Journal of Contextual Economics 140 (2020), 3–30
Duncker & Humblot, Berlin

Inside and Outside Perspectives on Institutions: An Economic Theory of the Noble Lie

By Cameron Harwick*

Abstract

If there exist no incentive or selective mechanisms that make cooperation in large groups incentive-compatible under realistic circumstances, functional social institutions will require subjective preferences to diverge from objective payoffs – a "noble lie." This implies the existence of irreducible and irreconcilable "inside" and "outside" perspectives on social institutions; that is, between foundationalist and functionalist approaches, both of which have a long pedigree in political economy. The conflict between the two, and the inability in practice to dispense with either, has a number of surprising implications for human organizations, including the impossibility of algorithmic governance, the necessity of discretionary rule enforcement in the breach, and the difficulty of an ethical economics of institutions.

JEL Codes: C73, A13

Keywords: Game Theory, Cooperation, Institutions

Leeson and Suarez argue that "some superstitions, and perhaps many, support self-governing arrangements. The relationship between such scientifically false beliefs and private institutions is symbiotic and socially productive" (2015, 48). This paper stakes out a stronger claim: that something like superstition is *essential* for *any* governance arrangement, self- or otherwise.

Specifically, we argue that human social structure both requires and maintains a systematic divergence between subjective preferences and objective payoffs, in a way that usually (though in principle does not necessarily) entails "scientifically false beliefs" for at least a subset of agents. We will refer to the basis of such preferences from the perspective of those holding them as an "inside perspective," as opposed to a functionalist-evolutionary explanation of their existence, which we will call an "outside perspective." Drawing on the theory of cooperation, we then show that the two perspectives are in principle irreconcilable, discussing some implications of that fact for political economy and the prospects of social organization.

* Department of Accounting, Economics, and Finance, SUNY Brockport, 350 New Campus Drive, Brockport, NY 14420, United States. The author can be reached at charwick@brockport.edu.

1. Inside and Outside Perspectives: The Example of Ordeals

For reasons that will be clearer in subsequent sections, the distinction between the inside and outside perspectives can be seen more clearly with some distance from the institution in question. Before proceeding to more familiar examples, therefore, we will use as our initial paradigmatic example something utterly foreign to most contemporary people's experience: trial by ordeal.

In both medieval Europe (Leeson 2012) and contemporary Liberia (Leeson and Coyne 2012), belief that God punishes the guilty is operationalized in criminal justice rituals – ordeals – designed to discover the judgment of God in ambiguous cases. In Europe, the accused plunged his hand into boiling water. If God protected him from scalding, it was a sign of innocence; otherwise he was guilty. Similarly, in Liberia, the accused imbibes the brew of a toxic bark. If the spirits cause him to vomit it up, he is innocent; otherwise his resulting illness indicates guilt.

Such is the inside perspective on ordeals, something like the account one would get from someone living in that society. From the vantage point of modern criminal justice on the other hand, which does not share the basic presuppositions of divine justice, such practices seem backward and arbitrary. But Leeson argues that such procedures are feasible second-bests in the absence of modern standards of evidence and the state capacity to act on them. Specifically, trial by ordeal results in a *separating equilibrium* whereby the innocent willingly undergo the ordeal and the guilty refuse, meaning that willingness to undergo the ordeal conveys valuable information on the innocence of the accused.[1] He then shows that a large proportion of ordeals were in fact successfully passed, establishing innocence, and implicating the administrators (priests) in consciously or unconsciously adjusting the severity of the ordeal conditional upon the accused's willingness to undergo it.

This is an *outside* perspective on the institutions of ordeals, analyzing its functionality from some distance and without committing to the society's presuppositions. The two perspectives are fundamentally incompatible, not only in the basic sense of their being alternative explanations of the same phenomena, but in the deeper sense that introducing an outside perspective to medieval Europe or contemporary Liberia would destroy the *possibility* of trial by ordeal. If criminals were to "see through" from the outside, their assent to the ordeal would carry no informational value, and the institution would be nonoperational. And indeed, Leeson shows that known nonbelievers were not subject to ordeals.

[1] Superstition also supports separating equilibria via a similar mechanism in Iannaccone (1992) and Leeson (2013a), where visible commitment to a burdensome superstition filters out noncooperators. These examples also result in separate inside and outside perspectives, and the remarks below on costless signaling will also apply.

Leeson demonstrates the stability of such an institution *provided* people are believers.[2] What is missing, however, is a *strategic* model of belief; an explanation of *why* susceptibility to manipulation would be a viable phenotype in the first place. This is not to say that one necessarily has a choice in one's beliefs, but rather that credulity will erode in a population if unbelief results in persistently higher payoffs, eventually leaving an ordeal system nonfunctional.[3] In such an environment believers will be outcompeted by nonbelievers – and particularly by nonbelievers who profess belief – regardless of their updating strategy. As is shown below, other institutional features intended to reverse the advantage (say, burning heretics) displace the unexplained element elsewhere, but ultimately do not account for it.

The following section shows briefly the inability of selfish and rational agents – the *homines æconomici* of neoclassical theory – to cooperate in large groups. Later sections then show where in such a model the inside/outside perspective distinction arises, how agents who *do* make this distinction can outcompete *homines æconomici* who do not, and some implications of the distinction for social science.

2. There Is No Incentive-Compatible Social Organization

More generally, a signaling game with a costless signal (or where the costs are immaterial) reduces to a social dilemma, a class of games that includes prisoner's dilemmas (two-person) and public goods, commons, and collective action problems (*N*-person). Social dilemmas are non-zero-sum games characterized by mutual gains from cooperation, but also a Nash equilibrium of mutual defection. In other words, the Pareto-optimum is not a Nash equilibrium. Social behavior is defined by cooperation in such games, against one's own narrow interests.[4]

Unfortunately much of the literature on cooperation and governance generalizes from two-person games to *N*-person games, and therefore concludes that repeated play

[2] This assumption is not quite taken for granted: Leeson (2013b) for example builds a Bayesian model of belief, so believers are not totally credulous in the face of crass manipulation, deriving an equilibrium quantity of manipulation.

[3] In this sense we are considering belief not as a Nash equilibrium, but as an evolutionarily stable strategy. Similar considerations also militate against the theory that language evolved for the purpose of manipulation or deception (e. g. Dawkins and Krebs 1978). It must be incentive-compatible not only to *send* a signal, but also to *receive* and act upon a signal (Fitch and Hauser 2002; Searcy and Nowicki 2005, 8). Knight (1998) takes these considerations and comes to a similar conclusion to this paper, with trust in the veracity of language vouchsafed by the costly rituals implicit in a normative community.

[4] *Sociality* in this sense is distinct from *gregariousness* (e. g. herding behavior), which is incentive-compatible under certain conditions. Sociality generally depends on a favorable mix of coordination games and social dilemmas in the environment of the cooperating group (Bear *et al.* 2017), but – as this section shows – the dilemma aspect is irreducible. Because coordination games have stable cooperative equilibria, we leave those to the side and focus on social dilemmas as the more difficult impediment to social behavior.

is sufficient to establish incentive-compatible governance structures even in the latter.[5] This section shows that this inference is not warranted. We use for our paradigmatic game, therefore, a public goods game (or, equivalently, a collective action problem) rather than a prisoner's dilemma. Governance, broadly speaking, consists in collective action of some sort or another. If rational and self-interested agents cannot cooperate in a public goods game, then governance and society more broadly will also be impossible for them.

The lack of a cooperative equilibrium, we argue, necessitates distinct "inside" and "outside" perspectives: that is, a divergence between the objective game structure, a social dilemma with an equilibrium of universal defection, and – for at least a subset of agents – a different reckoning of subjective costs, which transforms the game into one with a cooperative equilibrium.

2.1 Social Behavior Poses a Problem

The basic difficulty with sustaining social behavior can be seen in a one-shot public goods game, a limitation we will relax shortly. Suppose there are N agents, each with an endowment of 1. Each agent has the choice of contributing $c \in [0,1]$ to a communal pot, in which case γc (with $\gamma > 1$) is distributed equally to all agents. Agent i's payoff function, then, is

$$(1)\ p_i = 1 - c_i + \sum_{n=1}^{N} \frac{\gamma c_n}{N}$$

whereas the total payoff function, summing p_i over all i, simplifies to

$$(2)\ P = \sum_{n=1}^{N} (1 + (\gamma - 1)c_n)$$

The total payout is maximized at $c_n = 1$ for all n, but – provided c_n is independent of c_i for all $n \neq i$ – individual payoff is maximized at $c_i = 0$ so long as $\gamma < N$. There is a divergence between the private cost of non-contribution ($\partial p_i / \partial c_i = \gamma/N - 1$) and the social cost of non-contribution ($\partial P / \partial c_i = \gamma - 1$).

In any game of this structure, defection is the dominant strategy for rational and self-interested agents. The public good is not provided; the signal has no informational

[5] E. g. the classic simulation in Axelrod (1984) which showed the dominance of tit-for-tat when paired against other strategies for some number of periods. Hardin (1985) criticizes his generalization to N-person games. Kandori (1992), similarly, shows that cooperation-sustaining strategies exist for repeated pairwise games where the pairings are sampled randomly from a population, but not for non-pairwise interactions. Alger and Weibull (2013) examine the divergence between preferences and payoffs in a similar spirit to the present paper, but only for pairwise interactions with positive assortativity.

value; the commons is depleted; collective action is not undertaken; society does not get off the ground.[6] Ordeals may have been an important supporting institution in the medieval criminal justice regime, but they were not, by themselves, sufficient to establish peaceful society.

2.2 Repeated Play Is Not a Solution

It is well known that repeated play can sustain cooperation in two-person prisoner's dilemmas, provided the end of the game is not known, by allowing players to *punish* defectors. By playing a trigger strategy, for example, where one player responds to defection by defecting in all future games, one player can threaten the other with the loss of all future gains from cooperation. For narrower but still plausible ranges of discount rates, more forgiving strategies such as tit-for-tat (where the retaliation lasts only for a single subsequent period), or even tit-for-double-tat (where one period of retaliation is triggered only after two defections) can be cooperation-supporting equilibrium strategies (Axelrod 1984), especially where mistakes are made.

In a repeated N-person dilemma however, punishment is diffused over all agents, unlike the dyadic game that makes it possible to punish the defector and *only* the defector. Consider an infinitely repeated version of (1), with the simplifying modification that $c_i \in \{0,1\}$, meaning the agent has a binary choice of whether or not to contribute. A trigger strategy is not robust to the commission of any mistakes. If we introduce a parameter $\varepsilon \in (0,0.5)$ for the probability that an agent mistakenly fails to contribute where he meant to or vice versa,[7] a more forgiving strategy will be necessary if a single mistake is not to snowball into universal defection.

The more forgiving the strategy, however, the lower the difference in payoffs between cooperation and defection, and therefore the lower threshold discount rate necessary in order for cooperation to be feasible. Consider the N-person analogue to a tit-for-tat strategy, "contribute unless some fraction μ of the population failed to contribute in the last period." As N becomes arbitrarily large, any individual's choice of strategy matters increasingly less for the payoffs of his peers, making it increasingly difficult for him to punish defection and, in turn, for others to punish him. The strategy can therefore be invaded by a "never contribute" strategy. As Bowles and Gintis (2011, 63–7) show using a simulation, and Fehr and Gächter (2000) confirm in the

[6] For the signaling game in particular, if c_i is an unobservable or imperfectly observable cost that one may bear for the benefit of the group (say, refraining from crime), and the cost of signaling compliance is known to be zero (say, enthusiastically assenting to undergo an ordeal), then the signal's value as an indicator of c_i will be a commons which free riders will be motivated to deplete by falsifying the signal.

[7] Or, equivalently, the probability that any agent assesses another agent to have failed to contribute when he in fact did not, or vice versa.

lab, contribution to a public good drops off precipitously as N rises beyond about 5, even for very low error rates (0 – 0.02) and discount rates (0 – 0.04).[8]

2.3 Targeted Punishment Defers, But Does Not Solve, the Problem

The public goods game as set out so far is somewhat more limited than real-world public goods games. In particular, it is overly restrictive to assume that the only margin of choice is contribution or noncontribution. Real-world social behavior operates on many different margins, and choices along one can influence cooperation in another, for better or for worse (Reiter *et al*. 2018).

Suppose now that individual i can pay some cost v_{ij} to punish non-contributor j by subtracting v_{ij} from j's payoffs that period (the total product therefore falls by $2v_{ij}$). If $\sum_I v_{ij} > 1-\gamma/N$, there is no divergence between private and social cost for j, and j's dominant strategy is to contribute.

Things look different from the perspective of agent i, however. An agent facing the choice of whether to punish j faces the cost v_{ij}, but – because the good to which j failed to contribute is public – the benefits of j's future cooperation accrue to all agents. In other words, the "punish non-contributors" game is simply another public goods game superimposed upon the first (Yamagishi 1986). Even in a repeated game, it will be in i's interest to free-ride on the punishment of j.

Targeted enforcement of contribution brings us into strategic territory that begins to look like governance. And immediately we run into the fundamental problem of governance: who watches the watchmen? – which in our case we can reformulate as: who punishes the punishers who fail to punish? Second-order punishment is beset by the same problem on another level, along with third- and fourth- order punishment and so on.

We have, therefore, an unbridgeable gulf between Pareto optimality and Nash equilibrium in large groups, provided there is no independent authority, external to the system, to appeal to. And whether or not such an authority can be said to exist with respect to any particular subgroup, it is always true that for society as a whole, all authority must be regarded as endogenous to the social system. We will call this problem the *incentive gap:* the impossibility in broad classes of social dilemmas of eliminating the incentive to defect among some subset of agents, whose defection would eventually lead to the total unraveling of cooperation.

[8] The same argument applies to inclusive fitness explanations for cooperation, i. e. that altruistic genes can proliferate on the basis of kin selection. Relatedness enters into the structure of payoffs from a gene's perspective in exactly the same way as the probability of a repeated interaction, which is to say that the relatedness coefficient within human groups must be implausibly high in order for kin selection to support generalized altruistic behavior (Bowles and Gintis 2011, 60).

2.4 The Incentive Gap in the Firm

The problem of incentive alignment has been studied most systematically in the theory of the firm, where incentive mechanisms and organizational relationships are most explicit. If we regard the firm as a locus of joint production (Alchian and Demsetz 1972),[9] effort – to the extent that it is imperfectly monitorable – becomes a public good. The question then is: how can production be organized so that no agent has an incentive to shirk?

The conventional wisdom, per Alchian and Demsetz, holds that entrepreneurs provide monitoring services, and their own incentive to avoid shirking is ensured by their status as residual claimants. Holmström (1982), however, proved that there exists no structure of incentives that can motivate employees to avoid shirking in joint production so long as the budget is balanced. This is the problem of imperfect monitoring in Section 2.2: where effort is partially unmonitorable, any incentive system will face a tradeoff between capricious punitiveness and allowing scope for profitable defection (i. e., between Type I and Type II errors). Eswaran and Kotwal (1984) then showed that even incentive schemes which aligned incentives for employees by failing to balance the budget (i. e. enforcing effort via the threat of paying out less than the entire product) create perverse incentives for the residual claimant. This is the problem of second-order punishment in Section 2.3. They conclude that "the crucial necessity of monitoring the monitor is thus not met. […] the problem of moral hazard [i. e. defection in the firm's social dilemma] takes a different form but remains unsolved."

Relationships between firms are similarly fraught. The inability to write complete contracts, for example, is a common assumption in organizational economics, which is simply to say that the variety of choice margins open to two transactors precludes the ability to ensure incentive compatibility ex post, *even assuming perfect enforcement of written contracts.* As Alchian and Demsetz note, "it is hard to imagine any contract, which, when taken solely in terms of its stipulations, could not be evaded by one of the parties" (1972, 778). Trust is commonly taken as an exogenous feature of functional social systems, but it is underappreciated that the problem of trust consists in the fact that trust*worthiness* is not individually rational. Without special assumptions, the capitalized value of reputation is necessarily less than the value of a one-time defection in intertemporal markets (such as credit or money issue), as the value of defection rises in proportion with the value of reputation (Harwick and Caton 2020; Bulow and Rogoff 1989; Taub 1985). Where this is the case, to the extent that information is not public and reliable, reputation will be insufficient to ensure cooperation *even for dyadic interactions* if they are drawn uniformly from a larger population (e. g. in Kandori 1992).

[9] The importance of joint production is that it forecloses the possibility of paying by marginal product, as product exhaustion will not hold where each agent's marginal product is not independent of the effort of other agents. In this situation, compensation on the basis of inputs (i. e. effort) can be more feasible than compensation on the basis of value added.

2.5 The Incentive Gap in Society

The class of social dilemmas is vast, particularly in the context of governance. Besides the number of people involved, games can vary on the imperfection of information and monitoring, agents may have a wider or narrower range of choices in contribution or punishment, games may be conditioned upon the results of other games, and so on.

Depending on these various factors, many of the dilemmas encountered by people in a society can be adequately solved by repeated play, especially if interactions are dyadic. Indeed, many social institutions – most importantly, property rights and market exchange – have the function of transforming would-be *N*-person social dilemmas into soluble dyadic interactions. Nevertheless, the enforcement and/or voluntary respect of the rules constituting these transformative institutions are themselves irreducibly public goods. Despite the importance in the developed world and (especially) in economic theory of opportunities for dyadic exchange, the very existence of a market – and, for that matter, of a state – rests on the provision of a number of genuinely public goods on both micro- and macroeconomic levels. Similarly to the second-order punishment problem, even if we suppose that the provision of property rights could in turn be transformed into a dyadic game through some supervening institution, the establishment and maintenance of *that* institution would be a public good.

The open-endedness of human strategies can also be an impediment to cooperation and commitment (Stewart *et al*. 2016), analogous to the problem of incomplete contracts in organizational economics. This problem inhibits the establishment of both potential exchange relationships (Harwick 2018) and governance solutions.[10] The much-celebrated fact in economics that incentive-compatible Pareto-optimal resource allocations exist *given* well-defined property rights, complete contracts, and limited behavioral repertoire, should not blind us to the gulf separating the Arrow-Debreu world of general equilibrium from the real world, where open-ended behavior makes complete contracts impossible and property rights costly to establish.

The upshot is that, for plausible rates of discount and error, there exists no potential structure or set of strategies to ensure that every member in a large group has an incentive to cooperate in the face of social dilemmas, a problem as true for society broadly (Bowles and Gintis 2011, chapters 4–5) as it is for a firm. To the extent that sanctions can render cooperation the dominant strategy, they do so by placing the enforcers – whether the entire population or some specialized subset – into a social dilemma of their own.

[10] Ostrom's (2005: 259) famous design principles for the management of common pool resources, particularly those relating to monitoring, sanctions, and punishment, presuppose altruistic preferences of some form or another. In this crucial respect, Ostromian agents depart from the standard *homines æconomici.* See the following section.

What then of the infinite variety of equilibrium strategies possible under the folk theorem? Bowles and Gintis argue that most of these equilibria, even for dyadic interactions, are "evolutionarily irrelevant" – that is, there is no reason to expect the folk theorem to be actually operational under conditions of imperfect information, and there is no feasible path for the emergence of such strategies from a starting point of noncooperation. A viable strategy must be robust to error and be able to outcompete noncooperators under a wide variety of unfavorable situations. In other words, relevant cooperative strategies must – in addition to being Nash equilibria – be *evolutionarily stable.* This requirement binds even more tightly for *N*-person games. "Knife-edge" equilibria, trigger strategies, and other strategies that do not meet these criteria shed no light on the strategies actually employed by humans.

3. Self-Deception and Cooperation

The previous section has been for the most part deliberately ambiguous on the question of interpretation. There are two main ways the payoffs can be interpreted:

1. In classical game theory, payoffs are understood in terms of utility, or some proxy for it. Strategy selection is the result of a conscious and rational utility-maximizing choice.
2. In evolutionary game theory, payoffs are understood in terms of reproductive fitness, or some proxy for it. Choice is not necessary, and strategy selection results from the ability of fitter strategies to displace less fit strategies.

Some of the examples given thus far are interpreted more naturally under one or the other, and the tension between the two particular cases has been a matter of some controversy in the social sciences (Sugden 2001; Grüne-Yanoff 2011).[11] But so far, both are valid for the previous section: regardless of whether strategies are selected or chosen, if social life is structured as in section 2 – and it is generally assumed to be, at least on some critical margins – we should not observe cooperation.

This agreement between the two perspectives has sometimes been taken as an evolutionary foundation for the selfish and rational *homo œconomicus* (e. g. Binmore

[11] Cultural-evolutionary learning models, where the replicator is understood to be the strategy rather than the individual, represent a sort of hybrid between these two, with individual utility acting as the selector as in classical game theory, but with the replicator dynamics of evolutionary game theory. Such models have proven useful in some respects, particularly in their ability to explain features from section 2 that either "pure" interpretation alone could not. Nevertheless, such models pose severe interpretive difficulties. I do not address cultural-evolutionary models explicitly here, but treating classical and evolutionary game theory as analytical coequals in a group selection model, as this paper tries to do, can be understood as a way of resolving the difficulties of a hybrid model. The problem here, as below in section 3.2, is to explain: why would humans be such good vehicles for certain types of individual-fitness-reducing strategies in the first place?

1994). Any agent making conscious decisions on the basis of preferences must prefer and choose those things that maximize its objective fitness, for agents with such preferences will be the ones that reproduce and pass on their predilections. If this is the case, there is no issue in conflating objective payoffs and subjective utility, at least in equilibrium. For individuals *qua* individuals, cooperative strategies are ineluctably maladaptive.

Homo œconomicus' inability to cooperate in theory, however, stands in sharp contrast to the empirical Great Fact that functional firms and high-trust societies *do in fact exist.* If the incentive gap cannot be reconciled to this Great Fact either in terms of subjective utilities or objective fitness on their own, or with the former reduced to the latter, it will be necessary to show what kind of *non*-maximizing preferences might be selected for, and how. This section advances *self-deception* as just such an alternative: a divergence between subjective and objective payoffs, out of which arises the divergence between the inside and outside perspectives.

3.1 The Phylogeny of Self-Deception

One important difference between the stylized public goods game of the previous section and the actual social world is that the latter is characterized by a *group structure.* Within groups, assortativity can ensure cooperators reap enough of the benefits of cooperation to outcompete noncooperators (Alger and Weibull 2013; Bergstrom 2003). And competition between groups presents itself to members as a coordination game, the gains from which can outweigh the losses from cooperating with fellow group members in social dilemmas. In other words, as Sober and Wilson (1998) argue, selection for cooperative groups can more than balance the within-group selection against cooperative individuals.

This difference, however, entails a mismatch between the level of selection from which a particular behavior arises, and the level at which strategies are employed. For humans, strategies are employed by *individuals*, not by groups as such, and selection on individuals necessarily favors noncooperation. If the individual's cooperative predilections arise from selection at another level, and if we regard him as making conscious decisions on the basis of his interests, *he must deceive himself* regarding those interests if he is to live in a society capable of undertaking collective action.[12] In other words, selection operating under these circumstances will cause *subjective preferences to systematically diverge from the objective payoffs.*[13]

[12] This argument would apply to *any* mismatch between the level of selection and the locus of decision-making, including inclusive fitness explanations (see above, footnote 8), where altruism is selected for at the gene level. Thus eusocial insects are excluded, not because their altruism arose from a different selective process, but only because they do not make deliberate and conscious decisions.

[13] Bear, Kagan, and Rand (2017) show that deliberation (the process of rationally assessing one's interests) leads to lower levels of cooperation, and that cooperative strategies are never-

Note that both assortativity and group competition require mechanisms to enforce the assortativity and the group structure, mechanisms whose provision – like the supervening institutions of the previous section – is itself a public good. They do *not,* therefore, make cooperation individually rational, at least not for everyone; rather, they make irrationality (from the individual's perspective) viable. In either case, individuals must at minimum find intrinsic utility in punishing or excluding non-cooperators in order to solve the second-order punishment problem. If a subpopulation employing such a strategy manages to achieve a mass sufficient to impose its preferences on the remaining selfish rational maximizers, false beliefs motivating cooperation can even be self-confirming in the sense of Fudenberg and Levine (1993) so long as the threat remains credible – it really *will* be in the interest of most people to cooperate. Punishers, for their part, will have to do minimal punishing on the equilibrium path of play. In this way, cooperation can be stabilized and the incentive gap closed – both in the firm (Miller 1992, chapters 10 – 11)[14] and in society (Bowles and Gintis 2004; 2011, chapter 9) – allowing larger-scale collective action to get off the ground without depending on highly or uniformly altruistic preferences.

Furthermore, experimental evidence shows unequivocally that such a divergence between payoffs and preferences does in fact exist: humans are, on many margins, genuinely altruistic and pre-rationally inclined to cooperate against their own narrow interests (Bowles and Gintis 2011, chapter 1; Tomasello 2009).

In this light, the inside and the outside perspectives correspond, respectively, to looking at an institution from the perspective of its members' subjective preferences and beliefs, and of the objective payoffs. The distinction is sufficiently general, however, that it bears on any dynamic system where (1) relative frequencies of strategies are governed by some sort of selection dynamic, and (2) influences on the strategies employed by decision-making agents are selected at levels other than the agents themselves. As in the previous section, both biological and market competition fit the bill. Regardless of what the objective payoffs consists in – whether biological fitness, as in the sociobiological literature, money, as in economics, or even utility, to the extent that utility functions are deterministic in their inputs (see below, section 4.1) – collective action with imperfect information requires that subjective preferences diverge from them.

The inside-outside perspective distinction is not identical with the fact-value distinction, but the latter does follow straightforwardly from the former. If social organization necessarily relies upon individually maladaptive altruistic preferences in the breach, and if the function of human morality is to coordinate cooperative strategies (Curry 2016; Curry *et al*. 2019), then it will be impossible to derive a

theless pervasive, but generally non-deliberative. Alger *et al.* (2020) offer a formal model showing that exactly such a divergence can sustain social behavior.

[14] Miller is concerned here to show that the incentive gap in the firm can be closed by a non-maximizing "company culture" which allows credible commitments. This constitutes the divergence necessary to approximate Pareto-optimality in joint production.

morality that sustains human society from the nature of things (i. e. from the objective payoffs). To accept the broad and universal features of human moral life is *ipso facto* to deny the ability to derive normative force from the objective payoffs. Facts and values are related, of course – where else would morality come from if not the nature of things? – but the relationship cannot be a deductive one.

3.2 On Subjectivism

In contrast to the more naïve assumption that the incentive gap and the Great Fact can be reconciled straightforwardly, a number of strands of literature take an alternative route to obviate rather than to solve the problem. These strands together can be called 'subjectivist,' and can be divided into rational subjectivism and empirical subjectivism.

The rational subjectivist in the Chicago tradition sees no reason why subjective preferences ought to correspond to objective payoffs in the first place – after all, *de gustibus non est disputandum.* For the subjectivist, *homo œconomicus* is not incompatible with cooperation if we model him with a preference for altruism, a preference which must simply be taken as given.

Even the thoroughgoing subjectivist, however, has to assume *some* correspondence between preferences and payoffs in order for the analysis to escape from tautological formalism into any empirical relevance at all. In orthodox analysis, this correspondence is ensured by treating preferences as preferences *over consumer goods.* If all goods have a positive income elasticity of demand, a strict preference for income – i.e. concordance of subjective utility with objective payoffs – can be derived by transferring Binmore's selection logic to the economic sphere: agents *without* a strict preference for income get outcompeted in the marketplace, and do not ultimately affect the conclusions of the model (Alchian 1950; Becker 1962). Thus the dilemma: either we deny the correspondence of objective payoffs (in this case, income) with subjective utility and sacrifice the empirical relevance of price theory, or we affirm the correspondence and are immediately led back to the untenable selfish and rational *homo œconomicus. De gustibus* is a fine analytical decision when considering consumer decisions in a reasonably competitive marketplace, but generalizes poorly to arenas where we must consider strategic decisions.

Besides the *de gustibus* strategy, a number of ostensible repudiations of the rational choice model should be understood as subjectivist in the sense of stipulating a utility or learning function as well: for example, modeling altruism or beliefs as consumption goods (e. g. Bénabou and Tirole 2011), or the literature on team reasoning (Sugden 2003; Gintis 2016) where shared intentionality impresses group goals onto individuals.

The important limitation of these, however, is the fact that – like Leeson's opening examples – all such models are essentially static. A static theory of human behavior

does not attempt to *explain how* the incentive gap can be reconciled to the Great Fact; it simply observes – often correctly – that actual human decision-making bears little resemblance to *homo œconomicus* in many contexts. To consider altruism as a consumption good alongside other objects of preference is a valid analytical decision where we are not concerned with changes in the relative frequency of preferences or strategies. But altruism is of theoretical interest precisely because of its dynamic effect on these relative frequencies. Even if a subjectivist model with altruism is more empirically accurate as a model of human decision-making than *homo œconomicus* in many circumstances, the virtue of the latter is that it serves as a benchmark for dynamic stability.

Subjectivism, therefore – whether in its rational or empirical variety – is not an alternative reconciliation; it simply does not ask the question we are interested in.

3.3 The Ontogeny of Self-Deception: Preference vs. Belief

The human capacity to deliberate is the capacity to explicitly justify behavior; that is, to ground strategy choice in terms of a more basic objective function. For an organism with this capacity, a divergence between payoffs and preferences must consist either in failure to perceive the lack of correspondence, or a deliberate decision to ignore the objective payoffs. The former corresponds to an inside perspective as a *belief* phenomenon; the latter to an inside perspective as a *preference* phenomenon.

The initial example of ordeals was a belief phenomenon. In this case, individuals must be convinced that it is really in their interest to employ a strategy which is in fact dominated by another – a "Noble Lie," so to speak. Preference phenomena, on the other hand – as in many empirically informed subjectivist models – require individuals to voluntarily pursue goals at odds with their objective payoffs. All human societies and organizations rely on some mix of the two, though the ubiquity of motivated reasoning (Bénabou and Tirole 2011), and the near functional equivalence of the two in closing the incentive gap, might suggest that the distinction between preferences and beliefs is not quite so sharp as economists would have it.

To classify altruistic preferences as "self-deception" along with false beliefs is not a claim about the psychology of altruism. Rather, it is to take strict correspondence, as simple Darwinian logic demands, as the benchmark of strategic rationality for individuals *qua* individuals. In both cases, "self-deception" points to the fact that cooperative strategies such as humans in fact employ systematically fail to maximize individual fitness, and that the individual has adopted the fitness of something else (whether the group as a whole or other individual members) as a terminal goal.

There are tradeoffs to closing the incentive gap using predominantly beliefs versus preferences. Though less reliant on deliberate self-sacrifice, which can be difficult to motivate, belief-based inside perspectives are not necessarily robust to outsider contact, for example: it is more difficult to maintain rich factual beliefs when con-

fronted with other functional cultures maintaining incompatible factual beliefs (see Leeson 2013a for an example). Ecumenical polytheism and evangelical monotheism were both institutional technologies to deal with this problem, either by creating ideological space to preserve local norms, or through homogenization.

In principle, a population that preferred cooperation sufficiently strongly for its own sake could dispense with noble lies entirely, provided they were willing to punish defection wherever it did arise.[15] Nevertheless, in practice, a preference for altruism can only withstand so much defection. Humans do make deliberative choices on the margin, and in experimental public goods games, even groups highly inclined to cooperate at first will quickly decay to negligible contributions (Ledyard 1995). For this reason, any nonauthoritarian society – that is, one where overt punishment can be kept to a reasonable minimum – must rely on some combination of false facts and maladaptive preferences among the masses to maintain the divergence between objective and subjective reckonings of costs. The more intrinsically altruistic will be able to get by with fewer factual commitments, and (therefore) with more abstract religions and ideologies. A richer belief system can satisfy both groups with a single body of doctrine: metaphysics and theology for cooperators; the wrath of God for would-be defectors. Indeed, vengeful deities appear in the archaeological record to be strongly linked with the rise of large-scale political organization (Norenzayan *et al.* 2016). Finally, for those who nevertheless expect gains or derive pleasure from defection, there's overt punishment – which, when effective, itself relies primarily on the altruism of the first group.

3.4 Normative Drift and the Invisibility of the Inside Perspective

Belief-based inside perspectives are a deliberative blind spot, almost by definition. If the criminal from section 1 were to form his beliefs in full view of the objective payoffs, he could exploit the value of the signal and increase his own payoffs. There is an unexploited arbitrage opportunity which he systematically overlooks. Similarly for preference-based inside perspectives, altruistic preferences as an ultimate fact cannot be argued about. If one prefers helping others over one's own convenience in full view of the cost, there is no convincing him otherwise except on the basis of an even more fundamental preference or value.

[15] This suggests a novel interpretation of the rise of scientific rationalism (in the Weberian sense) in the West as a transition from belief-based cooperation to preference-based cooperation. This interpretation is supported by the facts that (1) scientific rationalism has been accompanied from the beginning by persistent worries of social decay, (2) that decay has so far failed to materialize, at least in terms of organizational capacity, and (3) that people from Weberian-rationalist cultures do seem to have a stronger preference for altruism (specifically, they are far more generous in one-shot dictator and ultimatum games than those from more traditional cultures – see Henrich *et al.* 2010). I will not pursue this line of thought here, however.

Both of these situations make it difficult to criticize a culture's norms from within that culture – again, unless this is done from the vantage point of another shared norm. An effective inside perspective must appear self-evident; in other words it must, whether through beliefs (e. g. in a moralistic deity) or preferences (e. g. the self-evidence of the Golden Rule among post-Christian Westerners), make itself invisible and present itself as an ultimate fact. Inside perspectives which fail to provide for their own survival this way, quite simply, do not persist.

Thus, in the absence of outside contact insular societies are prone to *normative drift*, which is to say there are no internal or external forces tending to select for prosocial rather than antisocial norms: no internal forces because its inside perspective remains invisible to its own practitioners, and no external forces by hypothesis. Phylogenetically, Bowles and Gintis (2011, chapter 10) show that, in the absence of strong external pressure, fitness-reducing norms can hitchhike on a more general norm-internalization capacity. And ontogenetically, the invisibility of the inside perspective indicates how exactly the human deliberative capacity can fail to weed out pathological norms.

Antisocial punishment is a particularly significant manifestation of normative drift (Hermann *et al*. 2008), which we may think of generally as the direction of altruistic punishment against the emergence of Pareto-superior norms, distinguished from mere selfishness or retaliation. There are numerous examples, especially – though not exclusively – in more isolated societies: self-destructive food taboos, human sacrifice (Edgerton 1992), female genital mutilation, forced marriages (Bicchieri 2016), and so on. And indeed, antisocial punishment has been shown both theoretically and experimentally to be a viable strategy in public goods games (Nikiforakis 2008; Rand *et al*. 2010). In some instances, practitioners have been more than happy to abandon such norms when given the opportunity to coordinate around new ones.

The same logic holds for hegemonic societies as well as insular societies: both cases lack effective inter-societal competition to weed out pathological norms, beliefs, and practices. Thus, the fact that many apparently backward cultural practices are rationalizable from an outside perspective as in section 1, should not be taken as an argument for unqualified cultural conservatism or relativism. That coordination around self-deceptive beliefs and/or preferences is necessary *in general* does not imply that any particular complex of norms is in any sense Pareto-optimal, even among close and feasible alternatives, except – perhaps – under circumstances of especially intense intercultural competition.

4. Implications for Political Economy

4.1 A New Light on Rules versus Discretion

The conventional wisdom in economics is that predefined rules are preferable to administrative discretion, in that the latter often precludes credible commitments, and introduces a social dilemma that an enforceable rule could solve (Simons 1936; Kydland and Prescott 1977; Root 1989).

This logic holds to the extent that one can rely on external enforcement somewhere in the logical chain, whether from the state or from preexisting norms. In the case of monetary policy for example (as in Kydland and Prescott 1977), this can be a valid assumption, as the central bank is generally already embedded in an enforcement apparatus. Suppose therefore that a society, desiring to pre-commit itself to cooperation now and for eternity, delegates the enforcement of cooperation to a disinterested and omniscient robot. There is no necessity of punishing the robot, as it necessarily obeys its programming. Is this sufficient to close the incentive gap?

The problem is this: if the structure of social life is such that there always exists some opportunity for profitable (in terms of the objective payoffs) defection from a society's norms, then it follows that any conceivable set of pre-announced and rigidly followed rules will be vulnerable to exploitation, either by rule-makers or rule-followers. One might think of malicious compliance as an example. As Boyd and Lorberbaum (1987) show, no deterministic strategy is the best response to all other strategies, and therefore no such strategy can resist invasion in a social dilemma.[16] To protect itself from novel defection strategies, therefore, an organization or a society must found its explicit rules upon some measure of discretionary administration in order to account for a variety of normative considerations without a monistic (and therefore exploitable) meta-principle.

A course of action is 'discretionary' if it is decided on the spot rather than according to explicit and pre-announced rules. There are two very different things this can entail. The traditional economics of commitment takes the lack of precommitment to entail rational maximization of the objective payoffs, i. e. the sort of behavior that – per section 2 above – would make society impossible if pursued consistently. But refusal to commit to an explicit rule does not necessarily entail rational maximizing – at least, not of the objective payoffs – for agents with altruistic preferences. Rather, a more basic feature of this refusal is the *tacitness* of the rules used to generate action. Discretionary action, in other words, is not deterministically related to, and cannot be fully justified in terms of, basic values.[17]

[16] Their model is in the context of a dyadic Axelrod-contest, but generalizes straightforwardly to the *N*-person case.

[17] Such actions may, however, be justified *post hoc* in ways that bear no relation to the actual decision rule, which is inaccessible to conscious reflection (see Henrich 2016, chapter 7

Taking discretion to mean tacit decision rules flips the logic of rules versus discretion on its head. Indeed, rational maximizing – being in principle fully explicable – should be considered a variety of rule-bound behavior. And in this sense, it is the very *lack* of a fully articulable decision rule that generates commitment power in certain contexts. Some tacit, inarticulable, and flexible "gut sense" of being exploited is necessary for identifying and punishing defectors, lest a rigid or preannounced rule set be invaded by novel defection strategies. Tacit decision rules can be thought of as meta-rules governing strategy-switching in order to maintain something like a best response to these novel strategies. And indeed, the necessity of such a "gut sense" governing strategy switching in social settings seems to have been a significant driver of the development of human intelligence (Cosmides *et al*. 2010). Hayek's (1952, 192) dictum that the mind can never fully explain itself is more than simply a limit in principle to self-knowledge; it is also a precondition of social behavior.

Robocop, therefore, is not a viable method of precommitting to open-ended cooperation, for the same reason that complete contracts are impossible. If it is impossible to close the incentive gap on the basis of explicit individual incentives, any mechanical enforcement of preannounced or pre-programmed rules, however nuanced those rules may be, must eventually fall to exploitation without discretionary judgment and tacit decision rules as a backstop.[18]

One straightforward policy implication is that the push toward uniformity in judicial sentencing, for example with mandatory minimum sentencing guidelines, is likely to ossify the ability of bureaucratized societies to maintain cooperation, at least in the criminal justice arena. While the logic of the law may be of necessity fully articulated, the rules governing its application in particular cases is, and must remain, tacit. Vague and imprecise legal concepts such as negligence, reasonableness, and so on, are important bulwarks against novel forms of defection that arise in response to existing complexes of articulated rules. Any functional legal system must rely to some extent on the legitimacy of the decision rule Justice Potter Stewart used to delineate the category of pornography in *Jacobellis v. Ohio:* "I know it when I see it."

Second, the recent development of algorithmic contract enforcement and trustless exchange with blockchain technology cannot substitute for judges and juries any more than an explicitly programmed Robocop can enforce cooperation upon a community. These developments can be valuable on certain margins, especially to the extent that they can relieve courts of the burden of cases where the mechanical application of a rule is sufficient. This may indeed be many cases, but it can in principle never be all cases. In this sense smart contracts, decentralized autonomous organizations (DAOs),

for examples). This rationalization may nevertheless be important for the external legitimacy of a discretionary decision.

[18] The opaqueness of the processes behind "deep learning" neural network AIs (Knight 2017) – despite typically being seen as problematic (e. g. Park *et al*. 2017) – could in principle suffice. Of course, given that the legitimacy of punishment is the key constraint, an opaque algorithm is not likely to command much assent.

and other algorithmic contracts are a complement to, and not a substitute for, traditional contracts enforced by the judgment of a human mediator.[19]

Finally, the necessity of opaque decision rules to the maintenance of cooperation indicates why *sacredness* is such a central feature of human experience. Something sacred resists analysis, resists being broken down into its constituent concepts; it must be apprehended as a whole (cf. Rappaport 1971; Hayek 1952, 76). In contrast to the common perception of sacred values as inflexible, this resistance to articulation is precisely what creates space for prosocial strategy-switching.

It is no wonder, therefore, that attempts to analyze sacred concepts frequently provoke offense (Tetlock *et al*. 2000). This reaction is especially familiar to economists, who are in the business of pointing out tradeoffs. It is, of course, valid to take an outside perspective on sacred values and point out that – for example – a human life can have a concrete opportunity cost that may not be judged worthwhile to bear, hence the much-maligned "value of a statistical life" (Simon *et al*. 2019). At the same time, economists ought to appreciate the functional role of sacred values in human cooperation. Particular sacred values may – and often should – be criticized and analyzed, but sacredness itself cannot be dispensed with or rejected as irrational.

4.2 Inside and Outside Perspectives in Political Economy

The fact that economists frequently find themselves on the wrong side of sacred values should not be taken to imply that economics as a discipline stands firmly in the outside perspective. There is a rich tradition of inside-perspective economics: as radical critics of neoclassical economics point out, economics has come to serve a minor legitimating function for the role of markets in modern life, and it relies on tautologies (e. g. utility maximization) and descriptively false simplifications (e. g. perfect competition) at precisely those points in the intellectual edifice where the danger of defection (self-interested rent-seeking) or antisocial punishment (radical zeal) is greatest.

Pace the radicals, however, to point out these "lies" is not sufficient to impugn the status and utility of mainstream economics. To the extent that they obscure opportunities for strategic rent-seeking from policymakers, such "lies" may be truly noble, cooperation-enhancing, and self-confirming in exactly the same sense as belief in the wrath of God, regardless of whether that intent played any role in their development.[20]

[19] Harwick and Caton (2020) develops this argument at greater length. This argument does not preclude smart contracts as the technical scaffolding for the application of discretionary human judgment, on which see Lesaege and Ast (2018).

[20] Krugman (1993) is a particularly self-aware example. The real point of classical trade theory, Krugman argues, is not that tariffs can never be welfare enhancing, but to obscure opportunities for rent-seeking that an "optimal tariff" policy would illuminate. Buchanan and

Indeed it is hard to imagine any social-scientific analytical framework – including the one underlying such critiques (see the following section) – that does not rely on tautologies or descriptively false simplifications to legitimate cooperation or collective action of some sort or another, whether for or against the existing social order.

The inside-outside perspective distinction also runs directly through the middle of the economics of institutions, with Nobel prizes on both sides. On the outside are economic historians such as North (e. g. 1990; 2005), and Acemoglu and Robinson (2005; 2012), who – though they have a normative goal of economic development – approach the question functionally and historically. On the inside are "rational reconstructions" such as Buchanan and Tullock (1962)[21] and Rawls (1971), who are concerned about connecting existing or potential institutions with widely shared moral intuitions (sacred values) using thought experiments rather than history. The same distinction can be traced very far back through the Western canon. Hobbes ([1668] 2012) and Locke ([1690] 1960), for example, were engaged in projects on decisively different sides of the divide. If it is true that the inside and outside perspectives are irreducible one into the other, it is hardly surprising that the arguments of Hobbes and Locke have both maintained appeal and plausibility in the ensuing centuries despite their basic incompatibility.[22]

4.3 Critical Theory and the Ethics of Political Economy

Even with a meaningful methodological distinction between analysis and legitimation, it would be a mistake to try to draw the lines of economics, or of science more broadly, to exclude the latter. Such is the goal of critical theory and its offshoots; perhaps the most salient example of normative drift in the developed world.

Critical theory can be understood as, among other things, a method for analyzing social institutions in terms of objective payoffs, with "power" understood either directly as the index of those payoffs or as the ability to obtain them at the expense of others. The beliefs and preferences that pry apart a community's subjective preferences from their objective payoffs are understood as an exercise of power against them, even in the absence of overt or threatened punishment.[23] In other words,

Wagner (1977) lament the eclipse of classical public finance principles by Keynesian aggregate demand management on the same basis.

[21] Buchanan, at least, seems to have been self-aware on his assumed role as Noble Liar: "Our normative role, as social philosophers, is to shape this civic religion" (Brennan and Buchanan 1985, 166). See also Brennan and Buchanan (1988) and Leeson (2018).

[22] From the perspective of this paper, it could be argued that the solution to Hobbes' dilemma is not an overawing Leviathan – which, per above, poses its own dilemmas – but the fact that humans are apt to generate and internalize Locke-esque ideologies. Locke's work is not an effective answer to Hobbes, but the *existence* of Locke's work is.

[23] Foucault (1978) is one of the ur-texts for this expansive understanding of power, one that attacks the inside-outside perspective distinction more directly than previous conceptions

supposedly oppressed or marginalized classes could do better for themselves by minding their own payoffs and declining to buy into their community's noble lies.

Per the foregoing analysis, this contention is correct – or at least, there always exists some such class. And yet, whether or not critical theory's formulation of objective payoffs is coherent, understanding social institutions as epiphenomena of power relations (or of any other objective payoffs) throws us back to the dilemma of section 2 and renders social cooperation impossible (cf. Hayek 1988, 68). We have argued that there will always be parties in a society whose dominant strategy is defection. Critical theory, especially with its liberationist bent, is simply a method for identifying those parties and alerting them to that possibility[24] – perhaps the most deliberate method of doing so, but far from the only method. Indeed, the game theory in this paper points to the very same possibility.

This poses an ethical dilemma for the student of society, no less for the game theorist and the new institutionalist than for the critical theorist. On the one hand, a functionalist outside perspective is valuable for identifying systemic problems in economic development and institution-building (e. g. Acemoglu 2003). Without the ability to accurately identify the source of institutional failure, efforts at foreign aid and development are likely to be Sisyphean, if not actively harmful (cf. Easterly 2001, chapters 2 – 7). On the other hand, given that approaching social institutions from an outside perspective (whether critically or not) can render them impossible to maintain, it may also be the case that a scientific approach itself will do more harm than good.

For the same reason that explicit rules have a limited ability to support cooperation, there can likely be no hard-and-fast set of prescriptions for dealing with this problem. Nevertheless, there is some reason for optimism. First, inside-perspective beliefs are typically resilient to disconfirmation, especially where sacredness is involved – a fact which has often consternated iconoclastic intellectuals, but which may limit any damage done by the scientist interested in understanding rather than activism. Trial by ordeal may be impossible to maintain in a population of atheists, but there is evidence that people believe in order to support such institutions, rather than the institutions existing to prop up belief (Chen 2010; Ager and Ciccone 2018; Auriol *et al*. 2020). It

which have often focused on coercive power (which, per above, is exercised relatively infrequently on the equilibrium path of play). Foucault himself can be read as having some appreciation for the functional role of the exercise of power in his sense, but subsequent literature has been predominantly liberationist. Critical theory should be distinguished from orthodox Marxism in its rejection of dialectical materialism: power, for the Marxist, is epiphenomenal to modes of production.

[24] As an illustration, in recent years critical theory has been aimed at the institutions of science, even the scientific method itself, as upholding certain power structures. The "objectivity" of science is derided as a self-serving myth. It is of course true that science does not grant the scientist an Archimedian vantage point from which to view the world. It is, rather, a process of replacing descriptions of objects in terms of our senses with descriptions in terms of other objects (Hayek 1952, 3) – a process which can in principle never reach perfection, and one which benefits some interests over others. The benefits of science, like the benefits of society in general, are vast, but predicated on a prosocial myth – in this case, objectivity.

may, therefore, be difficult to "disenchant" a population without obviating the institution, a fact which would give the scientist more latitude in inquiry.

Second, the scientist may resort to what Melzer (2014, chapter 6) called "protective esotericism" and self-censor in popular works, a tactic with a long history (as Melzer documents) among intellectuals dealing with contemporary inside perspectives – largely for their own protection, but also partly for the sake of their wider society. In less accessible technical work, to the extent that disenchanted academics can still be relied upon for preference-based rather than belief-based altruism,[25] it will not be necessary to censor.

Even so, this analysis complicates the task of deriving policy implications from social scientific work. As economists have long recognized, optimal policies may be outside the feasible opportunity set in the absence of commitment power. But in a fitness landscape riddled with local optima and varying distributions of altruists willing to take up the slack left by failing belief, the importation of scientific-rationalistic modes of thought may clear away the coordinating power that previous institutions offered. And without a sufficient proportion of preference-altruists to maintain Western-style liberal democratic institutions, more virulent ideologies may rush in to fill the gap.[26]

5. Conclusion

The logic of social behavior gives rise to a structure of human motivation that implies an irreducible distinction between inside and outside perspectives on social institutions – that is, between legitimating exercises on the one hand, and analytical exercises on the other. That same logic implies that the distinction will in normal circumstances be invisible to the member of a particular society, to the extent that invisibility aids the internalization of an inside perspective. Because of the incentives inherent in the social organization of distantly related agents, it is necessary that the subjective preferences of those agents diverge from their objective payoffs in precisely the places that support the provision of public goods and the punishment of non-contributors.

Thus, as the world continues to adjust to communication technologies that facilitate the transfer of knowledge, norms, and modes of thought, there are two existential and

[25] As suggested by Eisenberg-Berg (1979) and Millet and Dewitte (2007), though see Madison *et al.* (2017).

[26] A large proportion of Islamic fundamentalist leaders, for example, are Western educated (Devarajan *et al*. 2016), not traditionalists in any meaningful sense. Similarly, scientists and engineers (whom we may take as exemplars of education in the Western scientific-rationalist tradition) are dramatically overrepresented among extreme Hindu nationalists in India (Lutz 2007, 151). The West has also seen its intellectuals swept by waves of ideological extremism: fascism in the early 20th century, communism throughout the 20th century, and critical theory in recent decades.

potentially reinforcing dangers to be avoided: first, the ascendancy of institutions within which individual selection dominates, eventually leading to the nonviability and extinction of cooperative strategies; and second, the drift of altruism into anti-social punishment.

On the one hand, this analysis connects the game theory of cooperation with the logical structure of human morality and offers a number of important practical considerations for both economics and policy. These implications are especially significant when varying cultural norms become a significant factor, as in the economic development literature. On the other hand, if clearsightedness can itself be detrimental, if functional institutions do depend on "noble lies" at some critical juncture, the actionability of those considerations can be ambiguous and fraught. Such, perhaps, is the great tragedy of the human condition.

References

Acemoglu, D. 2003. "Why Not a Political Coase Theorem? Social Conflict, Commitment, and Politics." *Journal of Comparative Economics* 31(4): 620–52.

Acemoglu, D. and J. Robinson. 2005. *Economic Origins of Dictatorship and Democracy.* Cambridge: Cambridge University Press.

Acemoglu, D. and J. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown Publishers.

Ager, P. and A. Ciccone. 2018. "Agricultural Risk and the Spread of Religious Communities." *Journal of the European Economic Association* 16 (4): 1021–68.

Alchian, A. 1950. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy* 58 (3): 211–21.

Alchian, A. and H. Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 62 (5): 777–95.

Alger, I. and J. Weibull. 2013. "Homo Moralis: Preference Evolution Under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–302.

Alger, I., J. Weibull, and L. Lehmann. 2020. "Evolution of Preferences in Group-Structured Populations: Genes, Guns, and Culture." *Journal of Economic Theory* 185. doi: 10.1016/j.jet.2019.104951.

Auriol, E., J. Lassebie, A. Panin, E. Raiber, and P. Seabright. 2020. "God Insures Those Who Pay? Formal Insurance and Religious Offerings in Ghana." *Quarterly Journal of Economics* 135 (4): 1799–1848.

Axelrod, R. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Bear, A., Kagan, A., and D. G. Rand. 2017. "Co-Evolution of Cooperation and Cognition: The Impact of Imperfect Deliberation and Context-sensitive Intuition." *Proceedings of the Royal Society B: Biological Sciences* 284. https://doi.org/10.1098/rspb.2016.2436.

Becker, G. 1962. "Irrational Behavior and Economic Theory." *Journal of Political Economy* 70 (1): 1–13.

Bénabou, R. and J. Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126 (2): 805–55.

Bergstrom, T. C. 2003. "The Algebra of Assortative Encounters and the Evolution of Cooperation." *International Game Theory Review* 5 (3): 211–28.

Bicchieri, C. 2016. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms.* New York: Oxford University Press.

Binmore, K. 1994. *Game Theory and the Social Contract,* Vol. 1. Cambridge, MA: MIT Press.

Bowles, S. and H. Gintis. 2004. "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations." *Theoretical Population Biology* 65 (1): 17–28.

Bowles, S. and H. Gintis. 2011. *A Cooperative Species: Human Reciprocity and its Evolution.* Princeton: Princeton University Press.

Boyd, R. and J. Lorberbaum. 1987. "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game." *Nature* 327 (7): 58–9.

Brennan, G. and J. Buchanan. 1985. *The Reason of Rules: Constitutional Political Economy.* Cambridge: Cambridge University Press.

Brennan, G. and J. Buchanan. 1988. "Is Public Choice Immoral? The Case for the 'Nobel' Lie." *Virginia Law Review* 74 (2): 179–89.

Buchanan, J. and G. Tullock. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* Ann Arbor: University of Michigan Press.

Buchanan, J. and R. Wagner. 1977. *Democracy in Deficit: The Political Legacy of Lord Keynes.* Cambridge, MA: Academic Press.

Bulow, J. and K. Rogoff. 1989. "Sovereign Debt: Is to Forgive to Forget?" *American Economic Review* 79 (1): 43–50.

Chen, D. 2010. "Club Goods and Group Identity: Evidence from Islamic Resurgence during the Indonesian Financial Crisis." *Journal of Political Economy* 118 (2): 300–54.

Cosmides, L., H. C. Barrett, and J. Tooby. 2010. "Adaptive Specializations, Social Exchange, and the Evolution of Human Intelligence." *Proceedings of the National Academy of Sciences* 107: 9007–14.

Curry, O. S. 2016. "Morality as Cooperation: A Problem-Centered Approach." In *The Evolution of Morality*, edited by T. Shackleford and R. Hansen, 27–51. Cham, Switzerland: Springer International.

Curry, O. S., D. Mullins, and H. Whitehouse. 2019. "Is It Good to Cooperate? Testing the Theory of Morality as Cooperation in 60 Societies." *Current Anthropology* 60 (1): 47–69.

Devarajan, S., L. Mottaghi, Q. Do, A. Brockmeyer, C. Joubert, K. Bhatia, and M. A. Jelil. 2016. "Economic and Social Inclusion to Prevent Violent Extremism." *Middle East and North Africa Economic Monitor* (October). Washington, D.C.: World Bank.

Easterly, W. 2001. *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics.* Cambridge, MA: MIT University Press.

Edgerton, R. 1992. *Sick Societies: Challenging the Myth of Primitive Harmony.* Florence, MA: Free Press.

Eisenberg-Berg, N. 1979. "Relationship of Prosocial Moral Reasoning to Altruism, Political Liberalism, and Intelligence." *Developmental Psychology* 15 (1): 87–9.

Eswaran, M. and A. Kotwal. 1984. "The Moral Hazard of Budget Breaking." *RAND Journal of Economics* 15 (4): 578–81.

Fehr, E. and S. Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–94.

Fitch, W. T. and M. D. Hauser. 2002. "Unpacking 'Honesty:' Vertebrate Vocal Production and the Evolution of Acoustic Signals." In *Acoustic Communication*, edited by A. M. Simmons, R. R. Fay, and A. N. Popper, 65–137. New York: Springer.

Foucault, M. 1978. *The History of Sexuality, Vol. 1: An Introduction.* New York: Random House.

Fudenberg, D. and D. K. Levine. 1993. "Self-Confirming Equilibrium." *Econometrica* 61 (3): 523–45.

Gintis, H. 2016. "Homo Ludens." *Journal of Economic Behavior and Organization* 126: 95–109.

Grüne-Yanoff, T. 2011. "Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection." *Biology and Philosophy* 26: 637–54.

Hardin, R. 1985. "Individual Sanctions, Collective Benefits." In *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, edited by R. Campbell and L. Sowden, 339–354. Vancouver: University of British Columbia Press.

Harwick, C. 2018. "Money and its Institutional Substitutes: The Role of Exchange Institutions in Human Cooperation." *Journal of Institutional Economics* 14 (4): 689–714.

Harwick, C. and J. Caton. 2020. "What's Holding Back Blockchain Finance? On the Possibility of Decentralized Autonomous Intermediation." *Quarterly Journal of Economics and Finance.* doi: 10.1016/j.qref.2020.09.006.

Hayek, F. A. 1952. *The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology.* Chicago: University of Chicago Press.

Hayek, F. A. 1988. *The Fatal Conceit: The Errors of Socialism.* Chicago: Chicago University Press.

Henrich, J. 2016. *The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making us Smarter.* Princeton: Princeton University Press.

Henrich, J., S. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2–3): 61–83.

Hermann, B., C. Thönu, and S. Gächter. 2008. "Antisocial Punishment Across Societies." *Science* 319 (5868): 1362–7.

Hobbes, T. (1668) 2012. *Leviathan, or, The Matter, Forme and Power of a Common-Wealth Ecclesiasticall and Civil.* Oxford: Oxford University Press.

Holmström, B. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13 (2): 324–40.

Iannaccone, L. R. 1992. "Sacrifice and Stigma." *Journal of Political Economy* 100 (2): 271–91.

Kandori, M. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59 (1): 63–80.

Knight, C. 1998. "Ritual/Speech Coevolution: A Solution to the Problem of Deception." In *Approaches to the Evolution of Language*, edited by J. R. Hurford, M. Studdert-Kennedy, and C. Knight, 68–91. Cambridge: Cambridge University Press.

Knight, W. 2017. "The Dark Secret at the Heart of AI." *MIT Technology Review* 120 (3): 54–77.

Krebs, J. and R. Dawkins. 1978. "Animal Signals: Mind-Reading and Manipulation." In *Behavioural Ecology: An Evolutionary Approach*, edited by J. Krebs and N. B. Davies. Oxford: Blackwell Scientific Publications.

Krugman, P. 1993. "The Narrow and Broad Arguments for Free Trade." *American Economic Review* 83 (2): 362–66.

Kydland, F. and E. Prescott. 1977. "Rules Rather Than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85 (3): 473–92.

Leeson, P. 2012. "Ordeals." *Journal of Law and Economics* 55: 691–714.

Leeson, P. 2013a. "Gypsy Law." *Public Choice* 155 (3–4): 273–92.

Leeson, P. 2013b. "Vermin Trials." *Journal of Law and Economics* 56: 811–36.

Leeson, P. 2018. "Beneficent Bullshit." In *James M. Buchanan: A Theorist of Political Economy and Social Philosophy*, edited by R. Wagner. London: Palgrave Macmillan.

Leeson, P. and C. Coyne. 2012. "Sassywood." *Journal of Comparative Economics* 40 (4): 608–20.

Leeson, P. and P. Suarez. 2015. "Superstition and Self-Governance." *Advances in Austrian Economics* 19: 47–66.

Ledyard, J. O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, edited by J. Kagel and A. E. Roth, 111–94. Princeton: Princeton University Press.

Lesaege, C. and F. Ast. 2018. "Kleros." Accessed January 31, 2021. https://kleros.io/assets/whitepaper.pdf.

Locke, J. (1690) 1960. *Two Treatises of Government.* Cambridge: Cambridge University Press.

Luce, E. 2007. *In Spite of the Gods: The Strange Rise of Modern India.* New York: Doubleday.

Madison, G., E. Dutton, and C. Stern. 2017. "Intelligence, Competitive Altruism, and 'Clever Silliness' May Underlie Bias in Academe." *Behavioral and Brain Sciences* 40. https://doi.org/10.1017/S0140525X15002368.

Melzer, Arthur. 2014. *Philosophy Between the Lines: The Lost History of Esoteric Writing.* Chicago: University of Chicago Press.

Miller, G. 1992. *Managerial Dilemmas: The Political Economy of Hierarchy.* Cambridge: Cambridge University Press.

Millet, K. and S. Dewitte. 2007. "Altruistic Behavior as a Costly Signal of General Intelligence." *Journal of Research in Personality* 41 (2): 316–26.

Nikiforakis, N. 2008. "Punishment and Counter-Punishment in Public Goods Games: Can We Still Govern Ourselves?" *Journal of Public Economics* 92 (1–2): 91–112.

Norenzayan, A., A. F. Shariff, W. M. Gervais, A. K. Willard, R. A. McNamara, E. Slingerland, and J. Henrich. 2016. "The Cultural Evolution of Prosocial Religions." *Behavioral and Brain Sciences* 39. https://doi.org/10.1017/S0140525X14001356.

North, D. 1990. *Institutions, Institutional Change, and Economic Performance.* Cambridge: Cambridge University Press.

North, D. 2005. *Understanding the Process of Economic Change.* Princeton: Princeton University Press.

Ostrom, E. 2005. *Understanding Institutional Diversity.* Princeton: Princeton University Press.

Park, D. H., L. A. Hendricks, Z. Akata, A. Rohrback, B. Schiele, T. Darrell, and M. Rohrback. 2017. "Attentive Explanations: Justifying Decisions and Pointing to the Evidence." Accessed January 31, 2021. https://arxiv.org/pdf/1612.04757.pdf.

Rand, D. G., J. J. Armao, M. Nakamaru, and H. Ohtsuki. 2010. "Anti-Social Punishment Can Prevent the Co-Evolution of Punishment and Cooperation." *Journal of Theoretical Biology* 265 (4): 624–32.

Rappaport, R. A. 1971. "Ritual, Sanctity, and Cybernetics." *American Anthropologist* 73 (1): 59–76.

Reiter, J. G., C. Hilbe, D. G. Rand, D. G., K. Chatterjee, and M. A. Nowack. 2018. "Crosstalk in Concurrent Repeated Games Impedes Direct Reciprocity and Requires Stronger Levels of Forgiveness." *Nature Communications* 9. https://doi.org/10.1038/s41467–017–02721–8.

Root, H. 1989. "Tying the King's Hands: Credible Commitments and Royal Fiscal Policy During the Old Regime." *Rationality and Society* 1 (2): 240–58.

Searcy, W. and S. Nowicki. 2005. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems.* Princeton: Princeton University Press.

Simon, N. B., C. Dockins, K. B. Maguire, S. C. Newbold, A. J. Krupnick, and L. O. Taylor. 2019. "What's in a Name? A Search for Alternatives to 'VSL'." *Review of Environmental Economics and Policy* 13 (1): 155–61.

Simons, H. C. 1936. "Rules Versus Authorities in Monetary Policy." *Journal of Political Economy* 44 (1): 1–30.

Sober, E. and D. S. Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge MA: Harvard University Press.

Stewart, A. J., L. Parsons, and J. B. Plotkin. 2016. "Evolutionary Consequences of Behavioral Diversity." *Proceedings of the National Academy of Sciences* 113 (45): E7003–9.

Sugden, R. 2001. "The Evolutionary Turn in Game Theory." *Journal of Economic Methodology* 8 (1): 113–30.

Sugden, R. 2003. "The Logic of Team Reasoning." *Philosophical Explorations* 6 (3): 165–81.

Taub, B. 1985. "Private Fiat Money with Many Suppliers." *Journal of Monetary Economics* 16 (2): 195–208.

Tetlock, P. E., O.V. Kristel, B. Elston, M. C. Green, and J. S. Lerner. 2000. "The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals." *Journal of Personality and Social Psychology* 78 (5): 853–70.

Tomasello, M. 2009. *Why We Cooperate.* Cambridge, MA: MIT Press.

Yamagishi, T. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology* 51 (1): 110–16.