Werner Güth[*] and Hartmut Kliemt[**]

What Ethics Can Learn From Experimental Economics – If Anything

Abstract

Relying on the specific example of ultimatum bargaining experiments this paper explores the possible role of empirical knowledge of behavioural "norm(ative) facts" within the search for (W)RE – (Wide) Reflective Equilibria on normative issues. Assuming that "pro-social" behaviour "reveals" moral orientations, it is argued that these "norm-facts" can and should be used along with stated preferences in justificatory arguments of normative ethics and economics of the "means to given ends" variety. At the same time behavioural norm-facts are so heterogeneous that any hopes to reach an inter-personally agreed (W)RE in matters moral seem futile.

Key words: Meta-Ethics, Experimental Economics, "Methods of Ethics", Reflective Equilibrium

> "On any theory, our view of what ought to be must be largely derived, in details, from our apprehension of what is; the means of realising our ideal can only be thoroughly learnt by a careful study of actual phenomena; and to any individual asking himself 'What ought I to do or aim at?' it is important to examine the answers which his fellow-men have actually given to similar questions." (Sidgwick, 1907/1981, 2)

1. Introduction and overview

Lionel Robbins' "essay on the nature and significance of economic science" (Robbins, 1935) is a modern classic. Robbins believes that normative economic argument is constrained to pointing out "means to given ends". Rational normative argument is merely "hypothetical" or "relative" to ends. It has the general form: "if you want x you should do y!" A hypothetical imperative applies merely "to whom it concerns". It becomes irrelevant for its addressee whenever the ends, aims, or values to which it must be fixed cease to hold.

[*] Max Planck Institute of Economics, Jena, gueth@econ.mpg.de [**] Frankfurt School of Finance and Management, h.kliemt@frankfurt-school.de. We would like to express our gratitude to the CESifo institute, Vesa Kanniainen and Manfred J. Holler for organizing the excellent conference on the relationship of ethics and economics at which this paper was presented. We also thank the participants of the conference, anonymous referees and in particular Matthew Braham, Manfred J. Holler and Bernd Lahno for their very helpful criticisms and suggestions that led to major revisions – and as we hope – improvements of this paper.

To transcend the merely “subjective” character of “means to given ends” arguments some economists turned to philosophy as a source of “inter-subjectively valid” ends, aims or values. However, such economists might be reminded that a large group of philosophers concurs with Robbins in assuming that “an ought” presupposes “an is”. They accept that the argument that something should be done is relative to the *fact* that its addressees do as a matter of fact aspire to an appropriate aim, end or value (see on such a Humean perspective Hume, 1739/1978, Mackie, 1980, Hardin, 2007, Kliemt, 1985). In line with this “meta-ethical” view – a view *about* the limits of rational justification of substantive ethical claims – these philosophers all endorse on the *justificatory* level what may be called an “ethics in terms of hypothetical imperatives” (means to ends).

Endorsing the meta-ethical view that ethics should be framed in terms of hypothetical imperatives it becomes an obvious task of ethics to find out more about the prevalence of “normative facts” like the aims, ends, values as well as accepted norms and practices. Many disciplines could contribute insights concerning “norm(ative) facts”. Since a survey of findings is beyond the scope of a single paper our aims are much more modest. Intrigued by the question of what “laboratory experiments measuring social preferences *reveal* about real moral problems?” (Levitt and List, 2007, italics added) we restrict ourselves to the specific example of the ultimatum game. With a stylized account of experimental results on this game in hand, we ask the central methodological question how these findings could conceivably contribute to the “piece-meal” formation of prescriptive ethical theory.

We start with a brief rehearsal of existing proposals to frame ethical deliberation in close parallel to scientific deliberation (2.). Next we introduce a class of interactions which seem to be particularly interesting with respect to justice related behaviour (3.) and sketch some central insights into justice related normative facts that can be derived especially from ultimatum experiments as models of a specific sub-class of those interactions (4.). We then discuss to what extent the search for intra- as well as inter-personal coherence concerning matters of justice can make sense despite the striking heterogeneity of factual justice related behaviour (5.). We conclude with some remarks on utilizing experimental knowledge within a bounded rationality framework to form conceptions of "bounded justice" (6.).

2. The wide reflective equilibrium approach to ethics

Though other philosophers (e.g. Sidgwick, 1907/1981) have interpreted their own practices along the same lines the reflective equilibrium metaphor as used in more recent practical philosophy is rooted in Rawls' "Outline of a Decision Procedure for Ethics" (Rawls, 1951). The "procedure" is based on the assumption that in forming ethical theories it matters what individuals do accept from their moral points of view. As initially outlined it is a search process in which basic intuitive judgements concerning specific problems, "(judgemental) normative facts" so to say, are used to "derive" general ethical principles. Later Rawls embedded his theory of justice (Rawls, 1971) into the justificatory framework described by his reflective equilibrium metaphor RE. In this step the original unidirectional search for general principles became circular in that established general principles could also lead to revisions of specific judgements to eliminate incoherencies preventing the emergence of an equilibrium.

Searching for a RE an individual tries to construe a coherent view of the world in which his specific intuitive judgements concerning specific matters (either of a factual or a normative nature) cohere with his general principles. For instance, somebody observing the killing of a human being might intuitively consider the act as (morally) wrong concluding that it "should not be done". He might support this specific view by invoking a general principle "that the killing of human beings is morally wrong and is forbidden".

Assume, he is content to let it rest with that until confronted with a case of self-defence. Observing this he might be put out of his temporary reflective equilibrium since he may intuitively be convinced that in this case killing was justified. So he has to modify his general principle to the effect that he now believes that "killing of human beings is morally wrong unless in cases of self-defence". However, after a while, he may be confronted with an example in which somebody attacks another individual with a means that is recognizably not live threatening – say the content of a box of raw eggs. In this case he may have to specify the general principle even further to reach reflective equilibrium again.

After many precedents the principle prohibiting killings may be well specified. Then it will typically not only be supported by many single cases but also support specific rulings against intuitive judgements. If for instance self-defence against an attack with a weapon is legitimate according to the general principle and he observes that somebody attacks another person with an air pressure gun he might nevertheless believe initially that killing in self-defence is not justified in the case at hand because the attack is not life-threatening. However, the general principle may be so well entrenched in the reflective equilibrium he had reached before that it now wins out against the specific intuitive judgement. In the effort to reach coherence and a reflective equilibrium it is accepted on reflection that killing an attacker who comes with an air pressure gun can be a legitimate act of self-defence. This leads to a revision of the initial specific intuitive verdict and a new reflective equilibrium is reached.

Like a Cournot-Nash equilibrium in which behavioural expectations support each other a reflective equilibrium, RE, is like a Roman arch in which each of the stones is held in position by each other only. A wide reflective equilibrium, WRE, emerges if into the search for equilibrium more general background theories are included (see Rawls, 1974, Daniels, 1996, Hahn, 2000). The broader theories are factors exogenous to and besides the basic judgements that otherwise form the basic material for supporting the general principles and considerations. Following this lead we believe that besides general background theories further non-judgmental normative facts like established practices and results of experiments on human behaviour should be included in the search for reflective equilibria.

Since the whole point of the search for (wide) reflective equilibrium is to make normative ethical argument as similar as possible with scientific argument it is useful to reconstruct scientific procedure in the light of the reflective equilibrium metaphor (see Goodman, 1978 and again Hahn, 2000). After we have done so in a very rough and stylized way (2.1) we shall turn to the crucial issue whether basic judgments form the correct basis for reflective equilibrium or whether other kinds of normative facts should be used as well (2.2).

2.1. The search for an equilibrium in scientific methodology

In former times norms of good scientific practice were developed in a top down approach. Typically they evolved more or less on the basis of a priori arguments out of some epistemologically motivated philosophical conception. Such philosophical conceptions still do play some role. Yet nowadays due respect for established scientific practice – or for what the sciences in fact do – serves as the main springboard for normative methodological considerations of what the sciences should do. This leads to an a-posteriori or experience-based process of developing "norms" of "good scientific practice" out of a stylized account of scientific practice itself (see Fleck, 1935/1980, Kuhn, 1962, Lakatos, 1978; and closer to experimental economics and to the experiments discussed below, Binmore and Shaked, 2007).

The process of finding "best practice standards" is to some extent circular. It starts with a specific practice that prevails in the realm of science. To serve as authoritative evidence the practice must, first, be classified as “successful” according to some very broad evaluative standard. Then, second, a stylized account of the practice is given, or as philosophers tend to say, it is “rationally reconstructed”. Third, certain aspects are identified as likely causes of success. These are, fourth, presented in an idealized or stylized form to serve as a basis of normative standards of “good” science.

To put the same thing slightly otherwise, the established scientific practice gains a special normative status simply because it is an accepted established practice that is deemed successful by the “practitioners” and reasonably informed external observers. Though as in other cases of so-called “bench-marking” successful practices (auto-)determine what good practice is, an established practice can be corrected in a piece-meal way by the very generalizations and norms that are developed out of observations of that practice. In this broadly “critical rationalist” (Albert, 1978) – though somewhat more coherentist – account of scientific procedure the quest for substantial coherence is the driving force of the “rationalizing” process which can go back and forth between the general and the particular until coherence is reached.

Though it may temporarily come to a halt, reflection can nevertheless always start all over again. The search for reflective equilibrium will stop only temporarily once a “sufficient” level of coherence – meeting some aspiration level of coherence – is met (in the spirit of Simon, 1957, 1985). In this as in other aspects the approach is merely an idealized form of daily trial and error practices of justifying judgement on issues of scientific practice.

As far as rationality requirements are concerned, what is good enough for the paradigmatically “rational” practice of science should be good enough for other human endeavours. So against the background of the “bench-mark” of scientific practice let us turn to the analogous justificatory method (or metaphor) that Rawls proposes for the purposes of justifying normative judgements.

2.2. The quest for substantive normative coherence

In search for a coherent norm system of general and specific judgments the approach must start with some basic "normative facts". Traditionally these facts have been taken to be basic "normative *judgements*". For instance in the most simple case of his search for a personal reflective equilibrium on matters of justice Rawls relied basically on introspective evidence. He wondered what he himself – and, as he implicitly speculated, other competent addressees of his argument – would find intuitively appealing (see on the concept of competence Hoerster, 1977).

As Richard Hare objected early on, this seems a bit too much of circularity: "Rawls' POP [people in the original position] come to the decisions that they come to simply because they are replicas of Rawls himself … It is not surprising, therefore, that they reach conclusions that he can accept" (Hare, 1973, p. 249). And, with Frohlich and Oppenheimer, we may add that "(t)he traditional philosophical methodology for dealing with justice has called for introspection and argument about these issues. We believe that this narrowly introspective approach has limited progress in the field of ethics because it has not allowed philosophers to introduce the diversity and fine details to obtain the balance sought. For that a broader strategy is needed." (Frohlich and Oppenheimer, 1992, 2-3)

As part of their "broader strategy" Frohlich and Oppenheimer send the impartial spectator to the laboratory. This takes normative facts of real world practices more seriously than Rawls' arm chair empiricism. However, it is still rather close to the original Rawlsian ways. In particular, the participants of Frohlich's and Oppenheimer's experiments are exposed to "impartiality situations" in which they operate behind some veil of uncertainty. Individuals who do intend to deliberate from a moral point of view will accept this experimentally imposed veil as expressing their impartial intentions. It induces them to take into account all social positions in their joint deliberations in the laboratory.

As opposed to the Rawlsian veil of ignorance which is merely a tool in the search for RE and entirely fictitious the veil of uncertainty about their own later positions is real for the participants in Frohlich's and Oppenheimer's experiments. The experimental set up enables them to test the acceptability of moral principles by means other than introspection. The test is factual acceptance under idealized conditions. The individuals are forming their opinions in communicative situations of joint deliberation.[1] In the situations specific strategic aspects play a role because agreement must be found under a Buchanan type unanimity principle (see Frohlich and Oppenheimer, 1992, p. 28, p. 40) endowing every participant with veto-power (see critically on the ethical status of this endowment, Kliemt, 1994) yet the veil of uncertainty sees to it that the moral point of view is also taken into account.

As the experiments show it is not outrageously optimistic to expect human actors to agree unanimously in a "calculus of consent" (Buchanan and Tullock, 1962) manner on a constitutional decision as dictated by "the reason of rules" (Brennan and Buchanan, 1985). This is a possible way of framing decision making on ethical principles (in reasonably small groups). It can generate useful information concerning those given aims, ends, or values that represent the moral point of view of an individual. However, the analysis is biased in favour of the moral point of view. It is, so to say, built into the situation via rules introducing uncertainty and veto power etc. Contrary to that we suggest to confront theories of justice with real justice related behaviour in situations in

[1] Alluding to the fashion of our day, one might also refer to it as "deliberative democracy in the lab" operating under special knowledge and agreement conditions; for a collection, see Elster, 1998.

which impartiality along with partiality is operative in a more unbiased way. In line with Frohlich and Oppenheimer and on a route starting in Sidgwick we intend to “go for” revealed rather than merely stated “preferences” (see Louvierre et al., 2000) in search for RE. – Rather than discussing this in terms of philosophical abstractions we will try to outline what we have in mind by means of discussing a specific, exemplary class of interactions and their representations by game models.

3. Proposer-responder interactions

In the simple games of proposal and response that we consider here as models of a class of dyadic proposer-responder interactions, two actors, a proposer and a responder, can split a fixed sum or pie p (see also Manski, 2002, 883). Proposer X assigns shares $x,y \geq 0$ of the pie such that $x + y = p$. If assigned as proposed the share of the proposer X will be x, while y will be the share of the responder Y. After learning what the proposal (x, y) is, responder Y can accept or reject the proposal.[2] If she accepts, then the rewards are assigned as proposed to the two participants. Should the responder *reject* the proposal then the rewards will be (αx, βy) with $1 \geq \alpha, \beta \geq 0$.

First, look briefly at the four extreme parameter combinations (α, β), $\alpha, \beta \in \{0,1\}$. If α=β=1 then independently of its acceptance or rejection by the responder the reward allocation will be (x, y). The response of the responder is completely inconsequential for the material or substantive payoffs of both participants since the proposer is in a *dictatorial* position. If α=0 and β=1 then the responder can reject a proposal without forgoing any material payoff to herself. Her acts are substantially (as measured in material payoffs) inconsequential for herself while maximally consequential for her co-player. The first moving individual’s offer y is like a *bribe* for a responder Y who after the offer may “hit and run” or leave x to X. If, however, α=1 and β=0 expressing, say, “responder resentment” will

[2] There are also yes/no experiments in which the responder is not informed about (x, y), see Gehrig et al., 2007.

have no direct monetary impact on the proposer, while – relative to the proposal y – it is maximally consequential for the responder to say no. The proposer can do whatever he chooses with *impunity*. Finally, with α=0 and β=0 the responder can express her “resentment” but only at the full cost of entirely forgoing y. Her rejection of the proposal is maximally consequential for both proposer and responder since it will transform the proposed outcome (x, y) into the realized one of (0, 0).

The games emerging from the extreme parameter constellations have been studied experimentally in ways akin to searching for a reflective equilibrium in explaining behaviour in proposer-responder interactions. The territory of intermediate, non-extreme parameter settings has not been (and cannot be) fully explored. To take samples from the full range of intermediate cases $\alpha, \beta \in (0,1)$ of such games (Suleiman, 1996) should allow for even finer discriminations between possible normative convictions of individuals than already known (see for an summary account Bolton et al., 2008). This would help to identify the full range of normative principles actually guiding behaviour and thereby help to do justice to the complexities of real human moral behaviour.

Since our aims are methodological rather than substantive we are content to let it rest at that and merely sum up the preceding in the tabular form of Table 1.

Overview over the parameter constellations that give rise to different types of justice related interactions in simple proposer responder games

I. First mover X (proposer), second mover Y (responder)

pie, p,
proposal by X, (x, y) with $x+y=p$, addressed at Y
If Y accepts → (x, y) as payout
If Y vetoes → $(\alpha x,\beta y)$ as payout

II. Games that emerge from extreme parameter constellations

1.	$(\alpha=1,\beta=1)$	→	Dictator game
2.	$(\alpha=1,\beta=0)$	→	Impunity game
3.	$(\alpha=0,\beta=1)$	→	Bribe game
4.	$(\alpha=0,\beta=0)$	→	Ultimatum game

Table 1

We believe that including as much as possible of experimental knowledge of normative facts into the search for WRE must be the aim of substantive ethical theory formation. In principle a systematic exploration of all the proposer responder games of the preceding table should take place in search of an adequate descriptive and a supervenient normative moral theory for such situations. The great merit of a parametric presentation as the preceding is that other than in conventional ethical deliberations it creates a vision of the complete space of specific situations that may matter. This is clearly superior to leaving the invention of specific test situations for general normative principles simply to the philosophers. Experience teaches that they will be very creative in devising extreme or hard cases as tests. But it seems mistaken to assume that the full possibility space of more conventional cases is simply a kind of convex combination of extremes. Nor is it a good idea to build too much general theory on extreme or hard problems. We should recall that almost proverbially “hard cases make bad law". In all likelihood hard moral cases will lead to mistaken accounts of the working morals of boundedly rational individuals as well.

The testing that experimental economic modelling can provide forms an invaluable contribution on the way to a "moral science" that systematically rather than impressionistically explores normative systems. Of course, the knowledge of the "is" of "normative facts" will not directly yield prescriptive judgements. Yet the exploration of full parameter spaces of typical interaction situations, like proposer-responder interactions, voluntary contribution mechanisms etc., will form a much richer and more adequate basis for normative ethical theory formation than the existing judgement based methods. At least that is what we believe as subscribers to the quasi empirical methodological precepts of normative theory formation captured by the Rawlsian reflective equilibrium search metaphor.

The normative principles guiding behaviour in any of the proposer-responder interactions must always be understood in the context at least of the full class of such situations. For instance, running a dictator or an impunity game with similar payoffs as those of an ultimatum game is significant for understanding the motives of the proposer and the responder respectively in such interactions. Bearing this in mind it may nevertheless seem desirable to look at an exemplary parameter constellation and to ask what we can conceivably learn from experimenting on it. For the sake of specificity we will look at the extreme parameter constellation $\alpha=0,\beta=0$ of the ultimatum game. We use this example not merely for convenience of exposition. It is particularly well explored in a rather broad and varied literature.[3] Moreover it has a close relation to some central insights of philosophical and anthropological moral theory. For, the role of retributive dispositions and emotions that shows up so clearly in the

[3] Studies concerning such experiments in different societies can be found in Henrich et al., 2004. Overviews are given in Güth and Tietz, 1990, Roth, 1995 and more recently Camerer, 2003.

ultimatum game experiments has traditionally been identified as crucial for the proper workings of moral (and, for that matter, legal) institutions in general (see Mackie, 1982, Westermarck, 1906).

4. Stylized normative facts in ultimatum experiments

That some restriction of rational opportunism must be present in ultimatum game experiments is clear from observed rejections of substantial positive offers in such games. Since such "norm-bounded" behaviour violates what may be called the "efficiency axiom" it is not necessarily "moral" in the full sense of that term. However, it is at least akin to moral behaviour and therefore forms a natural starting point for an empirically based ethical study of the phenomenology of moral behaviour.

4.1. A brief account of responder behaviour

In "standard" ultimatum experiments (i.e., if $\alpha=0$ and $\beta=0$) many if not most responders (consistently more than 50%) reject offers, y, in the range $p/3>y$ and even more distinctively so if offers fall below 20% of the fixed pie "p". As questionnaires show this applies to responders who are fully aware that the interaction is anonymous. They seem to understand, too, that with practical certainty they never will interact again with the same proposer.[4] If so, any rationalization of the observed behaviour in terms of objective external incentives or extrinsic motivation is ruled out. Participants must be bound by some kind of intrinsic motivation to behave in the way they do regardless of the fact that their behaviour will not have any external causal consequences that will affect *themselves*.

The violation of the basic economic assumption of opportunism motivated by substantive payoffs is obvious. Less obvious and more interesting is it to find out why retributive behaviour is shown (expressed) in some instances while not

[4] There has always been a certain amount of scepticism concerning this premise since individuals might have deeper gut feelings adapted to repeat interaction. For a small group context see Huck and Oechssler, 1999.

in others: Why do actors actually yield to a retributive impulse sometimes and sometimes not?

One reason for behavioural differences might be that (for x, y>0) "punishment efficiency", x/y, matters to actors with retributive emotions who retain a sense of the costs involved.[5] The value of x/y=(p-y)/y would be increasing with decreasing y. However, in cases in which even "too" high offers have been rejected punishment efficiency can hardly be the motive. Another possible argument might be that actors want to express in some way or other their resentment against violations of equity per se. Deviations can go beyond the limits of the tolerable in all directions. What is tolerable is fixed by a kind of aspiration level located in a neighbourhood around the equitable solution of the problem. If the results are deemed intolerable the retributive emotion is aroused. It must somehow find a way to express itself. And it is expressed at the cost y>0, or so the speculative argument may run.

This reading is supported by experiments with impunity games (α=1 and β=0) in which regardless of the absence of punishment options responders nevertheless chose to reject too low offerings to themselves. It seems also to corroborate such an interpretation that offering an additional cheap talk option, in which individuals in the responder role could voice their complaint to a proposer by whom they felt unfairly treated, reduced rejection rates considerably (Xiao and Houser, 2005, Güth and Levati, 2007). But sticking to the (α=0 and β=0) constellation of the ultimatum case here let us simply note that the preceding sketch indicates how in principle finer morally relevant distinctions can be scrutinized experimentally and that the information is of obvious ethical interest.

[5] For a given proposal (x, y), x,y>0, the influence of punishment efficiency might be tested across games $(\alpha,\beta)\in(0,1)\times(0,1)$ by letting $\alpha\to 0$ or $\beta\to 0$ and considering $\alpha x/\beta y$.

4.2. Proposer behaviour

As opposed to much of responder behaviour the behaviour of proposers can at least conceivably be in line with the consequentialist forward looking rationality concept of standard economic theory. Though it seems rather clear that the individuals do not form beliefs and expectations along the lines suggested by expected utility theory they may still be acting in view of expected future consequences. And we think they often do act in such a "teleological" manner, yet only in a boundedly rational way.

If proposers go about their decision-making in terms of basic rules of thumb which express expectations about acceptance or rejection by the co-player, the next Figure 1 presents in a stylized way what is going on.

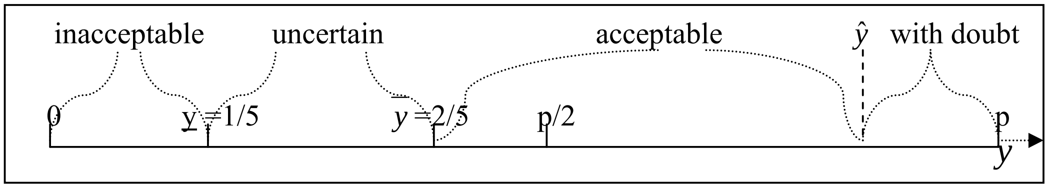

Figure 1

Assuming that $0<\underline{y}<\bar{y}<\frac{p}{2}$ we can offer the following comments on behaviour and its likely motives in the different intervals:

$y\in\left[\frac{p}{2},p\right]$, proposal $y<\hat{y}$ is expected to be accepted but since for $y\geq\hat{y}$ offerings may increasingly appear like charitable givings there may be some doubt about responder behaviour in case of offers close to p.

$y\in\left[\bar{y},\frac{p}{2}\right)$, in this realm X expects proposal y to be accepted by Y.

$y\in\left[\underline{y},\bar{y}\right)$, proposal y is such that the proposer just does not know what to expect.

$y \in \left[0, \underline{y}\right)$, proposal y is expected to be rejected.

When $y = \frac{p}{2}$, proposal y is expected to be accepted with practical certainty. This is in some Non-Bayesian way qualitatively different from all the other assignments. Likewise the inability of a decision maker to say what to expect in the range of $y \in \left[\underline{y}, \overline{y}\right)$ is clearly not in line with common Bayesian assumptions. Genuine uncertainty rather some probabilistic uncertainty prevails.

It is true, expected value formation could also explain statistical observations of real proposer behaviour. Assuming that individuals endorse heterogeneous beliefs about responder behaviour roughly the same statistics might emerge. Nevertheless, there seems to be convincing evidence that the model of boundedly rational decision making is more faithful to the cognitive processes underlying actual human behaviour than the expected value hypothesis.

Though it seems safe to rule out the expected value hypothesis as a potential explanation, additional experiments are necessary to understand the complexities of (in a wide sense) boundedly rational "moral" motivation more fully.[6] The results of such additional empirical research can, of course, not be foreseen. It is a rather safe prediction, too, that heterogeneity between individuals will persist. The same holds good with respect to cultural and situational differences. The generalization from one situation to others will raise additional complicated problems which may require the ability to make prudent judgements in one way or other. We have to take these facts of decision making into account when seeking a reflective equilibrium on justice or equity related matters rather than to insist on some streamlined rational choice model.

[6] In the present case all the reservations laid out in Binmore and Shaked, 2007, would kick in. It would perhaps be necessary to design a sequence of experiments to identify which of the rules of boundedly rational choice making are operative.

5. Towards behavioural reflective equilibrium?

Since a simple description and explanation of factually observed behaviour would not help in a justificatory enterprise like the search for (W)RE an additional step is needed: The observed behaviour must be put into a (normative) "rule" perspective. To that effect we need to postulate norms and rules such that an individual accepting those rules and norms as standards of her own behaviour would plausibly show the observed behaviour.

Clearly, norms cannot be tested directly against behavioural facts. Yet it can be tested whether or not the behaviour that should be shown according to the rules and norms imputed to the actors is in fact shown. The crucial "bridging-claim" is: If actors would follow the rules *from an internal point of view* (Hart, 1961) then they would show overt behaviour of a certain kind. If an actor who allegedly adopts an internal point of view to certain rules does not show the corresponding overt behaviour this falsifies the ascription of the normative theory as an accepted standard of behaviour – at least to some extent. The actor reveals in overt behaviour that she either follows different rules, that the opportunity costs of rule-following are too high etc.

5.1 Intra-personal incoherence in ultimatum game experiments

Let us start with intra-personal "heterogeneity" which seems to express itself in a kind of "role incoherence". For instance, if the proposer in an experiment employing the strategy vector method (as used for instance in the newspaper experiments Güth et al., 2003, Güth et al., 2007) is personally inclined to accept meagre offers as a responder and at the same time is willing to offer an equal split of the pie as a proposer, such behaviour, at least at first sight, seems to violate certain principles of (across) role coherence. Should not the morally coherent individual act in ways that would lead to the same result if the individual would adopt both roles in the ultimatum game?

Some ethical theorists as well as pedestrian moralists seem to tend to such views. They would require that an individual should in the proposer role offer what the individual would accept in the responder role and demand in the responder role not more or less than the offer the individual would make in the proposer role. However, already the very first ultimatum game experiment indicated that some participants would not have accepted their own proposal in the responder role. The offers of others would have gone well beyond the threshold demanded by them in the responder role.[7]

A moral philosopher who, faced with such results, argues that the consideration of trade offs between moral and other motives contaminates moral analysis behaves like the rational choice economist who intends to form a theory of rational behaviour without paying due respect to the facts and practices of actual boundedly rational behaviour. Akin to his “brother in guilt”[8] such a moral philosopher may want to render his claims definitional truths by identifying moral behaviour as being motivated by respect for the moral law per se. Such a move may be fine for the Kantian “homo noumenon”. Yet the morals of the “homo phenomenon” (Kant, 1798/1977) cannot abstract away everything but the “moral dimension”. To come closer to the boundedly rational and moral reasoning we must acknowledge that more often than not, there is in fact a trade off between requirements of impartiality and partiality. Day to day morals do not require that we grant no weight to motives other than moral ones. It requires that we give moral motives “acceptable” weight. This fits neatly not only with notions of boundedly rational (satisficing) behaviour it coheres well also with commonly accepted requirements to help others if that can be done at low costs

[7] See again Güth et al., 1982.
[8] See Sugden, 2004.

to oneself[9] while treating such acts as supererogatory if their performance implies costs beyond certain thresholds.

If we include trade offs then behaviour in the proposer and the responder role that otherwise seems incoherent may become quite coherent. Opportunism applies differently in the roles of the proposer and the responder and therefore different trade offs may seem justified.

5.2 Inter-personal (in)coherence in ultimatum game experiments

The existence of consent – or, for that matter, homogeneity – is not supported by observations of pro-social behaviour in the laboratory. There seems to be one exception, though. In the original class of experiments of the ultimatum game type an equal split by the proposer is seen clearly as unobjectionable by practically all in the responder role. So, perhaps here we can identify some minimal moral consensus on equality? Yet, it should be noted that the original situation is framed such that the pie comes as a kind of gift (manna). If the pie had to be earned in a preceding round of interaction then a proportionality norm would have kicked in (Hoffman et al., 1994). Likewise had there been some individual with special needs an equal split might have been rejected.

The moral philosopher may want to draw attention here to an Aristotelian version of proportional assignments of which the equal splits observed form a special case (see Frankena, 1966). If there was no preceding round of interaction in which the pie was rendered available both actors were equal in their (then zero) contributions. By imposing anonymity the individuals were made artificially equal in all other regards. Therefore proportionality would suggest an equal split of the pie as a special case of a proportionality norm, or so the argument might run (see recently Bergh, 2008, also on the role of entitlements).

[9] As opposed to Anglo-Saxon law not merely a moral but a legal obligation under German law. Particularly instructive on this Frellesen, 1980.

5.3 Behavioural heterogeneity in ultimatum game experiments

If one looks at actual raw data rather than statistical aggregates that often conceal rather than reveal it, then in economic experiments heterogeneity is “all over the place” (see e.g. Güth et al., 2001). On average, behaviour may be of a certain kind and the averages may be similar across time and place. However, for those who, as for example the contractarians do, emphasize respect for the separateness of persons averages do not matter, individuals do. At least prima facie the presence of different behavioural types indicates heterogeneity of “deeper” normative orientations. The notion of a consent or unanimity of “practically all” becomes a rather “other worldly” ideal that does not seem to have real applications.

We believe that for the moral philosopher in general and the applied ethicist in particular the demonstration of widespread heterogeneity is presumably the most relevant lesson from experiments, including those of the ultimatum type. There is pro-social behaviour but no homogeneity of the type of that behaviour unless artificially created. If ethicists seek to find agreement outside their fictitious worlds or settings like that of Frohlich and Oppenheimer they will seek in vein. For, if even in simple experiments homogeneity of ethical views on justice and equity does not show itself in homogeneous behaviour where else? And if it does not reveal itself in behaviour but is merely stated in judgement how much should we trust this as a basis of ethical theorizing then? If it is still claimed that seeming disagreement merely conceals a deeper agreement then additional experimental research can perhaps decide the issue and ethics can learn even more from experimental economics.

For the time being we can only observe that experimental economics shows that the heterogeneity of justice related behaviour is much greater than most philosophers – except for some hard-nosed cultural or phenomenological relativists – assume. In view of this observed heterogeneity we should at least prima facie suspect that any allegedly hidden consensus on normative principles is lacking. Homogeneity is found in conventions or practices. Yet this does not signify much since they may be followed in overt behaviour without being supported by a deeper consent.

6. Conclusions on bounded justice

In developing a theory of justice Rawls should have taken the quasi-empiricism of his basic methodology of reflective equilibrium search more seriously. Had he done so he might have looked at real actions of real people in justice related situations. However, the project of “a theory of justice” was too ambitious for this. Bringing in empirical facts is easier within a “local justice” (Elster, 1992) approach. Using “local” in a broader sense including specification along several dimensions like time, space, context and also context specific (ecological) rationality principles a “bounded justice” approach may apply (Schmidt, 1994).

In line with such an approach we explored a very specific case: the relationship between efforts to develop a boundedly rational theory of justice and normative facts as appear in the well-known game experiments. The search for reflective equilibrium is not anymore anchored merely in pseudo-empirical arm chair judgements that are derived introspectively but in normative facts explored experimentally. Moreover, it is not merely based on what people say they would do but on results about what they do or have in fact done when confronted with justice or equity related real choices as in particular in ultimatum experiments.

Emphasizing the role of *revealed as opposed to stated* “moral preferences” we do not intend to deny that moral language articulates justice claims. We also concede that this is in itself a normative fact. However, it is a fact concerning language use and *expressive* habits only. If we go beyond the stage where “my say so” is pitted against “your say so” and look at “your do so” as opposed to “my do so” we are addressing complementary issues of behaviour in situations with high opportunity costs. Here we agree that “(s)ince ‘actions speak louder than words,’ the information conveyed by actions may also be the most credible” (Bikhchandani et al., 1992). The moral theorist should set out to articulate the normative convictions that might be guiding the justice related actions that are actually observed.

As experiments show, the ethical theorist will find less agreement in the world than the official wisdom would have it. At least if ethical theorists would use revealed along with stated normative principles in their search for (W)RE there is no hope of inter-personal agreement. Within the broadly speaking sceptical tradition of normative argument – reaching from Hume to Mackie (see again Hume, 1739/1978, Mackie, 1980, Mackie, 1977, Harman, 1977) – this result is not too disturbing. For this tradition, all justificatory moral argument is ultimately “agent-relative” or “to whom it concerns”. If individuals have different concerns they will come up with different views on justice and act differently. And, different prescriptions will apply to their actions. However, for the sceptical ethical theorist it makes a difference whether an argument concerns many or few and whether it relates to deeper or more superficial concerns of its addressees. Though arguments from fictitious or conceivable consent may be irrelevant, the factual consent of as many individuals as possible can be most relevant. And, this is an empirical matter in which at least ethics – and for that matter Robbinsian normative economics – can learn a lot from experiments along with other empirical research.

Albert, Hans (1978), *Traktat über rationale Praxis*, Mohr: Tübingen.

Bergh, Andreas (2008), 'A critical note on the theory of inequity aversion', *The Journal of Socio-Economics*, 37, 1789-1796.

Bikhchandani, Sushil, Hirshleifer, David and Welch, Ivo (1992), 'A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades', *Journal of Political Economy*, 100(No. 5. (Oct.)), 992-1026.

Binmore, Ken and Shaked, Avner (2007), Experimental Economics: Science or What?, London, ELSE working paper 263; download: http://else.econ.ucl.ac.uk/newweb/displayProfile.php?key=2.

Bolton, Gary, Brandts, Jordi, Katok, Elena, Ockenfels, Axel and Zwick, Rami (2008), 'Testing Theories of Other-Regarding Behavior: A Sequence of Four Laboratory Studies', in: Charles R. Plott and Vernon L. Smith (eds.), Handbook of Experimental Economic Results, Vol. 1, North-Holland (Elsevier): Amsterdam et al., pp. 488-499.

Brennan, Harold Geoffrey and Buchanan, James McGill (1985), *The Reason of Rules*, Cambridge University Press: Cambridge.

Buchanan, James M. and Tullock, Gordon (1962), *The Calculus of Consent*, University of Michigan Press: Ann Arbor.

Camerer, Colin (2003), *Behavioral Game Theory*, Princeton University Press: Princeton.

Daniels, Norman (1996), *Justice and Justification*, Cambridge University Press et al.: Cambridge.

Elster, Jon (1992), *Local Justice. How Institutions Allocate Scarce Goods and Necessary Burdens*, Russell Sage Foundation: New York.

Elster, Jon (Ed.), 1998. Deliberative Democracy. Cambridge University Press, Cambridge.

Fleck, Ludwick (1935/1980), *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*, vol. 312, Suhrkamp: Frankfurt.

Frankena, William K. (1966), *Some Beliefs about Justice*, The University of Kansas: Lawrence. Kansas.

Frellesen, P. (1980), *Die Zumutbarkeit der Hilfsleistung*, Alfred Metzner Verlag.: Frankfurt/M.

Frohlich, Norman and Oppenheimer, Joe A. (1992), *Choosing Justice. An Experimental Approach to Ethical Theory*, University of California Press: Berkeley et. al.

Gehrig, T., Levati, V., Levínský, R., Ockenfels, A., Uske, T. et al. (2007), 'Buying a pig in a poke: An experimental study of unconditional veto power', *Journal of Economic Psychology*, 28, 692-703.

Goodman, Nelson (1978), *Fact, Fiction and Forecast*: Harvard University Press: Cambridge.

Güth, Werner, Kliemt, Hartmut and Ockenfels, Axel (2001), 'Retributive Responses', *Journal of Conflict Resolution*, 45(4), 453-469.

Güth, Werner and Levati, Vittoria (2007), Listen: I am angry! An experiment comparing ways of revealing emotions, Jena Economic Research Paper.

Güth, Werner, Schmidt, Carsten and Sutter, Matthias (2007), Bargaining outside the lab - A newspaper experiment of a three person ultimatum game, The Economic Journal, 117, pp. 449-469.

Güth, Werner, Schmidt, Carsten and Sutter, Matthias (2003), 'Fairness in the Mail and Opportunism in the Internet - A Newspaper Experiment on Ultimatum Bargaining', *German Economic Review*, 4(2), 243-265.

Güth, Werner, Schmittberger, Rolf and Schwarze, Bernd (1982), 'An Experimental Analysis of Ultimatum Bargaining', *Journal of Economic Behavior and Organization*, 3, 367-388.

Güth, Werner and Tietz, Reinhard (1990), 'Ultimatum bargaining behavior - A survey and comparison of experimental results', *Journal of Economic Psychology*, 11(3), 417-449.

Hahn, Susanne (2000), *Überlegungsgleichgewicht(e). Prüfung einer Rechtfertigungsmetapher*, Karl Alber: Freiburg i.Br.

Hardin, Russell (2007), *David Hume: Moral and Political Theorist*, Oxford University Press: Oxford.

Hare, Richard (1973), 'Rawls' Theory of Justice', *Philosophical Quaterly*, 21(April and July), 144-155, 241-252.

Harman, Gilbert (1977), *The Nature of Morality. An Introduction to Ethics*, Oxford University Press: New York.

Hart, Herbert L. A. (1961), *The Concept of Law*, Clarendon Press: Oxford.

Henrich, Joseph, Boyd, Richard, Bowles, Samuel, Camerer, Colin, Fehr, Ernst et al. (2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press: New York.

Hoerster, Norbert (1977), 'John Rawls' Kohärenztheorie der Normenbegründung', in: Otfried Höffe (ed.), Über John Rawls' Theorie der Gerechtigkeit, Suhrkamp: Frankfurt a. M., pp. 57-76.

Hoffman, E., McCabe, K. , Sachat, K. and Smith, V. (1994), 'Preferences, property rights and anonymity in bargaining games', *Games and Economic Behavior*, 7, 346–380.

Huck, Steffen and Oechssler, Jürgen (1999), 'The indirect evolutionary approach to explaining fair allocations', *Games and Economic Behavior*, 28, 13-24.

Hume, David (1739/1978), *A Treatise of Human Nature*, Clarendon: Oxford.

Kant, Immanuel (1798/1977), *Die Metaphysik der Sitten*, vol. VIII, Suhrkamp: Frankfurt.

Kliemt, Hartmut (1985), *Moralische Institutionen. Empiristische Theorien ihrer Evolution*, Karl Alber: Freiburg.

Kliemt, Hartmut (1994), 'The calculus of consent after thirty years', *Public Choice*, 79, 341-353.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press: Chicago.

Lakatos, Imre (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press: Cambridge.

Levitt, Steven D. and List, John A. (2007), 'What do laboratory experiments measuring social preferences reveal about the real world?' *Journal of Economic Perspectives*, 21(2), 153-174.

Louvierre, Jordan J. , Hensher, David A. and Swait, Joffre D. (2000), *Stated choice mehtods: analysis and application*, Cambridge University Press: Cambridge.

Mackie, John L. (1977), *Ethics. Inventing Right and Wrong.* Penguin: Harmondsworth.

Mackie, John L. (1980), *Hume's Moral Theory*, Routledge: London.

Mackie, John L. (1982), 'Morality and the Retributive Emotions', *Criminal Justice Ethics*, 1982, 3-10.

Manski, Charles F. (2002), 'Identification of decision rules in experiments on

simple games of proposal and response', *European Economic Review*, 46, 880-891.

Rawls, John (1951), 'Outline of a Decision Procedure for Ethics', *Philosophical Review*, 60, 177-190.

Rawls, John (1971), *A Theory of Justice*, Oxford University Press: Oxford.

Rawls, John (1974), 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association*, 48, 4-22.

Robbins, Lionel (1935), *An Essay on the Nature and Significance of Economic Science.*, Macmillan: London.

Roth, Alvin E (1995), 'Bargaining Experiments', in: John H. Kagel and Alvin E Roth (eds.), The Handbook of Experimental Economics, Princeton University Press: Princeton, pp. 253-348.

Schmidt, Volker H. (1994), 'Bounded Justice', *Social Science Information*, 33(2), 305-333.

Sidgwick, Henry (1907/1981), *The methods of ethics*, Hackett: Indainapolis.

Simon, Herbert A. (1957), *Models of Man*, John Wiley and Sons: New York.

Simon, Herbert A. (1985), *Models of Bounded Rationality (1&2)*, MIT-Press: Cambridge, MA.

Sugden, Robert (2004), 'What Public Choice and Philosophy Should *Not* Learn From Each Other', *American Journal of Economics and Sociology*, 63(1), 207-211.

Suleiman, R. (1996), 'Expectations and fairness in a modified ultimatum game.' *Journal of Economics Psychology*, 175, 531–554.

Westermarck, Edvard (1906), *The Origin and Development of Moral Ideas I and II*, MacMillan: London.

Xiao, E. and Houser, D. (2005), 'Emotion expression in human punishment behavior.' *Proceedings of the National Academy of Sciences*, 102(20), 7398-7401.