both the preference itself, and its moralized status in the utility function of the moralizing agent, must be stable enough to be potentially used to try to influence some strategic outcome or outcomes. The point of this condition is to exclude fleeting states of righteous anger that are suppressed before they effect behavior. These are just emotions, which typically accompany moralization, and are probably crucial cues in learning to moralize, but are not treated as equivalent to it.

The third necessary condition on moralization of a preference is that it must involve willingness to pay for the project of encouraging the world into line with the preference, in a way not indexed to any *ad hoc* set of specific individuals. For example, suppose I have (correctly, as it happens) a moralized belief about the goodness of various liberal democratic virtues. It is evidence that this *is* a *moralized* preference that I'd voluntarily, under a range of plausible circumstances threatening the persistence of these virtues, agree to the drastic behavioral step of going to war for them. However, willingness to pay dramatic costs is not necessary for moralization; many moralized preferences are held timidly. Here, as generally, non-cognitive motivators such as fear and lack of energy create noise between preference and the behavioral basis on which preference is inferred. Genuine moralization is thus usually revealed only by *patterns* of behavior, most of which will be personal, idiosyncratic and verbal rather than actively political. People will also often refrain from trying very hard to promote the projects recommended by their moralized preferences because their subjective evaluations of the probable efficacy of doing so suggest that it's pointless. And, of course, these evaluations are subject to processes of rationalization and self-deception. What *is* necessary for my preference to be a moralized one is that I don't favour maintenance of liberal democratic virtues merely in application to the governance of my friends, or for the duration of my own life or my spouse's life, or just in my own country, or otherwise restrict the scope of the endorsement in parochial ways.[13] Allowing for non-ad hoc restrictions (e.g., democracy can be restricted in war-time, or during plagues, or doesn't fully apply amongst groups including 6-year olds, etc.), then, the preference has categorical force, as Joyce argues that a moralized belief must do.

The fourth necessary condition on a preference being moralized is what I call the *lexicalization constraint.* This is that the holder of the preference behaves in a way suggesting she recognizes an incentive not to appear to reverse preferences as between two baskets X and X' just because the cardinal utility values of non-moralized items in X' are increased at the expense of utility values on moralized items in X.[14] Put crudely: if you can get someone to abandon behavioral expression of a

[13] It follows from this that *true* cultural relativists do not have moralized beliefs about any cultural values. Of course, most people who espouse some relativistic opinions are not *thorough-going* relativists, either because they are inconsistent or because they have sincere beliefs about real differences amongst groups that are not just rationalizations of parochial interests. Every reasonable person should be a relativist about *many* cultural differences, including quite important ones.

[14] Rawls (1971) grasps this point in having his social contract negotiators behind the veil of ignorance, who share a range of moralized preferences, maximize an index of primary goods rather than expected utility. I agree with Binmore (1994, 1998), however, that it vitiates the point of social contract theory to impose particular moralizations exogenously on the bargainers as Rawls does. Binmore even makes the

supposedly moralized preference by bribing her with reference to non-moralized aspects of her utility function, these are *prima facie* grounds for denying the agent's claim to have moralized the preference. If this dis-entitlement to claim to have moralized the preference itself represents a cost to the agent, as it typically will, then this is sufficient for moralization not to be cheap talk. Consider, to continue the previous example, how a non-liberal might try to talk me out of my moralized preference for liberal democracy. One route would involve showing that I have false beliefs about liberalism's consequences for, say, poverty. This would involve appeal to a reason, the kind of reason I take to be relevant in evaluating moralized preferences about political systems. But because my preference is moralized, a tactic that *won't* work is trying to show that the defeat of democracy in, say, Zimbabwe, won't harm me or my family. This is enough to show that moralization of preference makes a causal difference to behavior, since you can't change my behavior (at least, not *all* my relevant behavior, e.g. when the secret police aren't around) with respect to moralized preferences *just* by putting pressure on my environment of incentives, something that would work (given no restrictions on available pressures) for *any* of my non-moralized preferences.

I claim that the above necessary conditions are jointly sufficient for moralizing a preference. That is, in any human society if someone regarded as competent for agency in that society (so, perhaps not a child or a deranged person – in some unfortunate societies, not a slave or a woman) expresses a preference using that society's signaling conventions for moralization, then people will modify their strategic responses to the preference in question in ways characteristic (in that society) for moralized preferences. Because there are substantial similarities in strategic responses to moralized preferences across all societies, we have a basis for treating moralized preferences as a social kind.

So far I have claimed nothing with which, as far as I can see, Joyce provides grounds for quarrelling. However, according to him the idea that some preferences could be objectively imprudent objects of moralization, in all circumstances, is something that can be maintained only as a myth. He of course recognizes that it is prudent for people to moralize, mythically or otherwise, those preferences that most people in their communities moralize. But, to restate our main problem here, this seems to imply quietism. Let us see if we can find considerations that might encourage loudness instead.

As we saw, the lexicalization constraint says that you can't shake my moralized preference for liberal democracy just by bribing me. However, consider another tactic you might try. Suppose you could show me that my having a *moralized* preference that *other people* enjoy democracy offends those people – it strikes them as imperialistic. They, as it happens, have moralized preferences against imperialism that trump their moralized preferences against their dictators. As a result, if I continue to express my moralized preference for democracy I will, by doing so, undermine the prospects for the triumph of democracy that I prefer. Furthermore, suppose that I must interact on a regular basis with the people I'd like to see liberated and that if I continue to entertain my moralized preference on their behalf it will show in blushes

capacity to moralize endogenous in his model of the long run, which in my view would represent a methodological breakthrough in social contract theory even if all of the details of his theory were rejected.

of anger and irregularities of tone and diction that I can't control. (These people are very acute observers, or I'm a lousy actor, or I know I'll forget myself in the bar on Friday nights, or etc..) I might now be rationally motivated to de-moralize the preference. This need not entail *abandoning* the preference; the locals don't mind my saying "I like to live in democracies" or "Democracy would be good for your economy" or even "If you set up a democracy I'll pay you a million dollars," since these just express combinations of factual claims and statements of my personal preference without being ipsidexist. So, practical rationality will in these circumstances dictate that my concern for democracy should motivate me to de-moralize my preference for it, if I can.

Can I? It is crucial for the argument I'm engaged in that I can't do so *merely* by deciding. If moralization of preference consisted in nothing but using moral language to express standard preferences, then it would be subject to costless strategic use, and so would in fact be strategically useless; all moral expression would be taken as cheap talk by rational agents. To be strategically significant, moralized preferences must be partly constitutive of narrative selves, and coherence of the narrative self must be an (instrumentally) valuable *achievement*. As explained earlier, selves can be (are, indeed, expected to be) re-narrated over time, but this takes effort. Furthermore, if re-narration is done abruptly it signals insincerity and unreliability and so carries costs (often extreme costs). Finally, re-narration is not approved of when it is embarked on whimsically and unilaterally; it is expected to be a response to pressures from other people (that is, in my framework, to be a product of determining games). So: if you persuade me to de-moralize my preference for democracy because its moralization undermines its object, then I will have to work at doing so, and the project will take some time. What is crucial for my purposes is that people *can* be *rationally* persuaded to undertake such efforts at self-re-narration; and the fact that it takes effort shows that moralization is a source of real constraint that *calls for reasons* to be broken, when the constraints become too expensive.

This point is a route to a more general one. Moralization of preferences implies standing *projects*. Furthermore, these projects extend beyond the immediate surroundings – indeed, beyond the lifespans – of the particular moralizers. This is why, if my moralized preference for democracy really does undermine democracy, and the costs of self-re-narration are not impossibly high,[15] I should be rationally motivated to de-moralize it. Otherwise I'm caught in a contradiction detectable by application of practical rationality: I simultaneously morally prefer the triumph of democracy, but behaviorally reveal a preference that it fail. This might not look immediately like a *contradiction.* But moral preference, on the account I am giving, is *not* 'non-behavioral' preference – there is no such thing. Therefore: whether I *actually* de-moralize my preference for democracy or not, if (under the imagined circumstances) I fail to acknowledge that I at least have some reason to do so, I violate practical rationality.

Notice that we now have something Joyce argues doesn't exist: a way in which non-moral reasons govern distinctively 'moral' behavior, without appeal to any

[15] To see how they could be: suppose I'm Archbishop Tutu. I couldn't then demoralize my preference for democracy without destroying most of my publicly accessible self; and building a stable publicly accessible self is part of the point of building a self at all. Tutu's role is to tell his public what's worth moralizing, not to be just another political analyst.

distinctively moral facts. Because there are no moral facts independent of particular biographies, practical rationality can tell no one what to moralize independently of the properties of the self in which they have invested. Joyce's `relaxed Humeanism' described in section 2 is one way of recognizing this point. But practical rationality *does* constrain the objects of moralization *if* everyone rationally needs a self given (biologically normal) desires *and if* only some moralized preferences are compatible with selfhood.

Notice also that if an economist has a moralized preference that welfare not be wasted, and if most or even many other people participate in the institution of preference moralization (as of course they do), then we have arrived at *a* rational basis for the economist's sincerely moralizing his preferences for loud policies even if he doesn't believe in moral facts. In expressing his preference moralistically, he signals that (unless he is lying) his preference has all of the properties necessary for such preferences as were reviewed above. This information is relevant to others' behavior: they will at least have to pay the economist more to be quiet than they would if his preference were non-moralized. Now, if this were the only strategic consequence of preference moralization, then it still wouldn't amount to much in the context of our problem. It would imply that preference moralization can be highly useful in friendship networks, where people are motivated to care about how much it costs to influence the behavior of particular individuals. But our concern here is with moralization of public policy preferences. This has a possibility of achieving its normally intended effect if the economist can persuade others to moralize *their* non-vicious preference against welfare waste, and if expression of sincere moralization of his preference is partly responsible for this.

This brings us directly to the ring-of-Gyges issue. If Stalin in power can be given a reason to moralize some specific preference or other (even if he might not act on that reason because other psychological forces prevail), then I will have shown, in the hardest possible case, what I'm after: that moralization of preferences can be rationally, and not just rhetorically or emotionally, persuasive.

I begin this argument by reiterating some earlier points. All non-infant people without neurological pathologies moralize some preferences. They do this as a result of biologically evolved processes that raised the fitness of their ancestors by promoting social coordination. Joyce (2006) agrees with this claim. Put in terms of game determination theory: a self cannot be sufficiently stable to do its job, that is, regularly recruit other selves to cooperative projects and then achieve coordination with them, without moralized preferences. In their absence, others will have difficulty receiving crucial information about preference cardinality and the locations of contract curves in Edgeworth boxes.

Stalin was a stable, comprehensible (terrible) self; so Stalin had moralized preferences. It has been among the tasks of Stalin's army of biographers, as with all biographers, to try to figure out what they were. What Stalin moralized was, to a liberal person, as nasty as could be. He morally preferred a world of ruthless, effective maximizers of political and military power turning the planet into a sort of gigantic Ford plant. In his morally preferred world people work prodigiously (as he did), in the narrow sense of `work': they put in long hours at dedicated work-stations. They have significant disputes about nothing that it isn't purely technical and related to

maximizing productive output. They consume `art' not to deepen their personal subtlety – a quality they despise – but to help them coordinate on their output targets. They are earthy and straightforward in their personal manners – appreciation of the attractiveness of this was Stalin's most normal human trait, and his only (in moderation) attractive one – but (ruining the value of the quality) they are all uniformly like this. Finally, and crucially, they directly extract whatever labor-power they can from each other, and in any instance where the effort required for such exploitation exceeds the productive value obtainable from it, they kill the people in question without further thought, taking pride in their absence of sympathy or emotional consideration.

I claim that this portrait of Stalin's preferences is consistent with the biographical facts. I claim furthermore that these preferences of Stalin's were *moralized* in the conceptual sense I have articulated. That is, he cared, consistently and, by all available behavioral evidence, quite deeply, that the actual world should come to resemble his preferred one, both in his own lifetime and after his death. His policies, through all their zigs and zags, and his recorded conversations, were consistent with the lexicalization constraint as applied to the preference-content described above. This content was constitutive of Stalin's self, so it was not negotiable to even him, despite his being an inveterate liar and rationalizer. In case the reader worries that my summary is just a post-hoc rationalization of *all* Stalin's actual behavior – thereby trivializing the 'self' construct in application by identifying it with the totality of a person's behavior – note that it doesn't capture all of his actual preferences, including some important ones. Stalin was a sadist, deriving much enjoyment from the pain and humiliation of others. But he didn't moralize this preference; he inflicted gratuitous suffering only when he didn't anticipate paying a political price for it, and when he counseled others to engage in torture he did so on narrowly instrumental grounds. Indeed, he consistently tried to disguise his sadism as an expression of his favorite 'virtue,' hardness.

Stalin was spectacularly successful in causing a substantial part of the world, for a brief time, to approximate his ideal. Foreign communists assisted to power by him implemented it with surprising fidelity for a few decades, and the hard, rough, reliably motivated communist functionary was common enough for awhile to become a literary type. But, despite Stalin's prodigious efforts and almost perfect historical luck, the goal unraveled in farce within a short span after his death. Nothing did more to discredit 'scientific', industrial, engineering-oriented socialism – which had been a flourishing moralized political conception for a century – than Stalin's relentless attempt to force it into universal practice. One would like to hope that his project was bound to fail because it was ultimately incompatible with our species' natural sentiments (on a Humean-Smithian understanding of these). This is possible – though Rorty (1989) explicitly addresses himself to, and rejects, the claim – but I am doubtful that the necessarily empirical argument for it can be justified on the basis of the facts available to us. Enough communists, of enough different character types, were able to live coherent Stalinist lives to suggest that nothing in human moral nature made an industrial Bolshevik order, lasting for at least an historically significant amount of time, unachievable. But Stalin's methods made it impossible because they exhausted its participants and its victims alike and allowed them insufficient security to rationally plan for the medium run and to attempt *limited* experiments that could test better and worse ways of organizing the project. We may not know where or if

Humean sentiments set limits on the extent to which whole cultures can live by continuous mass murder. However, induction on the fates of industrial totalitarian socialist experiments in the past century does suggest binding economic limits. Stalin's project, like Mao's after him, required everyone to work intensely hard for very low expected personal returns, and people learned that marginal energy was more usefully deployed on subverting the state's capacity to aim its incentivizing knout in the right place than on producing things (Olson 2000). Stalin's moralized preference for demonstrative brutality and the refusal to acknowledge any role for rest subverted, and was bound to subvert, his ends. But because the preference was moralized he could not compromise with facts, could not, in the end, be practical.

My crucial claim here is that the argument above should have been relevant to a practically rational Stalin. As I stressed earlier, my contention doesn't require the stronger, and implausible, contention that it would actually have diverted Stalin given his general psychology. Skepticism about the extra-rhetorical force of moralization is defeated just in case there *is* a reason that can show the bearer of the ring of Gyges that some preferences cannot be consistently moralized. That argument, as Joyce's relaxed Humeanism, acknowledges, does not require that the reason in question be motivationally effective in the actual case. We need only get Stalin worried (perhaps annoyed *by* the philosopher's argument into a special dissonance-reducing murder of the philosopher) by an argument that appeals to the moralized character of some of his preferences.

It might be doubted that *moralization* of some of Stalin's preferences, even if it is agreed that they were moralized, really plays the central role in the above analysis that the argument requires. Why not just say that Stalin was too stubborn about some things for his own good? Of course we *can* say that, truthfully. But its meaning, in the absence of the above analysis, is unclear in two respects. First, what does the 'for his own good' refer to? That Stalin's project failed didn't harm Stalin, who died securely in power at the height of the project's apparent success, in any selfish sense. *If* we choose to say – as we surely at least reasonably might – that Stalin failed, this makes sense *because* we know he wanted, in a relatively impersonal way, the worldwide triumph of his preferred sort of social order, and we know (as he didn't) that it didn't happen. Second, what does 'stubbornness' refer to in this instance? What we have in Stalin's behaviour is not just a pique-ish unwillingness to admit mistakes during and after particular disputes; the case is that of a lifelong refusal to compromise as a result of a conviction about how the world – and not merely the circumscribed part of the world impacting the agent's narrow well-being – ought to be. This just *is* moralization of preference.

I will now draw the general point from the example.

Believers in the categorical force of morality – so, as Joyce argues, most people – make sense of such force by imagining that morality consists *directly* in commands issued to individuals from somewhere. Since there is no possible such somewhere, Joyce concludes that there are no rationally identifiable cross-cultural constraints on moralization.[16] In light of this, it must be emphasized that, keeping the popular image

[16] The logical dialectic between me and Joyce is quite complicated by this point in the discussion. We both agree that the case for error theory, which we both accept, doesn't depend on this premise – see note 3 above. My only – relatively tangential – point of contestation with Joyce is that I think

in play metaphorically, individuals don't get their moral commands independently of one another. People are social animals. Social animals must solve coordination problems. Certain sorts of preference structures make all solutions to such problems unstable. From this there arise *natural* constraints on what sorts of goals and actions can be rational objects of moralization. These constraints open the logical space for *objectively bad choices* (in the prudential sense) of preferences for moralization. If only some goals and actions are rational objects of moralization then it may be that some people – e.g., Stalin – have moralized *the wrong things*. In that case there is at least in principle a point to trying to persuade them of this.

For Stalin, loosening his grip would have been a very loud policy indeed, so much so that, given his other properties, no one appears to have advocated it to him during his last nineteen years in power. Fortunately, in liberal democracies one need not typically have much courage to advocate loud policies. It is one thing to recognize that such advocacy is unlikely to be politically effective, on any timescale, if it does not involve moralization. But this might be taken to reflect only the limited importance of rationality in real (as opposed to academic) political discourse (Teson and Picione 2006). However, I have here argued that there are non-cynical grounds for moralizing preferences for loud policies. Such moralization signals that the advocate partly defines herself by reference to her preferences for these policies. Once she has publicly moralized the preferences, concern for her reputation implies some level of commitment: there is then a sharply limited set of considerations that she can allow to influence non-incremental shifts of position, lest she be thought an unreliable social participant.

In addition, by moralizing a policy preference, and having this moralization understood by at least some others, one establishes, at least among those who pick up the signal, a social fact: *that* the policy preference in question is a potentially suitable object for moralization. This gives yet more people reason to at least imagine emulating the advocate's pattern of action – for that, not mere verbal declaration, is what preference-moralization consists in – and adding this story-line to their own self-narratives. Such emulation is an essential aspect of processes such as (for example) the one that took racial segregation in twentieth-century America from standard civil practice to something people are obligated to appear to regard as unacceptable. Many Americans probably first took the issue seriously when they discovered in early adult social life that failure to moralize a policy preference for racial integration limited their potential gains from social trade in one sub-community or another. By the point where this was *common* experience, the policy in question was no longer a loud one. But the process by which the policy moved from loud to quiet was through spreading diffusion of moralized preference for it from smaller to larger sub-communities; the policy was quiet on college campuses long before it was quiet in middle-class boroughs of Southern cities. Now the policy is loud only in tiny fringe communities.

I am not imagining that this progress was directly achieved by the reiteration of arguments; mass human preferences evolve largely though imitation. Rather, my point is that the line between moving social opinion by argument and moving it by

'fictionalism' would be a misleading label for the positive view I defend. The key to the subtle difference between us may lie in the fact that according to me moralization is a kind of *behaviour*, not the endorsement of some or other propositions. Thus it's not evident to me what *fictions* the sincere but philosophically sophisticated moralizer should be said to pretend to believe in.

encouraging imitation of certain moralized preferences is not sharp. Not only can one coherently argue for a preference and additionally moralize the preference in question, but one can (soundly) *argue for moralizing* the preference in question. Arguments of this sort need not depend on a metaphysical belief in moral facts. Advocates of loud policies need not be fundamentalists or Kantians in denial.

And so I conclude: it is outrageous that the world is simultaneously full of poverty and full of policy ideas that could greatly reduce it while requiring little or no sacrifice of expected utility by others. Conservative rent seekers (found in corporations, labour unions and political parties the world over) tirelessly resist such policies, often in overt bad faith. Though the universe is not offended by this behaviour, people should nevertheless denounce it by calling it immoral.[17]

References

Ainslie, G. (1992). *Picoeconomics*. Cambridge: Cambridge University Press.

Ainslie, G. (2001). *Breakdown of Will.* Cambridge: Cambridge University Press.

Binmore, K. (1994). *Game Theory and the Social Contract, Volume One: Playing Fair.* Cambridge, MA: MIT Press.

Binmore, K. (1998). *Game Theory and the Social Contract, Volume Two: Just Playing*. Cambridge, MA: MIT Press.

Binmore, K. (2005). *Natural Justice.* Oxford: Oxford University Press.

Bruner, J. (1992). *Acts of Meaning.* Cambridge, MA: Harvard University Press.

Byrne, R., and Whiten, A., eds. (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans.* Oxford: Oxford University Press.

Campbell, R. (1996). Can biology make ethics objective? *Biology and Philosophy* 11: 21-31.

Casebeer, W. (2003). *Natural Ethical Facts.* Cambridge, MA: MIT Press.

Danielson, P. (1992). *Artificial Morality.* London: Routledge.

Dennett, D. (1988). *Elbow Room*. Cambridge, MA: MIT Press.

Dennett, D. (1991). *Consciousness Explained*. Boston: Little Brown.

Dennett, D. (1995). *Darwin's Dangerous Idea.* New York: Simon and Schuster.

[17] I would like to thank Richard Joyce, Alex Rosenberg and an audience at the Max Planck Institute for Economics in Jena, Germany, for their helpful comments on earlier drafts of this paper.

Fingarette, H. (1989). *Heavy Drinking.* Berkeley: University of California Press.

Gauthier, D. (1986). *Morals By Agreement.* Oxford: Oxford University Press.

Gintis, H. (2004). Towards the unity of the human behavioral sciences. *Politics, Philosophy and Economics* 3: 37-57.

Gintis, H. (2006). Behavioral ethics meets natural justice. *Politics, Philosophy and Economics* 5:

Herrnstein, R. (1997). *The Matching Law.* Cambridge, MA: Harvard University Press.

Hutto, D. (forthcoming 2007). *Folk Psychological Narratives: The Socio-Cultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Joyce, R. (2001). *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, R. (2006). *The Evolution of Morality.* Cambridge, MA: MIT Press.

Kalderon, M. (2005). *Moral Fictionalism.* Oxford: Oxford University Press.

McClennen, E. (1990). *Rationality and Dynamic Choice.* Cambridge: Cambridge University Press.

Nichols, S. (2004). *Sentimental Rules.* Oxford: Oxford University Press.

Nozick, R. (1993). *The Nature of Rationality.* Cambridge, MA: Harvard University Press.

Nucci, L. (1986). Children's conceptions of morality, social conventions and religious prescription. In C. Harding, ed., *Moral Dilemmas.* Chicago: Precedent.

Olson, M. (2000). *Power and Prosperity.* New York: Basic Books.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Richards, R. (1986). A defense of evolutionary ethics. *Biology and Philosophy* 1: 265-293.

Rorty, R. (1989). `Orwell on Cruelty.' In Rorty, *Contingency, Irony and Solidarity*. Cambridge: Cambridge University Press, 169-188.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation.* Cambridge, MA: MIT Press.

Ross, D. (2006a). Evolutionary game theory and the normative theory of institutional design: Binmore and behavioral economics. *Politics, Philosophy and Economics* 5: 51-79.

Ross, D. (2006b). The economic and evolutionary basis of selves. *Cognitive Systems Research* 7: 246-258.

Ross, D. (forthcoming 2007). *H. sapiens* as ecologically special: what does language contribute? *Language Sciences.*

Seabright, P. (2004). *The Company of Strangers.* Princeton: Princeton University Press.

Seabright, P. (2006). The evolution of fairness norms: An essay on Ken Binmore's *Natural Justice*. *Politics, Philosophy, and Economics* 5:.

Teson, F., and Pincione, G. (2006). *Rational Choice and Democratic Deliberation: A Theory of Discourse Failure.* Cambridge: Cambridge University Press.