Mar. '97
Published in Social Justice Research, March '97: v. 10, #1, pp. 1-21

A Role for Structured Observation in Ethics[1]

by

Norman Frohlich[2] & Joe Oppenheimer[3]

suggested running head: Structured Observation & Ethics

Corresponding author:

Joe Oppenheimer, Professor
Department of Government & Politics
University of Maryland
College Park, Maryland 20742
(301) 405 - 4136
email: joppenhe@bss2.umd.edu

Abstract

Progress in the natural sciences has depended upon the collection and use of carefully controlled observational data. By contrast, ethicists have failed to agree upon a role for observational data in their enterprise. Although factors, embedded in the human condition obscure the role of observational data in ethical theory, barriers to the use of such data in ethics may be superable. Observation may not provide definitive answers to most ethical or metaethical questions. However observation of carefully constructed experimental conditions may provide the basis for cumulative progress in some branches of ethics.

Keywords: ethics, justice, experiments, methodology, impartial reasoning

---

1/ We would like to thank Jane Mansbridge whose very careful editorial assistance was generously given. We also are indebted to Thomas Schwartz who spent long hours discussing these questions with us, and Geoffrey Sayre-McCord who encouraged us to persevere, as well as Ron Terchek, whose comments on early drafts were very helpful.

2/ Faculty of Administrative Studies, University of Manitoba, Winnipeg, Manitoba R3T 2N2

3/ Department of Government & Politics, University of Maryland, College Park, MD. 20742

A ROLE FOR STRUCTURED OBSERVATION IN ETHICS

Progress in the natural sciences has depended upon the careful collection and use of observational data. By contrast, the role of observation in ethics has not been well defined. When Richard Boyd (1988, p. 185) asks: "What plays, in moral reasoning, the role played by observation in science?" he acknowledges the absence of evidence against which to weigh ethical statements and suggests, implicitly, a potential role for such evidence. Some philosophers claim, with Michael Waltzer, that different cultures produce different local understandings of justice. Others claim that philosophy can expect only to clarify beliefs: the foundations are beyond philosophical investigations. Structured observation could undoubtedly help both in identifying local understandings and in clarifying beliefs. This paper, however, focuses on the broader claims of some philosophers to something closer to truth itself. Quite a number of philosophers such as John Rawls and Jurgen Habermas identify hypothetical procedural situations which might be used to identify and justify substantial ethical conclusions. Rawls uses them to identify the principles of justice individuals would choose and Habermas uses them as a means for individuals to understand their true interests.

The conditions specified by Rawls and Habermas fall under the rubric of those "generally conducive to the formation of true belief" (Brink 1989, p. 132). Rawls' use of a "veil of ignorance" and Habermas' "ideal speech situation" may be viewed as attempts to specify hypothetical conditions conducive to the discovery of important ethical insights. Careful attention to conditions such as these may furnish clues as to how one might structure environments of controlled observation to provide data useful in evaluating ethical claims. If there is validity in the identification of these theoretically identified conditions then by attempting to replicate key components of their constructs we should be able to generate observations that illuminate the philosophical issues. Because the theoretically identified conditions do not occur naturally, the conjectured events and observations would have to be studied under carefully stylized experimental conditions. Once generated, such observations might provide data for cumulative progress in some branches of ethics.

How the Ambient Environment Can Impede the Use of Observational Data

To illustrate why the use of observational data in ethics is problematic and how observation might be made to play a role in the enterprise, we start with a fanciful analogy. Moral philosophers must operate in a viscous sea of self-interest, affect, and imperfect information in their attempts to find ethical truths. If they believe that some moral truths exist, then they need means for overcoming these disturbing factors and both identifying and justifying those truths.

An Analogy

Suppose that Isaac Newton had been born a dolphin (henceforth D. Newton). What extra impediments would he have faced in attempting to discover the laws of motion? And if he had discovered those laws, what role could data about his observable world play in helping to confirm them? A moment's reflection should convince the reader that to discover those laws under aquatic circumstances, D. Newton-dolphin would have had to imagine conditions far different from those which he experienced in his ambient world.[1] For example, in the sea he would not have had the experience of his earthly counterpart who could observe the planets and stars moving in apparently immutable orbits. Thus the experience of an object in motion remaining in motion unless acted upon by an external force would have been far from D. Newton's experience.[2] His problem is that he is immersed in a medium far more viscous than air: its viscosity makes it difficult to recognize generalizations about motion.[3]

Suppose, however, that he had by a monumental feat of imaginative intellect jumped to the conclusion: "Every object in motion will continue in motion unless acted upon by an external force." What evidence could he adduce to justify that proposition? What observational data could he use to test his hypothesized law? He would have difficulty mustering evidence for his claim.

In D. Newton's world everything is surrounded by water. Its viscosity impedes the progress of objects in motion. Indeed, he would probably not view this impediment as a force. He would perceive water all around as the natural condition. D. Newton would have to assert that the existence of water is really only a local phenomenon and that water exerts a force which impedes

any motion. He would have to convince his professional colleagues that it is possible to have conditions in which "no force" is posed in opposition to motion. He would have to convince them that this apparently strange condition is in some sense more representative of the world writ large than is the presence of water all around. The strange circumstance consisting of the absence of water, he would have to argue, constitutes the appropriate frame of reference for evaluating his hypothesized universal "law". The existence of water is only a local phenomenon. He would demand the testing of his hypothesis in a situation which perhaps no-one had ever experienced - one in which there was no ambient medium imposing a force on objects in motion. He would propose testing his conjecture in a vacuum.

How bizarre it might appear to his fellow dolphins. To test a candidate for a general law he would be demanding the creation of a completely artificial environment unlike anything they had experienced. Observation of naturally occurring phenomena simply would not do.

Impediments to the Direct Use of Naturally Occurring Observations in Ethics

This seemingly bizarre scenario can be used to gain insight into why ethicists may have such difficulty in establishing ethical theories and using observational data to confirm or disconfirm them. They face a problem parallel to that of D. Newton. The parallel between the ethicist and D. Newton can be made explicit on these points. The conditions of observation which D. Newton requires for the confirmation or disconfirmation of his hypothesized law do not occur naturally in his ambient world. The water presents a disturbing factor which must be eliminated so that the appropriate test condition (a vacuum) can obtain. He cannot use direct observation of naturally occurring phenomena as data. There are too many disturbing factors and they are too strong in relation to the phenomena at issue to ignore.

The ethicist faces the same problem. What water is to D. Newton, self-interest, affect, and imperfect information are to the ethical theorist. We all are immersed in a sea of overwhelming self-regardingness which leads us to give inordinate weight to our own interests. We have scarce

access to information on many relevant aspects of most situations. Finally, since we are more closely tied to some people than to others, our judgments and acts are obviously partial.

To use observational data as a basis for progress in an ethicist's program it is necessary to first have a theoretical template, identifying the disturbing factors to be eliminated or minimized from our everyday observational environment. The ethical theory must furnish the locus of ethically relevant contexts.[4] But within these theoretically defined contexts there must be the possibility of intersubjective evidence for ethical propositions. The evidence must go beyond the perspective of a single observer.[5]

We propose that potential progress lies as it does in the natural sciences, in experimentally removing those features which contaminate the empirical results of an observational inquiry. To get a sense of what kinds of controlled observations may be helpful in this regard, we offer, as an example, a set of observations that are potentially relevant to the evaluation of one of John Rawls's (1971) ethical propositions.

An Example: Creating a Controlled Observational Environment to Implement Impartial Reasoning

John Rawls has argued that one can of achieve insight into principles of justice by reasoning from a hypothetically structured state of impartiality.[6] The argument is built upon an idealized thought experiment developed to reveal principles of distributive justice. The principles which emerge from this procedure are said to come out of what Brink (p. 12) called "an impartial and imaginative consideration of the interests of the relevant parties". The standard of evaluation used to rank alternative principles includes aspects of the welfare of all individuals and is framed by the deliberative process of all of the individuals from behind a "veil of ignorance". The conclusions reached by the hypothetical individuals in reflective equilibrium are presumed to embody optimal (or perhaps only acceptable) tradeoffs among the competing interests of the parties involved.

In the theoretical exposition, the diverse and complex preferences and the various perspectives of different hypothetical individuals are brought to bear on ethically problematic situations by an

act of the theorist's imagination. Conclusions regarding such things as the just distribution of primary goods in society are then derived from two law-like assumptions. The first is a normative assumption that reasoning from an impartial point of view gives impartially arrived at conclusions a claim to ethical validity. The other is an implicit behavioral assumption about how hypothesized individuals would reason under specified conditions. Using a set of statements which include these two assumptions, Rawls arrives at the ethical claim that primary goods in a society should be distributed so that the worst off individual(s) have as much of them as possible. He concludes that this is the best rule for organizing distribution in society.

However, after all the pages are turned, Rawls conducts only a thought experiment. He provides no explicit role for observational data in the process and no set of protocols for adjudicating the dissension inherent in the competition of values and perspectives among any real representative individuals. While Rawls never envisioned the conduct of an experiment in a real setting, his assumptions and the line of his argument embody a number of empirical premises. These implicitly point to a possible role for observational data in evaluating his conclusions. For observational data may either support, or call into question, the implicit assumptions and ensuing conclusions.

In this regard, were Rawls seriously interested in bringing data to bear he would face some of the same problems that confront D. Newton. Just as watching objects move in water could yield poor tests of the laws of motion, so collecting data regarding how individuals reason everyday about distributive issues could yield observational data poorly suited to the evaluation of principles of distributive justice. Just as D. Newton would need to create an approximate vacuum, the ethicist must create an artificial environment to eliminate enough of the disturbing influences of the ambient world to test Rawls's conjecture. The test environment must permit observation of how individuals reason under approximations of the conditions Rawls posited - conditions designed to induce impartial reasoning. Observations under those specific conditions may provide some replicable and corrigible evidence useful in evaluating the ethical claims.

Earlier, we have attempted to approximate conditions relevant to the Rawlsian claims to induce appropriate impartial reasoning and to generate observations regarding reasoning and choices about distributive justice (Frohlich, Eavey & Oppenheimer 1987a, 1987b; Frohlich & Oppenheimer 1990, 1992.) Others have replicated the conditions (Lissowski et. al. 1991; Saijo and Turnbull, 1995; Jackson, 1995). While making no claim to be able to "replicate" Rawls' "veil of ignorance" we did attempt to set up an approximation of his situation in which subjects would seriously confront and have to choose from among principles of distributive principles from an impartial point of view. While full description of the methodology cannot be given in the confines of this paper (for full details see Frohlich and Oppenheimer 1992) a brief sketch of the procedure may furnish a sense of how laboratory environments can structure observations relevant to ethical inquiry.

Five subjects were brought into a room and were presented with a subject handbook which introduced them to the question of distributive justice in society by setting out four candidate principles of distributive justice: 1) maximum expected value, 2) Rawls' difference principle, 3) maximum expected value subject to a guaranteed minimum income and 4) maximum expected value subject to a constraint on the size of the range between the worst off and the best off.[7] Subjects were tested to ensure that they understood the implications of the different principles for the shape of income distributions in society. Each subject was then allowed, individually, on four different occasions to choose a principle to govern a cash payoff she would get by drawing a chit from a bag. The chit drawn was understood to randomly assigned the subject to one of five income classes in a hypothetical society. The payoff that the subject received was dependent both on the class drawn and on the principle the subject had chosen. So, for example, if the subject had chosen the principle of maximum expected value and drew a bottom class assignment she would receive a very low payoff. With the same draw, a subject who had chosen the difference principle would have received a higher payoff. Conversely, the draw of a top income class chit would yield an expected value maximizer a very large payoff, and a Rawlsian a substantial but lower payoff. The

idea was to familiarize the subjects both with the notion of random assignment to an income class and the distributional implications of the different principles.

In the second phase of the experiment impartial reasoning was induced. Subjects were told that they were to consider the task of choosing a principle of distributive justice for a hypothetical society in which they were to imagine themselves as members. They were to discuss the pros and cons of the principles to which they were introduced (as well as any others they might care to). They were told that if they could reach consensus on a single principle the substantial monetary payoff for the second part of the experiment would be distributed to them according to that principle. Five payoff classes would be established using their chosen principle and each would be randomly assigned to one of those classes. Each subject, therefore, had to consider the possibility of being in any one of the five income classes and hence had to give impartial consideration to the interests and entitlement of each income class. They were further told that if they could not reach consensus, one of the principles would be chosen at random and the payoffs would be distributed in accordance with that principle. The discussion could be terminated by unanimous agreement under secret ballot. Voting on a choice of principle was also conducted by secret ballot. Unanimous agreement was required for the selection of a principle.

Subsequent research showed subjects from Australia, Canada, Poland (then still Communist), Japan, and the United States, were virtually always able to reach unanimous agreement. Moreover, the results were relatively uniform across locations. The vast majority of groups agreed to what Rawls calls a mixed principle: maximum expected value subject to a constraint on the income floor. Rawls's difference principle fared badly indeed. The results were extremely robust across treatments and locations. Moreover, the subjects' discussions and behavior over the course of the experiments indicate that they were probably giving impartial consideration to the issue (as the underlying incentive structure would imply).

These observations furnish some preliminary evidence regarding what people choose under conditions of impartiality. Of course, this does not mean that Rawls's conclusions must be rejected.

Nor do the findings mean that the mixed principle must be accepted as ethically valid. As will be discussed below, conclusions drawn from laboratory experiments can be criticized on many grounds. But, precisely because of the theory, the experiments furnish observational data bearing on the claim that the difference principle is just. The fact that the vast majority of people unanimously agree to an income floor constraint principle and reject the difference principle under conditions designed to induce Rawlsian impartial reasoning does not unequivocally make the one valid and the other invalid. It does, however, provide intersubjectively testable evidence whose force depends on the two initial assumptions: first, that impartial reasoning imbues decisions with ethical validity; second, that an appropriate approximation of impartial reasoning was achieved in the laboratory. Both of these are contestable.

If one disagrees with the conclusions and with aspects of the experimental environment designed to induce impartial reasoning, one can try to structure better observational conditions. This potential for refinement and replication of a laboratory experiment opens the possibility for cumulative and correctable intersubjective consensus based on observation. Such controlled observation contrasts with the tradition of refining considered judgments via argumentation and reflection. Argumentation and reflection do not provide an empirical means whereby others may intersubjectively evaluate the truth value of the conclusions.

General Requirements for the Use of Observational Data to Confront Ethical Conjectures

Nihil ab nihilo fit - Nothing comes from nothing. Moral philosophers, like their natural science counterparts seeking a basis for formulating and testing their theories, must start somewhere. In both cases, simple *observation of naturally occurring* phenomena provides a minimal starting point.[8] Sets of observations lead to the construction of categories, generalizations, and conjectures. These are, initially and informally, compared with subsequent observations as a means of testing the fruitfulness of the constructed categories and the putative truth of the conjectures. Such an informal procedure may suffice to satisfy a single individual attempting to make sense of the world for herself.

However, to insure that the hypothesized generalizations are broadly applicable and not simply the idiosyncratic product of a singular perspective, we must broaden and make explicit procedures for testing the generalizations. Explicitly, some basis for intersubjective consensus must be provided, including agreement regarding the answers to the following question:

> What observations are to be considered germane to the judging of the validity of the generalizations?

Even in the natural sciences, this question is problematic: yet much progress has been made across broad areas of inquiry. Brink has noted the starting point for reliable moral beliefs shares much in common with that of reliable belief of other sorts. (Brink 1989, p. 132):

> ...[M]oral beliefs formed under conditions generally conducive to the formation of true belief will be more reliable than moral beliefs not formed under these conditions. A belief that is based on available (nonmoral) evidence and is thus well informed, that results form good inference patterns, that is not distorted by obvious forms of prejudice or self-interest, that is held with some confidence, and that is relatively stable over time is formed under conditions conducive to truth.

But Brink (ibid.) adds one additional condition which he believes to be particularly important in the formulation of moral beliefs:

> Because of the importance of impartiality in making moral decisions and the connection between morality and human good and harm, we are likely to obtain a more reliable class of moral beliefs by focusing on moral beliefs that have been formed not only under conditions of general cognitive reliability but also on the basis of an impartial and imaginative consideration of the interests of the relevant parties. We might call beliefs formed under such conditions *considered moral beliefs* (cf. Rawls 1951: 53-5, 1971:47-8).

The considered moral beliefs which emerge from such a process are the first candidates for ethical truth. Those who seek to improve the methodology of ethical inquiry will want to refine

intersubjective procedures for confirming or disconfirming such beliefs. For this purpose, some aspects of the standard of evaluation used in the arguments must, in principle, be potentially approximable in a constructed environment. Both the content and the procedures must be accessible empirically. But that is not enough. For cumulative progress to occur within any ethical research program, the *possibility* of observation is inadequate. *Actual* observation must take place: evidence must be marshalled. Individuals must sample evidentiary cases and be able to judge how the observations are at variance with theoretical expectations.

Controlled Observation as an Intersubjective Criterion of Truth Evaluation

Consider, Rawls' ethical claim that: "primary goods should be distributed in society so that the worst off individual in society is as well off as possible." Note that this claim is a considered moral belief about *all* cases concerning distribution under specific conditions (such as those where individuals reach conclusions by impartial reasoning). Hence the results of any such reasoning would be germane to the evaluation of the truth of the generalization.

Now we can see how one might specify a method for establishing the intersubjective identification of its truth. The results of any controlled observation would have to have the following two characteristics to constitute support for the difference principle: a) in any controlled observation all relevant individuals would have to exhibit similar support for the principle, and b) in different controlled observations, which are heterogeneous in context and subjects but which approximate the conditions specified in the theory, similar support for the principle must be forthcoming.

To evaluate the ethical validity of the principle, observations from the class of all evidentiary situations (whatever that might be) would have to take place, or be reported, or be filled in by the theorist (depending on whether the case was ongoing, reported, or hypothesized). And of course, there are no simple rules for determining which observations are best.[9] What is clear, however, is that some evidentiary cases are more conducive to the uncovering and reporting of *relevant* facts

which can be used in justifying the principle. These are precisely the ones to be found in a controlled environment structured to facilitate relevant observations and evaluation.

Of course, there may not be agreement on exactly what needs to be included in any given observational environment. Researchers in ethics may disagree, as do their counterparts in the natural sciences, on the importance of particular experimental settings and results. But the explicit specification of the parameters of the observations should facilitate movement towards intersubjective identification of corrigible truth claims.

At times one can combine the insight of the cognitive ethicist who prescribes the refinement of considered judgements with the methodology of the experimental scientist. The scientist creates the environment, usually an experimental setting, for bringing to bear observational data (i.e. abstract non-operationalized formulation of the theory
regarding representativeness and knowledge conditions). The theory may be used to identify some ideal situation particularly conducive to the identification of a true principle. Those ideal situations may involve conditions which induce or apply an appropriate standard of evaluation, but further: those conditions may be used to identify some empirical setting(s) which approximate the ideal situation. Such an empirical environment can provide a bridge between the ideal and the observable. Since the specification of the ideal conditions is likely to be incomplete, the gaps between the *ideal* and *realizable* conditions must be filled in to give weight to the theoretical conclusions of the cognitive ethicists. Much as it does in science, the experimental laboratory may furnish a bridge between the ideal conditions and the empirical substantiation of a theory. For example, selection of samples of humanity from widely divergent backgrounds might embody different conceptions of the good and hence might constitute a prerequisite for strong tests of deliberative ethical conjectures. Laboratory deliberations and their associated outcomes could be conducted by very varied representative groups.

What is required is a series of instantiated approximations of some of the ideal decision situations by the creation of a carefully controlled decision environment. *Observation of decisions*

taken under the controlled decision conditions would constitute the observational data necessary to corroborate or refute conjectured ethical truths. Just as a vacuum might allow D. Newton to test his conjectures about motion, a laboratory environment may enable ethicists to test some of their conjectures by eliminating or minimizing the impact of partiality, and limited knowledge.

On the surface, it may appear that this methodological prescription is made in opposition to the traditional methodology of subjective individual observation, introspection, consideration of prior written argument, conjecture, debate at a distance based on the written word, and reformulation. On closer look, it is not. Rather it is a supplement, which offers the prospect of enriching the process by providing access to sets of observational data that may be intersubjectively compared. Whether this supplementary methodology is worthwhile is, to some extent, an empirical question. We subscribe, partially, to the position of Laudan (1987):[10] "The criterion for evaluating any methodology is its fruitfulness. Fruitfulness is, in part, an empirical question."

A Second, Simpler Case: Collective Action

We have used Rawls "difference principle" in our discussions of how an ethical conjecture might be tested and why it is important to do so. One of our main arguments for testing is based on the fact that the difference principle has a very broad and complex set of extensions. It is presumably posited to hold over all societies of "moderate scarcity". And, in part, it is the extension to "all" such societies that makes a traditional analysis so difficult. To eliminate some of the complications involved in that example, and expand on some of the further implications of the methods, we now turn to a simpler illustration.

Political philosophers have long focussed on the problem of personal obligation to participate in social efforts. Although there have been numerous formulations of the problem, it has come to be referred to as the problem of the "Logic of Collective Action" (Olson, 1965) and is often thought of as an n-person prisoners' dilemma game (Hardin, 1971). In developing their prescriptions for how to calculate one individual's obligation to the group, some philosophers used the traditional impartial reasoning tools (Strang, 1960). Philosophers have argued that such thinking solves the

dilemma, and enables the group to obtain the services of its members. Again, we can ask, do real people, making decisions, under controlled laboratory conditions, which approximate impartial reasoning contribute to public goods as they ought to?

To develop a test of this simple conjectured ethical solution to the collective action problem, one must transform the model of the prisoner dilemma into a choice problem in an impartial reasoning framework. Frohlich (1992) accomplished this and showed that, indeed, the impartial reasoning solution to the choice problem in a prisoners' dilemma is to cooperate. This is quite easy to see: Consider the well known 2-person Prisoner's Dilemma. If players approach the game from an impartial reasoning point of view, they must consider the payoffs to both individuals impartially in deciding on their strategy choice. When they do this, cooperation becomes the dominant choice. If impartial reasoning is used as an ethical premise to generate the choice, cooperation is the ethically correct thing to do in a Prisoner's Dilemma.

This perspective on the 2 person PD can be extended to the typical n-person case. Once that has been done in general, a model can be developed for a specific n-person dilemma and one can set up an experimental design. The experiment could be used to test whether subjects, placed in conditions which induce impartial reasoning actually choose as ethical theory argues they ought. This was done and the results reported in a pair of articles (Frohlich and Oppenheimer 1995, 1996). Corroboration of predicted behavior would lend support to the notion that impartial reasoning is a way to understand and generate ethical solutions to problems. Of course, if experimentally one found that people did *not* choose to cooperate when thinking impartially (say under a large variety of experimental conditions), the theory relating ethical choice to impartial reasoning would suffer.

Table 1 about Here

Consider a simple linear 5-person prisoners' dilemma with the typical assumptions. The main requirements are that the payoff structure generates a dominant strategy for each player and the choice of those dominant strategies leads to a Pareto - inferior outcome. The game is depicted in Table 1. There each individual has a budget of 10 units, and can either keep the 10, or put any

proportion of it into a bonus fund.[11] Every unit in the fund yields .4 units to each person in the game. Since a contribution of any quantity, **x**, by a player results in a loss of **x** plus a gain of only .4**x** it is individually rational to contribute nothing at all. For example, in the contingency in which others' contributions amount to 10 units if the row player gives nothing (hence keeps her budget of 10) she will receive 4 out of the fund (as a result of the 10 units given by others) and will have a net return of 14. Contributing 10 reduces that sum to 8 units - and so on. Under each contingency of others' contributions, a single player is best off contributing nothing. If everyone is individually rational and no contributions are forthcoming the result is that each gets 10 units and the total group payoff is 50.[12] On the other hand, if all were to give their full 10 units, 50 units would be in the bonus fund and that would generate a payoff of 40% of 50 or 20 units for each player - a group payoff of 100.[13] All could do better under the latter outcome: full cooperation. But individual rationality yields the theoretical prediction of complete defection and the Pareto inferior result.

Table 2 about Here

Consider now how thinking about the problem impartially changes the incentive structure of the game and transforms it. Let each of the players in the game make a decision, and then let a randomizing device determine which player will actually get which payoff (see Table 2). For each level of expected contributions by others we can calculate the payoffs of any strategic choice, say for player one. For example, we can again imagine that one other player contributes 10 units. We can then contrast the payoffs to player one of either contributing nothing or contributing something.

Contributing 10 would produce 2 (out of 5) contributors, and 3 non-contributors. *Each* player would then receive an expected payoff consisting of 2 out of 5 chances of being assigned to a position which had contributed and 3 of 5 chances of getting a position which had not. The net expected value of those would be .4*(20*.4) + .6(20*.4 + 10) = 14. That is the expected value of contributing. On the other hand, contributing nothing leaves only one contributor and 4 non-contributors. Under that contingency, *each* individual has 1 out of 5 chances of getting that

contributor's payoffs and 4 of 5 chances of getting a non-contributor's payoff. The expected value of that strategy is:

$$.2*(10*.4) + .8(10*.4 +10) = 12.$$

The value of not contributing is smaller than the value of contributing. This is true under all contingencies. Thus, there is an incentive to contribute one's total resources. The transformed game has a dominant strategy of contributing 10 rather than contributing nothing.

The experiments we conducted placed five subjects in front of five linked computers which displayed the game in **Table 1**, However, players were told that after they had made their strategic choice they would be randomly assigned to one of the five computers. They would receive the payoff associated with the decision which had been made at that computer. This randomization induced individuals to give equal weight to the payoffs of each subject. In that sense it invoked impartial reasoning. Implicitly, subjects faced the game in **Figure 2** although the numbers they saw were those of **Figure 1**. They were hypothesized to choose the ethically correct strategy: cooperate and place all money in the bonus fund.

Experimental results of playing the game from an impartial point of view (both with and without communication) demonstrated that subjects do indeed choose, overwhelmingly, to cooperate when faced with the game complete with a randomizing device. By contrast, in similar games played in the ordinary way, researchers have found behavior which was clearly more self-interested (see Ledyard, 1995). Our findings are support for the ethical proposition that impartial reasoning can generate ethical behavior (Frohlich and Oppenheimer, 1995, 1996).

Laboratory Experiments: A Methodological Step Forward

To see the link between the set of ideal situations and an experimental approach, consider the generic form of the arguments with which we are concerned. Ethical theories may specify ideal situations, under which the ethical reasoning and choice are to occur.[14] The ideal conditions imbue any conclusions reached under them with ethical validity. The reasoning underlying their claim might be sketched in the following 4 steps:

1) There exist ideal conditions for choice of an ethical principle.[15]

2) Any principle chosen under ideal conditions is a valid principle.

3) Under the ideal conditions the theorist claims a particular principle would be chosen.

4) Therefore that principle is a valid ethical principle.

As noted, philosophers traditionally use **thinking** about how such individuals might choose under ideal conditions as a basis for determining what constitutes the chosen principle. They often they reach different conclusions because they implicitly project their personal perspective to fill in the blanks.[16] A theoretician working through a thought experiment such as Rawls's must answer questions such as 1) What preferences and knowledge do representative individuals bring to bear on the issue? 2) Whom do they represent? 3) How are the preferences and theoretical knowledge to be applied to consider the possible consequences of alternative choices? and 4) What standard is to be used to rank alternative choices? Even if those questions are answered explicitly, how is the theorist running this thought experiment to know whether that particular characterization of the thought experiment is appropriate?

No single philosopher can fill in the complex and diverse details of the preference, knowledge, and backgrounds of numerous representative individuals accurately. No one person can be expected, with a text authored from a single life-experience (however rich) to reliably identify what standard of evaluation would or should be used, or what principle should or would be chosen. Such identification requires the use of data. The traditional method of using a thought experiment provides the impetus for going beyond the introspective experience which a *single* mind can capture. The method suggests a basis for making explicit the combined experience of all relevant other (unknown) persons and so explicitly adjudicating among their competing interests.

The ethicist facing only hypothetical states will have difficulty predicting the outcome a group of people would arrive at were they to be placed in a difficult-to-realize and hypothetical environment. Each philosopher's particular life experience constrains her or his ability to comprehend and to assimilate impartially the wildly diverging life experiences, perspectives, and

preferences of others. Yet precisely those preferences and interests must play a major role in determining the choice. When the preferences of representative individuals are simplified and their knowledge base stripped down in a thought experiment, the austere theoretical structure is sure to be inadequate to the task. It will be unable to generate a definitive result. It will not identify a unique outcome. The ongoing disagreement between Rawls, Harsanyi and Nozick as to which principle would be chosen in such a process is testimony to the indeterminacy of the idealized conditions as a definitive generator of ethical results.

We suggest that mere hypothesizing and reasoning about what might happen under idealized conditions is doomed to ambiguity or worse. It is clearly impossible to replicate, in the real world, the *ideal* conditions called for by philosophers. But it is possible to model and *approximate* them.

Using this methodology it is possible to specify a relationship between an empirical inquiry and the abstract philosophical approach.[17] The general syllogism describing this relationship maintains the basic premises of the above 4 step philosophical argument but adds empirical content by introducing another two premises:

5) One can construct construct an experimental situation which approximates the ideal situation.

6) The actual choice of any principle, under the experimental approximation is intersubjective evidence regarding the ethical validity of the theoretically justified principle.

Thus the set of all experimental approximations of the ideal situations constitutes a locus for gathering observational data on the validity of the theory. Observations about what would happen in such experiments would constitute much of the crucial evidence for the theory.

How does this resort to the laboratory move towards intersubjective identification of the truth? In the first instance, the use of experimental subjects with diverse experiences, background, and preferences provides an obvious remedy to one of the problems identified above: the projection of a choice from the perspective of a single philosopher. Individuals, engaged in discussions and bringing to bear experiences from their diverse lives, can provide complex context to the decision

process. Further, actual discussions and tradeoffs can give the decisions a trajectory and nature an isolated philosopher could not possibly anticipate by introspection. And experiments offer the prospect of discovering whether a posited definitive choice does indeed emerge in the lab. And if the results of the tests are not the same as those conjectured, does some other principle come to the fore or are some set of alternatives always rejected? Furthermore, any such evidence is correctable or subject to additional confirmation via subsequent replication. Replicability furnishes the opportunity for intersubjective agreement regarding evidence.

Discussion

We have identified some methodological means for evaluating the truth values of some ethical theories. The experimental methodology we are proposing very closely parallels the experimental methodology used in the natural sciences. With that analogy in mind, possible problems with the experimental methodology proposed here can be put into context. They are the problems of correspondence and representativeness that are found in experimental literature in general.

Any particular experimental modelling of an ethical theory must provide answers to any number of key questions. How are the ideal conditions, expressed in the theory, interpreted in the experimental model? To what extent do the experimental conditions approximate the ideal conditions? Are there specifiable cognitive abilities that need to be present for an individual to be empowered to evaluate alternative principles for the class of ethically problematic cases under consideration? If subjects are to choose from among a number of alternatives, their choices must be informed by some preferences. How diverse need the preference sets be? What samples of humanity are needed to represent possibly different conceptions of the good underlying the standard of evaluation that will generate the choice? If the experiment is evaluating a general principle, then the principle has implication for a variety of situations. The sample of people and the sample of particular instances of problematic situations under consideration must be broad enough to justify confidence in any consensual conclusions reached.[18] Moreover, the individuals involved must conduct deliberations regarding possible alternatives actions and associated

outcomes relevant to the class of problematic situations. What procedures and processes are the individuals to use in this process. And in the end, what decision rule are they to use?[19]

Finally, it is necessary to identify standards for the evaluation of the observational data generated by the experiment. Does the failure of one group in a million to reach consensus negate the generalization? Does one in a thousand? Or is 95% agreement enough to justify its acceptance?

There are no simple *a priori* answers to these questions. The nature of the sampling of individuals, candidate principles, instances and decision rules, may vary as a function of the category of ethically problematic situations being addressed. Questions of individual responsibility for public and collective action may require different conditions for arriving at justifiable interpersonal identifications of consensus than do questions of distributive justice. Some categories of ethically problematic situations may not yield intersubjective consensus, and there may be no "truth of the matter" identifiable. But the appropriate parameters for the establishment of inter-subjective observational environments would likely emerge, as they have done in the natural sciences, as a result of the success and failure of differing protocols of evaluation. Nor, again in parallel with the natural sciences, are the protocols ever likely to be free of controversy.

An extension of the experiments discussed above shows how empirical results can be used to develop the theorizing fruitfully. Consider the contrast in perspective between Habermas and Rawls which we alluded to in the beginning. Which approach is likely to generate ethical decisions more consistently: "ideal speech situations," or "impartial reasoning"? If choosing impartially can develop ethical reasoning, we might conjecture that after a period of such choices the individuals involved would identify and then choose the ethical alternative. But when we ran experiments to discover whether individuals would exhibit such "ethical learning" we had some surprises.

The experiments involving the 5 person prisoners' dilemma had two phases: in Phase 1 each individual faced a repeated 5 person dilemma of the sort described in Tables 1 either directly or subject to randomization. In Phase 2 they all faced the same dilemma (without communication) of

the type shown in Table 1. The experiments were run both with and without communication in Phase 1. Subjects who decided under conditions of impartial reasoning in Phase 1 (both with and without communication) played Phase 2 *less* cooperatively than those who played Phase 1 in the normal way. The differences in cooperation were most pronounced when compared with the behavior of subjects who played the standard (non randomized) PD and were allowed to discuss the issue in Phase 1. (see Frohlich and Oppenheimer, 1995).

Although one can easily make too much of one set of experiments such findings help us understand the conflict between the theoretical arguments and positions of Habermas and Rawls. On the other hand the force of consensual results (should they be observed) are limited. They clearly are corroborative only in a conditional sense. They assume the relevance of the ideal situations to the identification of the underlying truths. Abstract or paradigmatic disputes regarding possibly different ideal conditions cannot be adjudicated easily. Choices between competing notions of ideal conditions may require concrete instantiations of cases in which the competing ideal conditions yield different ethical claims. If we are lucky, paradigmatic disputes will foster critical experiments. But we can not be sure of this.

What might be expected of the proposed methodology under a best case scenario? Perhaps it may be possible to identify one corner of ethical inquiry in which a few standards of evaluation enjoy broad support among theorists. In that realm, controlled observation of approximated idealized conditions might lead to correctable and cumulative knowledge claims regarding ethical matters. But Newton was correct: objects at rest will remain at rest unless acted upon by an external force.

Endnotes

1/ To emphasize the difficulty of a purely water-dwelling creature, we assume D. Newton never jumps into the air nor basks on the surface and stares at the world above the water.

2/ D. Newton could observe non-buoyant falling bodies in the ocean and conclude that they continue to fall until they hit bottom, and these observations might enable him to posit

something analogous to the general law, but it would be hard to see how he could generalize the law so that it was non-directional, and applied to motion in any direction (even up).

3/ On the other hand, given the resistance of water, objects floating in the ocean remain relatively at rest (unless moved by a current, or some other force). Thus, D. Newton might well have been able to arrive at one part of his first law: "Every object at rest will remain at rest unless acted upon by an external force." But that would only appear to apply to floating objects.

4/ After all, no amount of data generates an ought.

5/ This is true for both the moral realist who believes that one discovers ethical facts nested in an observable reality or, and the anti-realist who may hold that one synthesizes ethical statements and evaluates them in relation to their coherence. We are agnostic on this ontological point, and believe that our arguments apply in both contingencies.

6/ Another example is Roderick Firth (1952). Although his conditions seem to vary from those proposed by Rawls, it has been argued that there are close parallels in their constructions (Harrison, 1956 and Frohlich & Oppenheimer, 1992).

7/ Actually eight different variants of the experiments were run in five different countries over the course of 12 years by ourselves and other investigators. The variety of versions was necessitated by the robustness and persistence of our results and our attempts to determine that 1) it was not a peculiarity of the research design which drove the result and 2) that the result was indeed generalizable to different populations. The version detailed here is the original, simplest treatment.

8/ This is indeed the answer Boyd gives to the question he originally posed (see Boyd p. 206, and above p. 1)

9/ The identity of many of these details is, of course, mainly dependent upon one's ethical theory.

10/ Maxwell (1972) shows some of the difficulty in such a program. Also see Able and Oppenheimer for a view as to why empiricism can not be a complete methodological program.

11/ For ease of explanation the table displays only a discrete representation of the game, but it is conceptualized and implemented as a continuous game with strategy choice domain being the closed interval [0,10] and the payoffs [4,26].

12/ This is often referred to as the Nash outcome.

13/ This is often called the cooperative, or core outcome and is Pareto optimal.

14/ As indicated, this form of ethics is often referred to as part of the "cognitivist" school of ethics (see Frankena, 1963).

15/ These might specify knowledge and preference conditions as well as the context of the deliberation and choice.

16/ Consider an example from Harman (1988) as discussed by Sturgeon (1988): An individual observes young hoodlums pouring gasoline on a cat and setting it on fire. The dialogue between Harman and Sturgeon concerned whether this constitutes the possible direct observability of an ethical fact. In so doing, it points out the lacunae of the traditional method. In the dialogue, Harman claims one can "make a moral judgment immediately and without conscious reasoning say, that the children are wrong to set the cat on fire" (p. 122, page references here are to the Sayre-McCord volume). Harman goes on to say that "all we need assume ... you have certain more or less well articulated moral principles that are reflected in the judgments you make." And he continues, there is no "obvious reason to assume anything about 'moral facts,' such as that it is really wrong to set the cat on fire." Sturgeon takes issue with this position (p. 232-234) by noting that we do need to incorporate the argument into an explanatory mould, and thereby harness other assumptions about morality. But neither of them go further to note the obvious point: additional information is needed to evaluate the assertion that "It is wrong to set the cat on fire." One needs to fill in the parametric conditions of the situation more fully. Only then can the statement be evaluated. For example, the cat may already be dead. Or perhaps the children believe that cats were spreading a fatal disease and so needed to be burnt. Or imagine that they believed in a religion that called for cat sacrifice to bring healing to their sick relatives. Then our evaluation of the acts as *wrong* (rather than, let's say, *repulsive*) may have to be altered. Settling the matter (for a class of impartiality type theories) could call for the explicit presentation of the fine details of the case in such a way that numerous observers can reach intersubjective consensus on a conclusion. Indeed, it is often exactly because the fine details of the situation are *not* specified and an evaluation of them and their consequences are *not* opened to interactive debate that one observes apparent disagreements about the "truth" of an ethical claim in a particular instance. And of course, the problem is compounded when we are concerned about ethical

generalizations. Thus, structured observations are preferred to naive empirical analysis because it specifies more fully the conditions of observations. This permits us to find out when particular judgements will be biased or wrong.

17/ We are indebted to Thomas Schwartz for suggesting this particular form of the argument.

18/ If, as in the case of ideal observer theory, one individual is posited, that individual must be able to incorporate the preferences of **all** individuals in the real world who might fall subject to the choice to be made. For example, Firth (1952) has specified that an ideal observer needs to be: 1. omniscient with regard to relevant non-ethical facts; 2. omnipercipient (i.e. able to empathize perfectly) with regard to all the relevant parties; 3. disinterested; 4. dispassionate among the parties and toward the issues involved; 5. consistent (over time); 6. in other respects, normal. It is clear that such a task is well beyond the powers of any single individual. A philosopher might posit that the choice made by an ideal observer would be fair, but it would not be possible to operationalize that choice in most significant ethical choices.

19/ Theories posited on decisions made by a set of representative individuals (such as Rawls, 1971) generally require unanimity as the choice rule. The appropriate instantiation of a rule in an experiment is a matter that requires careful attention.

References

Abel, C. Frederick, and Joe A. Oppenheimer. (1982) "Liberating the Industrious Tailor: The case for Ideology and Instrumentalism in the Social Sciences." Political Methodology, 8, no. 1, 39 - 60.

Boyd, Richard B. (1988) "How to be A Moral Realist," in ESSAYS IN MORAL REALISM, (pp. 181-228) ed. Geoffrey Sayre-McCord, Cornell U. Press: Ithaca.

Brink, David O. (1989) MORAL REALISM AND THE FOUNDATIONS OF ETHICS, Cambridge Univ. Press: Cambridge.

Firth, Roderick (1952) "Ethical Absolutism and the Ideal Observer", PHILOSOPHY AND PHENOMENOLOGICAL RESEARCH, XII, No. 3, March, pp. 317 - 345.

Frankena, William K. (1963) ETHICS. Englewood Cliffs: Prentice Hall.

Frohlich, Norman and Joe A. Oppenheimer (1992). CHOOSING JUSTICE: AN EXPERIMENTAL APPROACH TO ETHICAL THEORY. California University Press: Berkeley.

Frohlich, Norman [and Joe A. Oppenheimer]. (1995) The Incompatibility of Incentive Compatible Devices and Ethical Behavior: Some Experimental Results and Insights. Public Choice Studies, V. 25: 24-51. (Incorrectly published without Oppenheimer's name on it).

Frohlich, Norman and Joe A. Oppenheimer (1996)."Experiencing Impartiality to Invoke Fairness in the n-PD: Some Experimental Results." Public Choice, 86 (117 - 135).

Hardin, R. (1971) "Collective Action as an Agreeable N-Prisoners' Dilemma" Behavioral Science 16 (no. 5): 472 - 479.

Harman, Gilbert (1988) "Ethics and Observation," in ESSAYS IN MORAL REALISM, (pp. 119-126) ed. Geoffrey Sayre-McCord, Cornell U. Press: Ithaca. Originally published as Chapter 1 of Harman (1977), THE NATURE OF MORALITY. New York: Oxford (p. 3-10).

Harrison, Jonathan (1956) "Some Comments on Professor Firth's Ideal Observer Theory", PHILOSOPHY AND PHENOMENOLOGICAL RESEARCH, XVII: pp. 256 - 262.

Harsanyi, John C. (1953). Cardinal Utility in Welfare Economics and in the Theory of Risk - Taking. JOURNAL OF POLITICAL ECONOMY, 61: 434 - 435.

Harsanyi, John C. (1955) "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," JOURNAL OF POLITICAL ECONOMY, v. 63: pp. 302 - 321.

Jackson, Michael, "A Fair Share" JOURNAL OF THEORETICAL POLITICS, Vol 7 (2): pp. 169-179.

Laudan, Larry "Progress or Rationality? The Prospects for Normative Naturalism," AMERICAN PHILOSOPHICAL QUARTERLY, Vol 24, No. 1 (January, 1987): pp. 19 - 31.

Lissowski, Grzegorz, Tadeusz Tyszka and Wlodzimierz Okrasa. Principles of Distributive Justice: Experiments in Poland and America. Journal of Conflict Resolution v. 35, No. 1, March, 1991: 98 - 119.

Maxwell, Nicholas "A Critique of Popper's Views on Scientific Method," PHILOSOPHY OF SCIENCE, (June, 1972), 131-152.

Nozick, Robert (1974) ANARCHY, STATE AND UTOPIA, New York: Basic Books.

Olson, Mancur, (1965) The Logic of Collective Action. Harvard University Press: Cambridge.

Rawls, John (1951) "Outline for a Decision Procedure for Ethics." Philosophical Review 60, 2 (April): 177-97. Reprinted in J. Thomson and G. Dworken (eds.) ETHICS. New York: Harper & Row. 1968. (Page numbers are to the reprint.)

Rawls, John (1971) A THEORY OF JUSTICE, Cambridge: Harvard University Press.

Saijo, Tatsuyoshi, and Steven Turnbull (1994) Personal correspondence regarding experiments conducted at Tskuba University in Japan in 1994.

Sayre-McCord, Geoffrey (1988) "Introduction: The Many Moral Realisms," in ESSAYS IN MORAL REALISM, (pp. 1-26) ed. Geoffrey Sayre-McCord, Cornell U. Press: Ithaca.

Strang, Colin (1960) "What if Everyone Did That?" Durham University Journal, 53 (1960), pp. 5-10. Reprinted in Baruch A. Brody, ed. MORAL RULES AND PARTICULAR CIRCUMSTANCES. pp. 135 - 144. Prentice Hall: Englewood Cliffs, New Jersey. 1970.

Sturgeon, Nicholas L. (1988) "Moral Explanations," in ESSAYS IN MORAL REALISM, (pp. 229-255) ed. Geoffrey Sayre-McCord, Cornell U. Press: Ithaca. Originally from MORALITY, REASON AND TRUTH, ed. D. Copp and D. Zimmerman (Totowa, NJ: Rowman & Allenheld, 1985), pp. 49-78.

TABLES AND FIGURES

Table 1: 5 Person Prisoners' Dilemma (Showing Payoffs Only to The Row Player)	Amount Given by Others				
1 Person's Strategies	**40**	**30**	**20**	**10**	**0**
give 0	26	22	18	14	10
give 10	20	16	12	8	4

Table 2: Impartial Transform of the 5 Person Prisoners' Dilemma (Showing Payoffs Only to The Row Player)	Amount Given by Others				
1 Person's Strategies	**40**	**30**	**20**	**10**	**0**
give 0	18	16	14	12	10
give 10	20	18	16	14	12

Contents

Insert Page - Additional (as requested) References

Reference insert #1:

Frohlich, N., Oppenheimer, J. A. and Eavey, C. (1987a). Laboratory Results on Rawls' Principle of Distributive Justice. British Journal of Political Science 17: 1-21.

Frohlich, N., Oppenheimer, J. A. and Eavey, C. (1987b). Choices of Principles of Distributive Justice in Experimental Groups. American Journal of Political Science 31(3): 606-636.

Frohlich, N. and Oppenheimer J. A. (1990). Choosing Justice in Experimental Democracies with Production. American Political Science Review 84(2): 461-477.

Reference insert #2:

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. In Kagel, J. H. and Roth, A. E. (eds.), The Handbook of Experimental Economics, Princeton University Press: Princeton, NJ, pp. 111-194.