Public Reason 9 (1-2): 61-78

Context Dependence in Gaus's Evolutionary Account of Public Reason

Luca Costa
University of Pavia

Abstract: One of the distinctive features of Gerald Gaus's public justification theory is his extensive use of the empirical data from the social sciences to support his normative claims. One such claim which stands out for its importance, within the context of Gaus's theory, is the evolution of *strong reciprocity*: over time, members of large societies develop a tendency to follow social rules and punish defectors. This claim, in turn, is supported by several studies in experimental economics involving mixed motive games, which show how subjects are inclined to punish, even at a net cost for themselves, any perceived violation of social norms such as fairness. While critics of Gaus's theory focused mainly on whether the use of empirical evidence in a normative theory is *prima facie* admissible, in this paper I address two different issues. The first concerns the accuracy of the empirical evidence used by Gaus's theory, and whether the conclusions advanced by the social scientists on the grounds of this evidence are actually sound. The second issue, on the other hand, concerns the consistency between the empirical assumptions underlying these experiments and the claims of Gaus's theory. I argue that both concerns are warranted. On the one hand, there is empirical evidence that the rule-following behavior observed in experiments involving mixed motive games is *context dependent*: the tendency of subjects to follow rules and punish defectors is strongly correlated to the epistemic constraints, such as anonymity, commonly imposed during the experiments. On the other hand, these constraints hardly reflect the circumstances of life in modern societies, which is the context assumed by Gaus's justificatory theory. I conclude that Gaus's account of rule-following punishers is not altogether invalidated, but the empirical evidence from experimental economics is insufficient to support it.

Key words: experimental economics, Gerald Gaus, naturalistic fallacy, public justification, social morality, ultimatum games.

Interest in the contributions of social sciences has been all but a recurring feature in the contemporary debate on public reason. For example, John Rawls's *Political Liberalism* (1996) based his theory of public reason on a conception of citizens as reasonable and rational, which is described as a philosophical scheme rather than an accurate account of humans' moral psychology (1996, 81-87); and this choice, in turn, is motivated by a sceptical stance towards social sciences, which are claimed to be incapable of providing useful insight beyond what historical knowledge and common sense would already suggest (1996, 88). In other words, social sciences are unable to accurately specify facts about human nature, which could shed light on what limits there are to the viability of a philosophical conception of citizens.

However, social sciences have recently known an unprecedented development. Different research fields, such as evolutionary psychology (Barkow et al. 1992), gene-culture coevolution theories (Richerson and Boyd 2005), and sociology (Bicchieri 2006) have produced an impressive bulk of data about humans' biological and social nature, undermining the grounds for Rawls's scepticism.

Among the first ones to understand the relevance of these results for political philosophy[1], Gerald Gaus did not follow Rawls in his distrust. Rather, one of the most important arguments of the first part of his *The Order of Public Reason* (Gaus 2011) is that recent studies in the fields of experimental economy, evolutionary psychology, and game theory provide a significant support for the evolution of *strong reciprocity* – i.e. the claim that individuals have a tendency to cooperate when enough others show a cooperative behaviour, and to punish those who violate social norms and rules (2011, 105). This argument represents a two-fold novelty in the contemporary debate on public reason. On the one hand, it defends the claim that recent experimental data from the social sciences can actually have a role in a theory of public reason. On the other hand, it draws the attention of political philosophy on insofar neglected fields of research whose subjects of interest are relevant for normative political theory, as well.

However, an appeal to empirical data, in order to support normative claims, can raise several suspicions. Even if one were to find out that, for example, individuals actually cooperate and punish defectors, as the hypothesis of strong reciprocity predicts, how would this tell us anything in regards to how they *ought* to act? In other words, how could the descriptive claim that people act in a certain way hold any normative weight? Any serious attempt to ground a normative theory on an empirical account will first need to dispel these doubts.

The relation between descriptive and normative levels is not the only issue at stake, either. It might be the case, in fact, that even if the logical step from one level to another is overall sound, the descriptive data this step relies on are inadequate. On the one hand, these data may simply be incorrect, underdetermined, or seen as controversial by the contemporary scientific community. On the other hand, even if these data were substantially correct, they could nonetheless rely on theoretical premises (regarding the properties of the subject of study) which are at odds with the ones assumed by Gaus's empirical account.

Before I address these points, however, I will first provide a brief overview of Gaus's empirical account: its theoretical background, its basic factual claims, its role within Gaus's normative framework. This will be the subject of the first section of this paper. In the second section, I will examine the data referenced by Gaus in support of his empirical account. In the third section, I will address the aforementioned concerns about the relation between empirical data and normative claims. Then, in the fourth and fifth sections, I will explore these empirical data in further detail, in an attempt to assess their plausibility within the context of Gaus's theory. I will conclude with an assessment of Gaus's evolutionary approach, in the light of the criticism advanced in this paper.

1] On the other hand, interest of social sciences in political philosophy and political consequences in a wider sense has been mixed. While sociobiologists have defended the role of social sciences in normative thought (Ruse and Wilson 1986), evolutionary psychologists have more or less explicitly refused any involvement in the political side (Thornhill and Thornhill 1992), out of fear of the "naturalistic fallacy" (Moore 1903). For a critical review of the latter position, as well as a more detailed account of this debate, see Dupré (2001).

I. MORALITY AND EVOLUTION

I mentioned above that one of the central arguments of Gaus's work is that there is significant empirical evidence for the evolution of strong reciprocity. But the reader might still feel unconvinced. Why is it so important to actually rely on empirical evidence, rather than a more intuitive or idealized account? And what does evolution have to do with public justification? In this overview of Gaus's evolutionary account, I will seek to address both questions.

In order to answer the first issue, concerning the relevance of empirical evidence, one must first turn to Gaus's approach to morality. Gaus is not interested in morality and ethics in general, but only in that subset of moral rules which require, allow, or prohibit certain courses of action when members of society interact with each other: this subset is what Gaus calls *social morality* (2011: 2-3). Gaus's view of morality – and, specifically, of social morality – is *functional*: according to Gaus, social morality is of value to us (and we consider its claims to be authoritative) because of its essential function in allowing cooperation among human beings in a large-scale society (2011, 2-6, 145-47, 191-93; 2015a, 1081-82).[2]

For social morality to perform this function, however, it is crucial that members of society actually endorse it, and see its imperatives as binding: a social morality which nobody conforms to is no social morality at all (Gaus 2011, 163-64). This does not necessarily require actual acceptance of the rules of social morality, as people can withhold their acceptance due to ignorance, confusion, or stubbornness (1996, 123). At the same time, though, social morality cannot be justified by something like a hypothetical agreement between highly idealized counterparts of members of society (like in Rawls 1999), as actual people cannot be bound by an agreement struck between strongly idealized versions of themselves (Dworkin 1973). Therefore, Gaus opts for a moderate idealization of members of society in the deliberative process, which gives them a few cognitive restraints (such as the impossibility to lie about their own preferences), but otherwise lets them retain all the psychological and social features of their actual counterparts (Gaus 2011, 276).[3]

Within such a perspective, empirical data about human nature assume renewed relevance. In order to build a moderate idealization of actual members of society, in

2] The focus on this subset of the moral sphere, as well as the functional approach to morality, is by no means unique to Gaus's theory, though. It dates at least back to the work of John Stuart Mill (1963), and can be found in more recent philosophers, as well (Strawson 1961; Gauthier 1986; Baier 1995; Kitcher 2011).

3] It is important to note here that Gaus is not claiming that all and only the pre-existing rules of social morality (what he calls *positive morality*) should be justified. To the contrary, the content of a justified social morality (what he calls *true morality*) can be quite different from the one of positive morality. But, at the same time, due to the function social morality needs to perform in society, the content of true morality cannot stray too far from the one of positive morality (Gaus 2011, 56). For a more articulated motivation behind this stance, see his argument about the *path-dependency* of morality (2011, 242-43).

fact, it is important to be aware of what these actual members of society look like; their different approaches to morality; their attitudes towards social rules; and so on. I will say more about this point in the following, but first, I need to address the second issue this section started with: what is the point of evolution?

As I mentioned before, Gaus endorses a functional view of morality, according to which the practice of social morality has a point – allowing cooperation among members of society. But there is also another approach to public justification which shared such a view of morality: the instrumentalist approach, pioneered by the seminal work of David Gauthier (1986). This is why Gaus dedicates a significant part of his own book to evaluate the merits and the limits of this approach (2011, 53-100), and why it is useful to briefly review the main points of Gaus's criticism here.

According to Gauthier, the justification of social morality is grounded in the rationality of persons: if they follow the rules of social morality, rather than solely trying to maximize their respective utilities, everyone can do better: it is thus rational to show restraint in one's own selfish motivations and comply with the demands of morality, because doing so is the best way for everyone to advance his or her own ends (1991, 22-23).

Consider for example a typical Prisoner's Dilemma scenario: two instrumentally rational people, Alf and Betty, are faced with a situation where they have to decide whether to cooperate or not. For each person, not cooperating is always the best ("dominant") choice. If the other party cooperates, not cooperating allows to "free-ride" and reap the benefits of cooperation without suffering the costs. And if the other party does not cooperate, not cooperating at least allows each person not to suffer the costs of cooperation without getting its benefits. The best choice for both parties is thus not to cooperate, to avoid being free-ridden. But Alf and Betty will get a lower payoff, by acting in this way, than what they would have got if they cooperated.

Following moral rules poses a similar dilemma. Members of society who only try to maximize their own utilities will always choose to "not cooperate" – which in this case means that they will defect from the demands of social morality. But if everyone defects, the result is that everyone is worse off than if each of them opted to comply with moral rules instead. So, how can Alf and Betty reason themselves out of this dilemma?

According to Gauthier, the mistake in this scenario is to assume that both parties ought to endorse a *modular* form of rationality. A rational course of action is considered to be modularly rational if, at any time t_i between the beginning and the end of the action, complying with the course of action decided at t_0 is still a better choice than defecting from it. So, in a Prisoner's Dilemma scenario, if Alf and Betty are modularly rational, they will never manage to agree to not defect, even if they promise each other to do so: once the promise is done, in fact, the best course of action (the one yielding the highest pay off) for each of them is still to defect.

Gauthier argues that it is wrong to assume that rationality requires modularity. Instead, he embraces a notion of rationality as *effectiveness*, according to which a course of action is rational if employing it is conducive to one's life going as well as possible, even

if at some point of the performance of that action it might appear that doing so is not in the agent's best interest (1994, 701).

Can past really hold such a sway over the present, though? Even assuming both parties are in good faith when they promise to cooperate, their information about what makes their courses of action good or bad may change through time. Is it really rational – in any intuitive, commonly shared sense of the term – to ignore such information and stick with the original course of action because, at the time the agent formed it, it was "conducive to his or her life going as well as possible"? And at the same time, is it really rational to assume good faith such easily in the first place? In the aforementioned Prisoner's Dilemma scenario, for example, unless we assume Alf and Betty know each other well enough to have good reason to believe in each other's promises, they are taking a risk which may well not pay off at all. This is why Gaus argues that, ultimately, the instrumentalist approach failed. In his attempt to implement it, in fact, Gauthier ended up revising the concept of rationality in a way that undermined the plausibility of his solutions (Gaus 2011, 100).

How can members of society reason themselves out of this Prisoner's Dilemma, then? It might seem that, after all, an independent motivation such as reasonableness is required – a motivation that cannot be derived from mere instrumental rationality (Rawls 1996, 50-51). Gaus opts for a similar solution, but he appeals to evolutionary forces, rather than reasonableness, to solve the dilemma. Following a proposal originally advanced by Bryan Skyrms (1996), Gaus claims that, unlike rationality, evolution is not bound by constraints such as modularity: evolutionary forces can select a strategy *T* on the grounds that those employing *T* outperform those who employ different strategies, even if *T* sometimes instructs people to act in ways that do not best promote their own goals (Gaus 2011, 104-05).

Following the rules of social morality is one such strategy (2011, 105-12). Members of society who adopt this strategy sustain a cost, based on how complying with rules limits their ability to pursue their own ends. In return, however, they get a reward for their troubles: cooperating with other rule followers allows them to reap benefits they would not have been able to achieve on their own. And these benefits may improve each member's ability to pursue his chosen end. If the benefits outweigh the costs, we can say that each member has a *payoff* from adopting the rule following strategy, based on the difference between costs and rewards. We can call this strategy *simple rule following*.

Simple rule following, as a strategy, has the advantage of a positive payoff through the benefits of cooperation. However, it has a crucial weakness: it can be *invaded*. An instrumentally rational agent who only follows rules when doing so does not thwart his own ends, for example, may be able to reap the benefits of cooperation from simple rule followers; but he does not sustain the same costs, because he follows rules only when it is contingently convenient for him to do so. In a society of simple rule followers, therefore, the instrumentally rational agent who defects rule following

would have a higher payoff than other simple rule followers. Over time, more and more defectors would successfully invade that society, until the majority of its members are all rule defectors.

However, suppose that, in a different society, its members not only follow rules, but also punish all members who does not comply with rules. We can call these members *rule-following punishers*. The presence of rule-following punishers suddenly makes defecting a less advantageous strategy, because the former's punishment imposes a cost on the latter which reduces the latter's overall payoff. Of course, the cost of punishing also reduces the payoff of rule-following punishers, compared to simple rule followers. However, if defectors are not present in significant numbers (which is likely, given their reduced payoff), the additional cost incurred by rule-following punishers may be low enough to allow this strategy to thrive alongside simple rule followers.[4] Rule-following punishers would thus be an evolutionarily stable strategy (ESS) under these conditions.[5]

Unlike instrumental explanations, Gaus's evolutionary account does not thus explain how can agents reason themselves into following rules. Rather, it aims to show that rule followers can be more "fit" than agents devoted only to their own ends (2011, 112). But this still leaves a crucial question unanswered: what role does this evolutionary account play in Gaus's normative theory of public justification?

Gaus's proposed model of public justification is a convergent process that selects and implements a subset of moral rules out of a morally eligible set (2011, 321-25), under the assumption that members of society (or, more precisely, their moderately idealized counterparts) will recognize there is no other way to secure the benefits of social cooperation. This convergence of a single option from the eligible set is neither foreseen nor constructed in advance by any of the members, but nevertheless each of them has sufficient reason to accept the outcome of this process (2011, ch. 19).

What makes this process plausible, though, is the hypothesis that people are actually capable of internalizing and following rules, rather than acting as narrowly instrumental agents. The eligible set, in fact, includes several options that, while still superior – in the eyes of some members of society – to not having any social rule at all (that is the basic condition for them to belong to the eligible set, in the first place), are still far from optimal for those members. If members of society were not capable of developing a motivation to comply with rules, even when this hinders their chosen goals, convergence on any social rule would be actually impossible.

4] The advantage of simple rule followers over rule-following punishers may be further reduced by "second-order" forms of punishment, which punish members who follow rules but do not punish defectors (Boyd and Richerson 2005, 166-79).

5] An ESS is a strategy S if there is no other strategy T which can get a better payoff by playing against S. For a classical study on ESS, see Smith (1978).

II. EXPERIMENTAL DATA

However, is this a realistic scenario? Do we have a reason to think that we, or at least the majority of us, are indeed rule-following punishers? As I noted in the previous section, the plausibility of this evolutionary account rests on the assumption that a sufficient number of rule followers arose in the population. For this reason, Gaus shows recent experiments that, he argues, provide a strong case for the existence of rule-following punishers (2011, 119).

The large majority of these studies are experiments involving several variants of the *Ultimatum Bargaining Game*. In its simplest form, it is a one-shot game played by two subjects – a *Proposer* and a *Responder* – who have to bargain over a certain amount X of some good (usually money). The Proposer moves first and has to offer a share n of X to the Responder, where the offer can range between X and zero. The Responder has two choices: to accept the offer, or to reject it. If the offer is accepted, the Responder receives n while the Proposer receives $X - n$. If the offer is rejected, each player receives nothing.

If the Responder were an instrumentally rational agent, and since instrumentally rational agents are assumed to always choose more over less of anything they value, he or she would accept any offer higher than zero. Likewise, if the Proposer were an instrumentally rational agent – and if he or she expected the Responder to act like an instrumentally rational agent, as well – he would make an offer higher than zero, but as small (as close to zero) as possible. So, if we suppose that $X = 100$ and that n must be a natural number, the instrumentally rational Proposer would always offer 1, and the instrumentally rational Responder would always accept.

This is not the observed outcome, though. Studies of one-shot Ultimatum Games involving university students from the United States and other countries showed that median offers of proposers to responders are significantly more common than mean offers (40-50% against 30-40%, respectively), and mean offers (below 20% of the share) are refused roughly half the time.[6] The divergence from the scenario predicted by the hypothesis of instrumentally rational agents is thus significant: not only proposers tend to offer more than expected, but responders do not accept all offers either (in fact, they refuse mean offers with surprising frequency). Many social scientists thus interpreted these results as an evidence of an aversion to unfair results from the subjects (Fehr and Schmidt 1999; Bolton and Ockenfels 2000).

However, further experiments challenged this explanation. In another study (Falk et al., 2000), a variant of the ultimatum game has been tested, where proposers

6] These data come from the following studies: (Bolton and Zwick 1995; Cameron 1999; Croson 1996; Eckel and Grossman 2001; Hoffman et al. 1994, 1996; Güth et al. 1982; Roth et al. 1991; Forsythe et al. 1991; Harrison and McCabe 1996; Larrick and Blount 1997; List and Cherry 2000; Rapoport et al. 1996; Schotter et al. 1996; Slonim and Roth 1998. A review of these studies can be found in Camerer 2003).

could only choose one of two possible offers (and the responders were aware of the choices available to their proposers). When the proposers' alternatives were between a mean offer and a fair offer (80-20 vs 50-50), the rejection rates of the mean offer were on par with the previous experiments (44.4% rejected the 80-20 offer). However, when the proposer was forced to choose between a mean and an altruistic offer (80-20 vs 20-80), the rejection rates of the mean offer were significantly lower (18% rejected the 80-20 offer). When the only choice available to the proposer was to either make a mean offer or an unreasonably altruistic one (80-20 vs 0-100), the rejection rates of the mean offer were extremely low (8.9% rejected the 80-20 offer). It would seem, therefore, that responders change their behaviour according to the options available to the proposer, and show different acceptance rates of equally unfair shares.

A study which does not involve ultimatum games arguably produced similar results (Falk et al. 2005). In the first stage of this experiment, subjects engaged in a 3-player Prisoner's Dilemma, where they had to decide simultaneously whether to cooperate or defect. If both other players defected, cooperating yielded a payoff of 12 while defecting yielded 20. If one of the other two players cooperated, defecting yielded 32 while cooperating yielded 24. If both other players cooperated, defecting yielded 44 while cooperating yielded 36. In all three scenarios, defecting was the dominant strategy, but if all people cooperated they yielded higher payoffs than if they all defected.

In the second stage, each player was informed about the other players' individual decisions for stage 1, and could choose to punish them by reducing the payoff of either or both of the other players, by a maximum of 25. However, the cost of the punishment greatly differed in the two treatments of second stage. In the *low-sanction* treatment, punishing was equally costly for the punisher and the punished. For example, if a player decided to spend 6 to punish another, the other player would have seen his payoff reduced by 6. To the contrary, in the *high-sanction* treatment, punishment was more effective against defectors and even more effective against cooperators. So, in this treatment, a player who decided to spend 6 to punish another would reduce the other player's payoff by 15 (2.5 factor) if he were a defector, or by 19 (3 1/3 factor) if he were a cooperator. The results showed that, even in the *low-sanction* treatment, the majority of cooperators (59.6%) decided to punish defectors, even if doing so would not reduce the inequality of the outcome (as the cooperator's payoff would be reduced as much as the defector's).

While studies evidence a tendency, from subjects, to cooperate and punish defectors, it seems therefore that they do not do so because they are motivated by a distaste for unequal outcomes. The best explanation, Gaus argues, is that the subjects "endorse norms about fairness in certain sorts of interactions and are willing to forgo material benefits to punish those who do not comply" (Gaus 2011, 122).

III. THE PROBLEM OF FACT SENSITIVITY

As noted at the end of the first section, the role of this empirical account in Gaus's public justification model is permissive.[7] It does not dictate which normative claims one must adopt; rather, it limits the range of admissible normative models, by making some more or less plausible than others. Nevertheless, it may still draw criticism, insofar as it draws normative conclusions from descriptive premises. In this section, I will thus consider two different arguments against the use of descriptive evidence in normative theory, and how Gaus's model fares against them.

First, the contribution of empirical data may be accused of being redundant. According to this line of criticism, the issue is not that the normative claim is false, *per se*. Rather, its validity does not actually rest on the empirical evidence, as assumed by the criticized argument, but on a third, different (and often hidden or implicit) normative premise, which explains why the empirical evidence supports the normative claim, in the first place.[8]

Suppose, for example, that a person claimed that we have a standing duty to respect promises (we shall call this normative claim *P*); and that we have this duty because there is empirical evidence that respecting promises improves the promisees' chances to pursue their own conceptions of the good (we shall call this descriptive claim *F*). In other words, the claim is that *F* is sufficient to support *P* – i.e. *P* has moral authority because *F* is true. However, this still does not clarify why the person should believe that *F* actually supports *P*, unless she adds an additional premise *P'* along the lines of "we have a standing duty to improve people's chance to pursue their own conceptions of the good". It is thus *P'* which causes *F* to "matter", and that causes *F* to support *P*. But *P'* itself is not supported by *F*: even if the person were to believe that *F* is false, she would still believe that *P'* is true. In the end, therefore, it would seem that it is the normative framework (the principles *P* and *P'*) that are "doing the work", and the appeal to *F* would be largely redundant. This line of criticism is not especially problematic for Gaus's argument. In fact, consider his aforementioned functional view of morality. Is it a descriptive or a normative claim? Were it to be a descriptive claim, we would have made no progress, as one could still press the question of why this mere fact, that social morality exercised an essential function in large-scale societies, supports Gaus's justificatory model. And it seems apparent that Gaus's functional view of morality is definitely stating something about the empirical world. However, it would be a mistake to conclude that Gaus's stance is merely a descriptive one. The claim that "*of course* social morality has a point – providing the foundations for social life – and this fact must shape our understanding of it" (Gaus 2011, 56) does not just state that social morality provides the foundations for social life, but that our understanding of

7] This definition of the permissive character of accounts of human nature follows closely Rawls (1996, 87).

8] This line of criticism draws primarily from G. A. Cohen's analysis of normative principles and their sensitivity to empirical facts (Cohen 2003).

social morality should be shaped by how it provided such foundations. This latter part is a normative claim. Not only is it not itself supported by any fact, but it may very well be the case that one could believe the fact that social morality provided the foundations of social life, while denying that this should shape our understanding of it (Enoch 2013).

Does this mean, then, that empirical evidence does not play any relevant role, because it is only the truth of this added normative claim (the functional view of morality) which makes them support Gaus's justificatory model? Again, such a conclusion would be arguably mistaken. Suppose, in fact, that the empirical evidence Gaus appeals to in support of his theory is somehow flawed, or otherwise wrong. Even if this had no weight on whether or not the functional view of morality is valid, it still could have significant consequences in regards to the validity of Gaus's other normative claims – consequences which would be likewise supported by his very view of morality. To say that Gaus's empirical evidence is redundant, simply because some of his normative claims are not supported by them, would underestimate the extent to which his other, not any less important, normative claims depend on them for their validity.[9]

According to a second line of criticism, though, the issue is not just that there is a third premise which is required to show that the empirical evidence actually supports the normative claim. Rather, the charge is that this third premise is false, or otherwise implausible. In other words, this line of criticism rejects Gaus's functional view of morality as a valid normative claim. David Enoch, for example, objects that morality itself is hardly the kind of thing which can have a function – in the same sense in which we do not see physics (rather than the study of physics) as having a function (Enoch 2013, 149).

Gaus's reply to this objection is that taking physics as a fitting model for morality is itself questionable, and law would make for a more appropriate comparison (2015a, 1081). If we were to truly believe that morality has no function at all, Gaus argues, all the costs moral life imposes on us – such as guilt, blame, and rebuke – would look more like a neurosis than something rationally justifiable. Furthermore, a functional view of morality has the added benefit of making sense of all the work of ethnographers and social scientists, who commonly assume morality to be an essential form of social adaptation (2015a, 1082).

This answer is hardly conclusive. It could still be replied, in fact, that a realist view of morals could be just as capable of explaining the purported role social morality played in the evolution of large-scale societies (Enoch 2013, 149); and that furthermore, such a view would make sense of our actual moral attitudes and responses in a better way than Gaus's functional perspective (2011, ch. 1). However, it should be noted that this debate, while crucial, does not bear specifically on the use of empirical evidence we are concerned with, at this point. When Enoch says that the functional view of social morality cannot vindicate Gaus's appeal to studies in the social sciences (2013, 148), he is right in saying

9] And in fact, to be fair, the purpose of Cohen's argument was not to claim that there cannot be principles (normative claims) which depend on the support of facts for their validity, but simply that «*a principle can reflect or respond to a fact only because it is also a response to a principle that is not a response to a fact*» (2003, 214). And this more modest claim is, as I have argued, fully compatible with Gaus's use of empirical evidence.

so from the point of view of a moral realist. However, unless further reasons for why one should dismiss a functional view of morality are presented, this does not show that Gaus incurs in the naturalistic fallacy as Enoch claims (2013, 150), but merely that the normative principle which justifies the support of empirical evidence for Gaus's theory is incompatible with a certain metaethical stance.

IV. CONTEXTUAL ISSUES WITH THE EXPERIMENTAL EVIDENCE: THE IDENTITY OF SUBJECTS

The conclusion of the last section is somewhat promising for Gaus's argument. While it does rely on a controversial view of morality, the criticisms about fact sensitivity that I discussed do not undermine its internal consistency. As long as Gaus's normative premises about the nature and role of social morality are accepted as valid, the use of empirical evidence within the context of his framework seems, at least in principle, admissible.

However, as noted before, there is another possible source of concern that we ought to consider: is Gaus's empirical evidence actually sound? And most importantly, does this evidence actually support the conclusions Gaus draws from them? The objective of this section, as well as the next one, will thus be to assess this concern – with particular attention to the evidence supporting the existence of rule-following punishers. I will thus begin with a discussion regarding the nature of the subjects involved in the experiments I presented in section two.

Who are these subjects? In each and every of the aforementioned studies, they are always graduate or (most commonly) undergraduate students from universities, generally from economics and MBA classes, that are invited to participate in exchange for a fee (such as in Harrison and McCabe 1996; Forsythe et al. 1991; Schotter et al. 1996), or to get credits for their final exams (such as in Güth et al. 1982).

It is hardly obvious that university students as experimental subjects can be representative of humanity as a whole. After all, there are billions of humans who never attended a university, in the first place. So, how can we be sure that the latter would display the same kind of behaviour we observed in the former? How can we be sure, in other words, that the former's behaviour is not a result of the unique social environment they are part of?

In order to verify this hypothesis, a group of anthropologists went on to test Ultimatum Games and other mixed motive games in several small societies located in Asia, Africa, and South America (Heinrich et al. 2001). The results were surprising. Group differences were significantly larger here than in the previous studies involving university students, both on account of the proposers and of the responders. In some groups, rejection rates of mean offers were significantly lower than the ones observed in industrial societies. In other groups, on the other hand, rejection rates were significantly higher, and included frequent rejections of offers above 50% (2011, 75). As a consequence, it becomes significantly harder to claim, as some social scientists suggested, that cooperation evolved

in the environment of ancestral hunter-gathering societies (Fehr and Gächter 2002), when these very ancestral hunter-gathering societies show such an abysmal cooperative record, compared to their Western counterparts.

At the same time, though, the data from this cross-cultural study suggested a different hypothesis, based on a significant correlation between the differences in offers among groups and their respective degrees of market integration (Camerer 2003, 72-74; see also Chibnik 2005). In particular, the data show that societies with a higher degree of market integration show acceptance and rejection patterns which more closely resemble the ones typically observed in large-scale societies, such as the USA. This observation is arguably in line with Gaus's own analysis, as he describes the societies, which he based his justificatory model on, as large-scale societies where people frequently confront each other as strangers (Gaus 2011, 268); and these anonymous relations primarily take the form of market relations (2011, 474).

While there are certainly sensible reasons for restricting the scope of a public justification theory to large-scale societies, though, there are still some criticisms which are left unanswered. On the one hand, while the data from Heinrich et al. (2001) may not clash with Gaus's empirical account *per se*, it might be at odds with some of the other empirical sources Gaus relies on for his theory.[10] On the other hand, there is evidence for differences in people's behaviour even among market societies (Roth et al. 1991), so market integration is probably not the only variable involved in these differences.

V. CONTEXTUAL ISSUES WITH THE EXPERIMENTAL EVIDENCE: KNOWLEDGE AND ANONIMITY

There is another set of issues, which does not have to do with the identity of the subjects. Instead, it concerns the contexts in which the experiments are performed. The main contribution to Gaus's account which comes from experimental economics is, in fact, the evidence that actual people behave like rule-following punishers. In light of this conclusion, one may thus expect these studies to try and simulate an environment as close to real life as possible, to make sure that the results observed in the experiment make sense of the behaviour we are familiar with in actual social experience.

This is though not the case, at least as far as the experimental evidence considered by Gaus goes. In these experiments, the subjects perform under severe cognitive constraints, which have very little to do with anything resembling real life experience. For example, in these experiments, subjects are generally prevented from engaging in any sort of verbal communication with each other (Güth et al. 1982; Hoffman et al. 1996), and are restricted to one-shot interactions, never meeting the same player twice (Cameron 1999).

This is generally done to reduce the confounding effect of irrelevant variables which are not the focus of the study, and is necessary to understand the influence of a specific

10] See, in particular, Gaus's account of *deontic reasoning* (2011, 122-30).

factor over the focus of the study. For example, if the study is focused on testing whether the motivation behind informal sanctions is aversion to unfair results or retaliation against norm violators (such as in Falk et al. 2005), it is important to control for any other variable which could influence a person's tendency to inflict an informal sanction – such as previous experience, the other person's outlook, specific features of the environment, and so on. However, it has the disadvantage of reducing the study's ability to predict actual behaviour: even if a certain behaviour is observed in the experiment, there is no assurance that the same behaviour will be observed in an actual, real life scenario, which includes all those variables the experiment controlled for. And even if, in such a scenario, the same behaviour witnessed in the experiment were to occur, one could not tell (on the grounds of that study alone, at least) whether the basic explanation of the observed behaviour is the same as the one tested in the experiment.

Unlike the previous set of issues, this criticism is arguably less of a worry from the point of view of experimental economics. This is not because the discipline has no interest in predicting actual behaviour, but because no single study expects to test *all* the possible variables which could influence a certain behaviour. Even if a study ends up eliminating a possibly relevant variable from the context of the experiment, the impact of that variable can always be checked for in another study. The criticism does call for special care in circumscribing the scope of the outcome of a study, but it does not, all things considered, undermine the discipline as a whole, or its methodology.

Conversely, though, this very criticism poses a bigger challenge to Gaus's empirical account of social morality, and to the role played by these studies in his account. According to Gaus, the outcome of these studies applies not just under the artificial constraints of the experiments, but to actual social environments as well. However, real life includes a myriad of variables, which could either affect the behaviour of rule-following punishers or provide an alternative, stronger explanation for them than the one proposed in the studies. This worry would be somewhat lessened, if one had reason to believe that the circumstances of real life do not change significantly the behaviour of people who would otherwise behave like rule-following punishers. However, evidence from experimental economics suggests a different picture.

A study by Eckel and Grossman (2001), for example, arguably shows that gender has an influence on behaviour in ultimatum games. In this study, while other conditions of anonymity were present – such as random pairing of subjects – the subjects of the experiments were aware of the gender of the other player (even if they were unaware of the specific identity of said player). The results showed that women's offers are, on an average, slightly more generous than the ones advanced by men. Furthermore, women's offers were much less likely to be rejected than men's: even for a given offer amount, the given offered size coming from a woman was more likely to be accepted than the same size offered by a man. Women were also more likely to accept a given offer than men. Conversely, in the control group with mixed genders (so that proposers and responders

could not determine each other's gender), there were no appreciable differences between the rejection rates of the two genders.

Now, the fact that social variables may influence people's behaviour is still not exceedingly problematic for Gaus's account. As long as people are still shown to reason like rule-following punishers – that is to say, as long as they show a tendency to follow rules and punish defectors, even at a net cost for themselves – it is not too much of an issue that external factors may affect the likelihood and intensity of such punishments.

However, even this assumption is put into question by Zamir (2001). More precisely, the assumption Zamir questions here is the interpretation of Ultimatum Games and their results as a case of irrational behaviour. According to Zamir, subjects in the experiments do not tend towards fair offers because they endorse rules of fairness, even when doing so may end up thwarting their own ends. Quite the contrary, in fact, they act so because it best advances their interests, and they change their behaviour whenever this is no longer true. More specifically, what causes players to advance fair offers (around 50% of the share) is because "they pay well; *i.e.* they respond best to the rejection patterns of the responders" (2001, 19).

To put this claim to test, an experiment was conducted, where players were randomly paired for several rounds of Ultimatum Bargaining either with real players, or with virtual players – that is to say, computer programs employing fixed strategies decided at the beginning of the games.[11] Some of these programs acted like "fair players", making offers in the vicinity of 50% and rejecting any offer below that threshold. Other programs, though, acted like "tough players", more willing to accept mean offers but also more likely to offer low shares to the responders. The result was that, even when virtual players constituted a minority (around 40%) of the overall number of players, the rest of (real) players quickly adapted their behaviour within the first ten rounds of play. In environments populated by "tough" virtual players, even real players began acting "tough", as they learned that they could get away with meaner offers and they had more reason to also accept such mean shares; vice versa, the presence of "fair" virtual players strongly stabilized the average offer around the fair, 50:50 split. This goes to show, Zamir argues, that real players – far from being "irrational" or "endorsing rules of fairness" – act according to the basic rational behaviour of maximizing their income: their tendency to fairness is merely *"context dependent"* (2001, 20).

Are these results an issue, for Gaus's account? At first sight, it may appear that they are not. Indeed, one could be even tempted to see Zamir's study as the confirmation of Gaus's claim that, from the point of view of a functional approach to social morality, it is important that members of society actually endorse its rules. However, if members of society were merely rational agents, who endorse rules only when doing so maximizes

11] In some cases, participants were informed that they could be matched against virtual players, but they were unaware of how likely this match-up was, and of the nature of the computer programs. In other cases, participants did not know about the presence of virtual players altogether. Neither scenario showed significant differences from each other in its outcome (Zamir 2001, 9).

their interests, defecting to them whenever they can get away with it, they would not be the sort of members of society who could look up to rules as binding, in the sense intended by Gaus. What allows his evolutionary model to escape the Prisoner's Dilemma introduced in section one, in fact, was the assumption that members of society would endorse rules, even when doing so hinders their ability to pursue their chosen ends.

What is needed, thus, is a background theory which gives us an independent reason to believe that the actual, real life context (both in its biological and in its social dimension) actually support robust, internalized rule following. Without such a background, the results of these studies are too underdetermined to successfully support, alone, Gaus's empirical account of social morality.

VI. CONCLUSION

The use of empirical evidence in a normative argument has generally been met by political philosophy with either scepticism (Cohen 2003; Enoch 2013) or approval (D'Agostino 2013). However, both the criticism and the approval focused primarily on the logical admissibility of empirical evidence in normative discourse, in principle. This left out two important questions. First, the data Gaus presents might be insufficient to support his specific normative claims – not because of an *a priori* impossibility of employing empirical evidence to validate normative statements, but because the specific assumptions behind the data employed by Gaus are incompatible with his justificatory model. And second, it could be the case that these empirical data are flawed in the first place, and the conclusions researchers drew from them are themselves unsubstantiated.[12]

In this paper, I argued that this both concerns are warranted. On the one hand, the general scope of the claims that researchers have drawn from the aforementioned experiments is somewhat undermined by the contextual features of their experiments, such as the identity of the subjects involved. While this does not necessarily invalidate such claims, it still leaves them significantly undermined: as compelling as the hypothesis of strong reciprocity is, other explanations might be compatible with the results of these studies. On the other hand, the epistemic constraints commonly imposed to subjects in experiments make it difficult to draw conclusions that canapply to an empirical account to social morality, as Gaus assumes.

It should be noted, though, that Gaus does not rely solely on these studies, in order to defend his empirical account of social morality. A large part of his empirical evidence, in fact, depends on other disciplines, such as evolutionary psychology (Barkow et al. 1992; Cummins 1996a, 1996b) and theories of gene-culture coevolution (Boyd and Richerson 1985, 2005; Richerson and Boyd 2005), which offer an articulate explanation

12] Enoch (2013, 149) actually showed scepticism towards Gaus's claim, that there is extensive empirical evidence about the role social morality would have played in the evolution of large-scale cooperation. However, Enoch himself does not give any reference in support of his own claim, whereas Gaus's hypothesis is, indeed, supported by a significant number of sources (as noted in Gaus 2015b).

that can provide independent support for the claims advanced by experimental economics. However, on the one hand, these research fields have not been immune to criticism, either.[13] On the other hand, evolutionary psychology relies on empirical assumptions that are partially contested by theories of gene-culture coevolution (Richerson and Boyd 2005, 44-48), but the ramifications of this fact are not given any consideration within Gaus's work. Most importantly, though, other theories which could offer a plausible explanation for the evolution of culture and morality exist (Jablonka and Lamb 2005; Sterelny 2003, 2012).

The evolutionary approach to public reason developed by Gaus offers a valuable model of explanation for the justification of morality and social norms. As I have argued in §1, it is based on a moderately idealized account of members of society, which arguably constitutes a preferable alternative to the highly idealized model advanced by Rawls (1996). Moreover, unlike the instrumentalist approach developed by Gauthier, Gaus does not rely on a controversial notion of rationality, but rather on an appeal to evolutionary forces that can find support in extensive empirical evidence. However, for this evolutionary approach to overcome its weaknesses and flaws, a more comprehensive and critical assessment of its empirical account may be required.

lucagenova@fastwebnet.it

REFERENCES

Baier, Kurt. 1995. *The Rational and the Moral Order. The Social Roots of Reason and Morality.* La Salle IL: Open Court.

Barkow, Jerome H., Leda Cosmides, John Tooby. 1992. *The Adapted Mind.* New York: Oxford University Press.

Bicchieri, Cristina. 2005. *The Grammar of Society. The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Bolton, Gary E., Ockenfels Axel. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review* 90 (1): 166-93.

Bolton, Gary E., Rami Zwick. 1995. Anonymity versus Punishment in Ultimatum Bargaining. *Games and Economic Behavior* 10: 95-121.

Boyd, Robert, Peter J. Richerson. 1985. *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

———. 2005. *The Origin and Evolution of Culture.* Oxford: Oxford University Press. Buller, David J. 2005. Get Over: Massive Modularity. *Biology and Philosophy* 20: 881-91.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton: Princeton University Press.

Cameron, Lisa A. 1999. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry* 37 (1): 47-59.

13] For a criticism of the inheritance mechanisms assumed by theories of gene-culture coevolution, see Jablonka and Lamb (2005, ch. 6). For a criticism of evolutionary psychology, with particular attention to its massive modularity hypothesis, see Sterelny and Griffiths (1999), Sterelny (2003), Buller (2005). For a more general criticism of evolutionary psychology, see also Dupré (2001).

Chibnik, Michael. 2005. Experimental economics in anthropology: A critical assessment. *American Ethnologist* 32 (2): 198-209.
Cohen, Gary A. 2003. Facts and Principles. *Philosophy and Public Affairs* 31 (3): 211-45.
Croson, Rachel, T. A. 1995. Information in ultimatum games: An experimental study. *Journal of Economic Behavior & Organization*, 30: 197-212.
Cummins, Denise D. 1996a. Evidence of deontic reasoning in 3- and 4-year-old children. *Memory & Cognition* 24 (6): 823-29.
———. 1996b. Evidence for the Innateness of Deontic Reasoning. *Mind & Language* 11 (2): 160-90.
D'Agostino, Fred. 2013. The Orders of Public Reason. *Analytic Philosophy* 54 (1): 129-55.
Dupré, John. 2001. *Human Nature and the Limits of Science.* Oxford: Clarendon Press.
Dworkin, Ronald. 1973. The Original Position. *The University of Chicago Law Review* 40 (3): 500-33.
Eckel, Catherine, Philip Grossman. 2001. Chivalry and solidarity in ultimatum games. *Economic Inquiry* 39 (2): 171-88.
Enoch, David. 2011. *Taking Morality Seriously. A Defense of Robust Realism.* New York: Oxford University Press.
———. 2013. The Disorder of Public Reason: A Critical Study of Gerald Gaus's The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World. *Ethics* 124 (1): 141-76.
Falk, Armin, Ernst Fehr, Urs Fischbacher. 2000. Informal Sanctions. *IEER Working Paper* No. 59.
———. 2005. Driving Forces Behind Informal Sanctions. *IZA Discussion Paper* No. 1635. Fehr, Ernst, Simon Gächter. 2002. Altruistic Punishment in Humans. *Nature* 415 (6868): 137-40. Fehr, Ernst, Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114 (3): 817-68.
Forsythe, Robert, Horowitz Joel L., Savin N. E., Sefton Martin. 1991. Fairness in Simple Bargaining Experiments. *Games and Economic Behavior* 6: 347-69.
Gaus, Gerald. 1996. *Justificatory Liberalism: An Essay on Epistemology and Political Theory.* New York: Oxford University Press.
———. 2011. *The Order of Public Reason. A Theory of Freedom and Morality in a Diverse and Bounded World.* New York: Cambridge University Press.
———. 2015a. On Dissing Public Reason: A Reply to Enoch. *Ethics* 125 (4): 1078-1095.
———. 2015b. The Egalitarian Species. *Social Philosophy and Policy* 31 (2): 1-27.
Gauthier, David. 1986. *Morals by Agreement.* New York: Oxford University Press.
———. 1991. Why Contractarianism?. In *Contractarianism and rational choice: essays on David Gauthier's Morals by agreement*, edited by in P. Vallentyne, New York: Cambridge University Press, pp. 15-30.
———. 1994. Assure and Threaten. *Ethics* 104 (4): 690-721.
Güth, Werner, Rolf Schmittberger, Bernd Schwarze. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3: 367-88.
Harrison, Glenn W., Kevin A. McCabe. 1996. Expectations and Fairness in a Simple Bargaining Experiment. *International Journal of Game Theory* 25: 303-27.
Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath. 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* 91 (2): 73-78.
Hoffman, Elizabeth, McCabe Kevin, Shachat Keith, Smith Vernon. 1994. Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior* 7: 346-80.
———. 1996. On Expectations and The Monetary Stakes in Ultimatum Games. *International Journal of Game Theory* 25: 289-301.

Jablonka, Eva, Marion J. Lamb. 2005. *Evolution in Four Dimensions. Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life.* Cambridge, MA: MIT Press. Kitcher, Philip. 2011. *The Ethical Project.* Cambridge MA: Harvard University Press.

Larrick, Richard, Sally Blount. 1997. The Claiming Effect: Why Players Are More Generous in Social Dilemmas Than in Ultimatum Games. *Journal of Personality and Social Psychology* 72 (4): 810-25.

List, John A., Todd L. Cherry. 2000. Learning to Accept in Ultimatum Games: Evidence from an Experimental Design that Generates Low Offers. *Experimental Economics* 3: 11-29.

Mill, John Stuart. 1963. On Liberty. In *The Collected Works of John Stuart Mill*, vol. 2 and 3, edited by in Robson J. M., Toronto: University of Toronto Press.

Moore, George E. 1903. *Principia Ethica.* Cambridge: Cambridge University Press.

Rapoport, Amnon, James A. Sundali, Darryl A. Seale. 1996. Ultimatums in two-person bargaining with one-sided uncertainty: Demand games. *Journal of Economic Behavior & Organization* 30: 173-96.

Rawls, John. 1996. *Political Liberalism.* New York: Columbia University Press.

———. 1999. *A Theory of Justice. Revised Edition.* Cambridge MA: The Belknap Press of Harvard University Press.

Richerson, Peter, Robert Boyd. 2005. *Not by Genes Alone. How Culture Transformed Human Evolution.* Chicago: The University of Chicago Press.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, Shmuel Zamir. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *The American Economic Review* 81 (5): 1068-1095.

Ruse, Michael, Edward O. Wilson. 1986. Moral Philosophy as Applied Science. *Philosophy* 61: 173-92.

Schotter, Andrew, Avi Weiss, Inigo Zapater 1996. Fairness and survival in ultimatum and dictatorship games. *Journal of Economic Behavior & Organization* 31: 37-56.

Skyrms, Brian. 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press. Slonim, Robert and Alvin E. Roth. 1998. Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica* 66 (3): 569-96.

Smith, John M. 1978. The Evolution of Behavior. *Scientific America* 239: 176-92.

Sterelny, Kim. 2003. *Thought in a Hostile World. The Evolution of Human Cognition.* Malden MA: Blackwell.

———. 2012. *The Evolved Apprentice. How Evolution Made Humans Unique.* Cambridge MA: The MIT Press.

Sterelny, Kim, Paul E. Griffiths. 1999. *Sex and Death. An Introduction to Philosophy of Biology.* Chicago: The University of Chicago Press.

Strawson, Peter F. 1961. Social Morality and Individual Ideal. *Philosophy* 36 (136): 1-17.

Thornhill, Randy, Nancy Wilmsen Thornhill. 1992. The Evolutionary Psychology of Men's Coercive Sexuality. *Behavioral and Brain Sciences* 15: 363-421.

Zamir, Shmuel. 2001. Rationality and Emotions in Ultimatum Bargaining. *Annales d'Économie et de Statistique* 61: 1-31.