Blind Justice: Fairness to Groups and the Do-no-harm Principle

JONATHAN BARON
University of Pennsylvania, USA

ABSTRACT

People are reluctant to harm some people in order to help others, even when the harm is less than the forgone help (the harm resulting from not acting). The present studies use hypothetical scenarios to argue that these judgments go against what the subjects themselves would take to be the best overall outcome. When the outcomes in question are income gains and losses for two groups of farmers, subjects judge the harm they would not impose through their action to be smaller than the harm they would impose through inaction. Some subjects refuse to reduce cure rates for one group of AIDS patients in order to increase cure rates more for another group, even when group membership was unknowable to anyone, so that, from each patient's point of view, the change would increase the probability of cure. Likewise, they resisted a vaccine that reduced overall mortality in one group but increased deaths from side effects in another group, even when, again, group membership was unknowable. Some people apply a do-no-harm principle to groups without apparent understanding of how such a principle might be justified in terms of its consequences. The capacity for such judgments makes them vulnerable to learning principles that have no justification at all.

INTRODUCTION

Judgments about the just distribution of costs and benefits seem to involve the use of a few basic principles: distribute outcomes equally; maximize some quantity (such as monetary value or utility); or distribute outcomes in proportion to contribution, need, or desire (Bar-Hillel and Yaari, 1993; Baron, 1993a, 1994a; Deutsch, 1975). Application of such principles need not — and often does not — produce the best overall outcome (Baron, 1993; Elster, 1993). For example, people sometimes want to punish harmdoers (in proportion to the harm they did) even when the punishment will not deter or prevent future harm (Baron and Ritov, 1993).

One principle of distribution is that it is wrong to harm some people in order to help others, even when the benefits outweigh the harm. The studies reported here concern mainly this 'do-no-harm' principle. This principle can be understood as an application of a rule against taking an action that leads to a worse

outcome for someone than would have occurred with no action. This is a reasonable rule in many situations, but it leads to undesirable results when it is not pitted against an equivalent rule against causing harm through *omission* (Spranca *et al.*, 1991).

For example, Ritov and Baron (1990) asked subjects to imagine that children had a 10 out of 10,000 chance of death from a flue epidemic, a vaccine could prevent the flu, but the vaccine itself could kill some number of children. Subjects were asked to indicate the maximum overall death rate for vaccinated children for which they would be willing to recommend a general policy for vaccinating children. Most subjects answered well below 9 per 10,000. Some would not tolerate any risk at all. (The same results were found when subjects were asked whether they would vaccinate their own child at various levels of risk.) These results and others (Baron, 1994b) suggest that many people are more concerned about the deaths that result from action — vaccinating — than about those that result from omission. Baron and Ritov (1994) found that this asymmetry between action and omission is limited to bad outcomes, that is, outcomes that are worse than the outcome produced by an another option. Hence the characterization of the results as a rule against 'doing harm'.

The vaccination study, and others reviewed by Baron (1994b), suggest that moral intuitions are sometimes nonutilitarian, that is, that people can *knowingly* favor options that will lead to something other than the best outcome aggregated over all those affected. Two objections have been made to this conclusion (e.g. Railton, 1994; Tetlock, 1994). First, the subjects themselves may think that the outcomes of their favored option *are* utilitarian, taking into account various consequences that the experimenter has not considered (and that the subjects did not mention when they were asked for justifications). The simplest way to deal with this objection is to ask the subjects to judge the outcomes directly. The first experiment here does that. The second objection — a specific form of the first — is that the utility of outcomes for those affected may differ as a function of the knowledge of those affected of how the outcome occurred. For example, death from a vaccine may be worse than death from a disease because those affected (the child, the parents, etc.) may be more emotionally upset by the former. The remaining experiments in this article deal with that objection by presenting scenarios in which the experience of those affected cannot account for the judgments found.

The examples here all concern the application of the do-no-harm principle to identified groups of people rather than individuals. People may be particularly reluctant to harm one group in order to help another group. For example, US President Clinton wanted a large tax on energy to reduce the budget deficit but finally settled for a small increase in the gasoline tax. Opposition to the energy tax seemed to be based on its harmful effects being concentrated in certain parts of the country. The winners and losers were identifiable for the energy tax, less so for other taxes. Of course, President Clinton may have been concerned about losing the support of certain legislators or the electoral votes of certain states. Real policy decisions are typically determined by many factors. If we want to isolate the effects of the processes of human judgment, laboratory studies are helpful.

Principles that concern fairness to groups — as distinct from fairness to individuals — are difficult to justify, even for nonutilitarians (e.g. Elster, 1993). Group membership is itself arbitrary. Each of us belongs to several groups, and each of these groups may be differently affected by some particular proposal such as an energy tax. For any proposal, arbitrary groups can be designated so that the proposal harms some of them more than others. (Again, I put aside political considerations here. For example, one reason for trying to avoid harming certain groups is to allay their historically well-founded fears of discrimination.)

Baron (1993b) found that the do-no-harm principle was applied to groups. Subjects were asked to put themselves in the position of a benevolent dictator of a small island populated by equal numbers of bean growers and wheat growers. The decision was whether to accept or decline the final offer of the island's only trading partner, as a function of its effect on the incomes of the two groups. Most subjects would not accept any offer that reduced the income of one group in order to increase income of the other,

even if the reduction was a small fraction of the gain. This effect was specific to changes from the status quo: subjects were also reluctant to make the same changes in reverse. Subjects thus showed a status quo effect (like that found by Samuelson and Zeckhauser, 1988, and others).

This result suggests that subjects favor options that do not yield what they would consider the best outcome. However, subjects may think that the harm from small losses is greater than the harm from large forgone gains, so they may think that their refusal to permit harm to one group was in fact the way to bring about the best overall outcome. The first experiment tests this possibility by asking subjects for their judgment of consequences.

Baron and Jurney (1993), applying this technique, presented subjects with six proposed reforms, each involving some public coercion, such as compulsory vaccination and tort reform involving elimination of lawsuits. Most subjects judged the reforms to be beneficial on the whole, but many of *these* subjects said that they would not vote for the reforms. Grounds for opposing such beneficial proposals included *unfairness* in the distribution of costs or benefits and *harm* to some people despite benefits to others. In one study, 39% of subjects said they would vote for a 100% tax on gasoline (to counter global warming), but 48% of those who opposed the tax thought that it would do more good than harm on the whole. Subjects would thus make decisions that were nonutilitarian by their own judgment of consequences. Of those subjects who would vote against the tax despite thinking that it would do more good than harm, 85% cited the unfairness of the tax as a reason for voting against it, and 75% cited the fact that the tax would harm some people. We also found a status quo effect: subjects were less likely to vote to repeal proposals than to vote against them if they were not already scheduled to go into effect.

The first study here extends this method, asking subjects both for favored decisions and judgments of their consequences, to decisions that involve tradeoffs of outcomes to identified groups. The consequence judgments involved comparison of effects on the two groups rather than overall judgments of the sort used by Baron and Jurney. The present study also gives subjects specific information about outcomes rather than asking them to imagine it. It therefore provides a different kind of demonstration that decisions can conflict with judgments of best consequences because of moral principles.

Another way in which apparently nonutilitarian judgment may be utilitarian after all is by taking into account the emotional reactions of those affected: people may be more upset about being harmed by acts than by omissions (as suggested by the results of Baron, 1992); people may be more affected by monetary losses than by equivalent gains (Kahneman and Tversky, 1979); or people may find harms to be more upsetting when the harm occurs because of their membership in an identified group than when it is randomly distributed among different groups. Subjects' understanding of these emotional effects may lead to apparently nonutilitarian judgments that are really utilitarian (in subjects' eyes) once the subjects' beliefs about emotional effects are taken into account. Of course, subjects are free to include emotions in their overall evaluation of consequences in the first study.

The remaining studies presented here, after the first, address this problem in a different way, by presenting subjects with hypothetical policy decisions in which the main factor that could lead to different emotional reactions (and therefore justify otherwise nonutilitarian judgments), group membership, was hidden from those affected. This method — making people 'blind' with respect to certain information — results in scenarios that are necessarily somewhat hypothetical. However, in a medical context, such situations are by no means unknown. Decisions about fetal testing, for example, must often be made without knowledge of important characteristics (Asch *et al.*, 1993).

EXPERIMENT 1: INCOME

The first study builds on the experiment of Baron (1993b) concerning judgments of the distribution of income between two groups (presumably equally deserving). Subjects were reluctant to harm one group

in order to help another, even when the increased income of one group was much greater than the lost income of the other. The present study, like the earlier one, asked subjects whether they would accept a decrease in average income for one group, e.g. $4000, in order to insure a larger increase of (e.g. $10,000) for the other group. Baron (1993b) found that subjects were reluctant to accept much decrease in one group's income even when the other group's increase was large.

Those results could be caused by loss aversion (Kahneman and Tversky, 1979). Such loss aversion would be consistent with utilitarianism if, in the subjects' judgment, the loss was experienced much more intensely by the losers than the gain by the gainers. A similar explanation could be applied to many other results in which subjects are reluctant to impose losses on one group for the sake of others. Even when this reluctance is affected by the way in which a situation is framed — e.g. as a wage cut versus a failure to raise wages to keep up with inflation — subjects may believe that framing affects actual emotional effects, and hence overall well-being (Kahneman *et al.* 1986). To test this account, I asked subjects to make direct judgments comparing the effects on well-being of various decreases and increases. I hypothesized that subjects would reject some decreases but would judge these decreases to have less effect on well-being than the foregone increases.

Method

Subjects were 39 students (20 females, 18 males, 1 unknown) from the University of Pennsylvania and the Philadelphia College of Pharmacy Science, solicited by advertisements and paid $6/hour for filling out questionnaires under supervision in a quiet room during specified hours. The questionnaire read

> Imagine that you are the president of a small island republic. You have the power to make treaties by yourself. You are engaged in annual trade negotiations with your sole trading partner, a much larger nation. Your entire economy is dependent on agricultural exports. Crops are grown by farmers who own their own farms. Half are bean growers, and half are wheat growers. Bean growers and wheat growers are equally diligent. They cannot switch products because of the conditions of the land. It is not their custom to help one another when one group is better of.
>
> In each of the following cases, imagine that negotiations have come to an end, and your trading partner has made a final offer for the next year, which you can accept or reject. For each offer, you are shown the present average annual income of each group and the average annual incomes that would result from accepting the offer. If you accept the offer, the new average incomes will be in effect for another year. If you reject the offer, the average incomes of the two groups will stay as they are now for another year. These income figures have about the same significance as they do in the U.S.

One form of the questionnaire began with Exhibit 1 called the *Harm table* henceforth (because it measures willingness to harm the bean growers in order to help the wheat growers).

Exhibit 1.

Bean growers			Wheat growers				
Present income		Income from accepting	Present income		Income from accepting	Accept	Reject
$30,000	⇒	$20,000	$30,000	⇒	$40,000		
$30,000	⇒	$22,000	$30,000	⇒	$40,000		
$30,000	⇒	$24,000	$30,000	⇒	$40,000		
$30,000	⇒	$26,000	$30,000	⇒	$40,000		
$30,000	⇒	$28,000	$30,000	⇒	$40,000		
$30,000	⇒	$30,000	$30,000	⇒	$40,000		

After checking the Accept or Reject column in each case, subjects were told 'Please check your answers. If you accepted an offer for one case and then rejected an offer for a case lower in the list, you may have misread the numbers.' (One subject was inconsistent in this way on all tables and was omitted. Otherwise, data were used whenever this consistency test was met.)

We might expect subjects to reject the first offer if they thought that the change from $30,000 to $20,000 had more effect on well-being than the change from £30,000 to $40,000. The question is how many offers they will reject. If they reject the third offer, for example, then they are giving up $4000 in order to avoid hurting the bean growers. This could result from a belief that moving from $30,000 to $24,000 has a greater effect on well-being than moving from $30,000 to $40,000, or it could result from a do-no-harm heuristic. To distinguish these explanations, a later question asks directly about well-being.

In a second table, the Pareto table, the 'Income from accepting' was identical to the Harm table, but the 'Present income' was $20,000 for bean growers and $40,000 for wheat growers. All offers except the first were therefore improvements for the bean growers, and the wheat growers were unaffected. (Hence, all these offers were Pareto improvements.) I expected subjects to accept all offers after the first, unless they based their judgments on final outcomes rather than changes. Most subjects did as I expected, so this table has little importance.

In a third table (Exhibit 2) the numbers in the 'Income from accepting' and 'Present income' columns were reversed from those in the Harm table. This is called the *Status quo table* because comparison with the Harm table measures the status quo effect (as found by Baron, 1993b).

Exhibit 2.

Bean growers			Wheat growers				
Present income		Income from accepting	Present income		Income from accepting	Accept	Reject
$20,000	⇒	$30,000	$40,000	⇒	$30,000		
$22,000	⇒	$30,000	$40,000	⇒	$30,000		
$24,000	⇒	$30,000	$40,000	⇒	$30,000		
$26,000	⇒	$30,000	$40,000	⇒	$30,000		
$28,000	⇒	$30,000	$40,000	⇒	$30,000		
$30,000	⇒	$30,000	$40,000	⇒	$30,000		

In particular, if subjects judged only according to *levels* of income, ignoring the direction of *changes*, then their responses to the Status quo table should be the mirror image of the responses to the Harm table, i.e. with 'Accept' and 'Reject' reversed. If, on the other hand, subjects regard losses as more serious than gains, then they would accept fewer offers than they rejected in the Harm table. The loss from $40,000 to $30,000 for the wheat growers would be the main factor in their decision. (On the other hand, a desire for equality of final outcome should incline subjects to accept more offers in the Status quo table and reject more offers in the Harm table, the opposite pattern of results.)

In two more tables, subjects were asked 'Finally, consider the following changes in income for one group or the other. In each case, one change increases income and the other decreases income. Indicate *which of the two changes has a greater effect on the overall well-being of the group it affects*. Changes of the same amount of money do not have to have equal effects on overall well-being.' The two tables that followed contained all items from the Harm table and the Status quo table, respectively, except the last item in each (where the answer was obvious). These tables were called Harm-judgement and Status quo-judgement respectively: Instead of 'Accept' and 'Reject', the columns at the right were called 'First

change greater' and 'Second change greater', and a final column was called 'Effects equal'. From the judgments made here about which change was greater, we can infer what subjects would have said in the corresponding tables if they based their choice on their judgments of effects on well-being rather than basing it on their role in bringing about those changes. For example, if a utilitarian subject said that the loss from $30,000 to $22,000 had a greater effect than the gain from $30,000 to $40,000, this subject should reject the agreement that causes both of these changes.

In a second form of the questionnaire, the order of the Harm and Pareto tables was reversed, as was the order of Harm-judgment and Status quo-judgment and the order of items within each table. The difference between forms had no effect on the responses.

Results

Subjects were, on average, willing to reduce the income of bean growers by $4154 (s.e. $500) in order to increase the income of wheat growers by $10,000 in the Harm table, but they judged that a reduction of $6758 (s.e. $608) was equivalent to an increase of $10,000 in the Harm-judgment table. The difference was significant ($p = 0.001$, one-tailed Wilcoxon test; the test was done after all differences between the two measures were reduced in absolute value by $1000, to take into account the fact that judgments of equal effects were permitted in the judgment tables). Eighteen of the 39 subjects showed this effect, and five showed the reverse effect. The magnitude of the effect did not vary with sex. In sum, many subjects' reluctance to harm the bean growers was not accounted for by their own well-being judgments for the two groups. It is also of interest that 16 subjects showed agreement between judgments and choices.

The equivalent mean willingness to sacrifice total income for the sake of equality in the Status quo table was $1272 (s.e. $476). This was significantly less than in the Harm table ($4154; $p = 0.000$, Wilcoxon test), thus indicating a status quo effect. In contrast to the Harm table, mean judgments in the Status quo table ($1272) were not significantly different from those predicted from the Status quo-judgment table (mean $2190, s.e. $502). Moreover, the difference between Harm-judgment and Status quo-judgment was not significant, and the magnitude of the status quo effect for Harm versus Equality was greater than than for the corresponding judgment tables ($p = 0.001$, one-tailed Wilcoxon test). The status quo effect for choices thus cannot be accounted for in terms of beliefs about the effects of losses versus gains on well-being.

Ten subjects were unwilling to accept offers that improved the bean growers' income without hurting the wheat growers, in the Pareto table. All the results just described were still significant when these subjects were eliminated.

The main result can be seen as kind of preference reversal between choice and judgment. Subjects rejected some agreements when they made a choice but then judged that the net benefit from accepting is greater than the net harm. In this regard, the result is analogous to that of Bazerman *et al.* (1992), who found that choices involving allocation between self and other were more affected by the level of payoff to the self than by the difference between payoff to self and payoff to other, while judgments of satisfaction or acceptability were relatively more affected by the self–other difference. In the present study, however, the question is one of conflicting principles of fairness (utilitarian versus do-no-harm) rather than fairness versus self-interest.

In sum, subjects are unwilling to accept some agreements despite judging that the harm from them is less than the benefit. The do-no-harm heuristic appears to be used even when it goes against people's own judgments of utility. People are knowingly nonutilitarian.

EXPERIMENT 2: AIDS

The remaining experiments address the same issue in a different way, by manipulating the distribution of gains and losses to groups, independently of the distribution over individuals. To make sure that group distribution is not seen as affecting individual well-being, subjects are told that the individuals do not know their group membership. Group membership thus cannot affect anyone's utility. If subjects take it into account, they are, again, nonutilitarian in a way that would be obvious if they thought about effects on individuals.

In Experiment 2, subjects evaluated hypothetical treatments for AIDS. Two types of AIDS are equally likely, but patients cannot know their type. In some cases, a current treatment cures some patients of both types, and subjects decide whether to accept a new treatment will cure fewer patients of one type but more of the other type. For example, a treatment that cures 50% of each type could be replaced with a treatment that cures 80% of one type but 30% of the other type. Subjects who consider overall cure rates only would accept this change. More individuals are helped, and the individual patients would all favor the change if they cared about their overall probability of cure, since they do not know their type. Subjects who want to avoid excessive harm to one type would reject this change, but they would accept a change from 70%–30% to 80%–30%; here, the overall probability of cure is the same before the change as in the original example, and the same after the change. Judgments that take type into account in this way cannot be justified in terms of the reactions of individual patients, who know only whether the treatment works or not, or in terms of the distribution of outcomes over individuals (ignoring type).

In other cases, no current treatment is available, and subjects evaluated new treatments that cure some patients but kill others through side effects. Patients who die from the side effects are either in the same group as those who are cured or in the other group. At issue is whether the maximum tolerable risk of death depends on whether it occurs in the same patient group that benefits or in the other group. Note that both kinds of cases involve the sacrifice of *individual* patients for the benefit of others, since patients do not know their group.

Method
Seventy-three subjects, recruited and tested as in Experiment 1, completed a brief questionnaire, which began

> You are a government official in charge of approving new treatments for AIDS, a few years from now. You are asked to consider several hypothetical treatments. Treatments can be used on people infected with the AIDS virus as well as on those who show symptoms of the disease. By this time it is known that the disease has different types, A and B, with *equal numbers of patients having each type*. Some of the treatments have different effects on one type or another. However, for various complex reasons, *it will be impossible for anyone to discover which type each patient had*. For each patient, we will know only whether the disease was cured or not.

Subjects were then told to make a decision about whether to approve each treatment. They were told that any approved treatment would be used by some patients. Finally, the main points were summarized.

Subjects rated four sequences of treatments. In the first two sequences, the 50–50 sequence and the 70–30 sequence, subjects saw cure rates for current and proposed treatments. Each treatment was presented in a table like the following:

1	Type A	Type B	Overall
Current treatment	50%	50%	50%
Proposed treatment	80%	0%	40%

In the 50–50 sequence, the current treatment cured 50% of each type, as in the example just given. In the 70–30 sequence, the current cure rates were 70% and 30% for Types A and B respectively. In both sequences, the proposed-treatment cure rate for Type B went from 0% to 55% successively in steps of 5 (and the overall cure rate went from 40% to 67.5% in steps of 2.5%). Presumably most subjects would not approve the new treatment until the overall cure rate was 50% or more. The hypothesis is that they will require a higher cure rate in the 50–50 sequence than in the 70–30 sequence because Type B is harmed more in the 50–50 sequence.

The other two sequences were called same-groups and different-groups. Subjects were told 'In the following cases, side effects of some treatments can cause death within a year. The life expectancy of an untreated patient is 7 years. There is no current treatment.' Again, subjects decided whether to approve each treatment. The first case in the same-groups sequences was as follows, and thereafter the death rate increased in steps of 5%, up to 50%.

1	Type A	Type B	Overall
Cure rate	40%	40%	40%
Death rate	0%	0%	0%

The first different-groups case was:

1	Type A	Type B	Overall
Cure rate	80%	0%	40%
Death rate	0%	0%	0%

Here, the death rate for Type A remained at 0% for all cases, but the death rate for Type B increased in steps of 10% up to 100%, so that the overall death rate increased up to 50% (as in the same-groups sequence). Thus, the two sequences differed in whether the harms occurred in the same group as the benefits or in a different group. The hypothesis is that subjects would accept a lower death rate in the different-groups condition, because one group had to be harmed in order to help the other group. From the point of view of individuals, the situation in the two sequences was the same, however. Each individual would either be cured or die, and the probabilities were those in the Overall column, since the individuals did not know their types.

Thirty-five subjects did the sequences in the order just described, and 38 did them in the order: 70–30, 50–50, different-groups, same-groups. (Ten other subjects were excluded because they approved better cases and disapproved worse ones.)

Results

As hypothesized, subjects required a higher Type-B cure rate for the proposed treatment when the current treatment had a Type-B cure rate of 50% than when it had Type-B cure rate of 30%. Less harm was done to Type B in the latter case. The mean minimally acceptable Type-B cure rate was 22.0% for the 70–30 sequence and 24.3% for the 50–50 sequence ($p = 0.004$, Wilcoxon test), with 21 subjects giving higher values in the 50–50 sequence and nine giving higher values in the 70–30 sequence. (The remaining subjects all gave the same answer for both conditions, which is why the difference of means was so small yet still statistically significant.)

The mean minimally acceptable overall death rate was 18.6% in the different-groups condition and

20.7% in the same-groups condition ($p = 0.012$, Wilcoxon test for the difference). Sixteen subjects gave a higher value in the different-groups condition, five in the same-groups condition. (Again, the remaining subjects all gave the same answer for both conditions.) Lethal side effects were again more tolerable when they occur in the same groups that benefit, even if they occur to different individuals.

Both of these effects were also found in another study (not reported in the interest of space), in which no overall statistics, were provided and in which subjects were asked to indicate the minimum cure rate or maximum death rate at which they would approve each treatment. The differences of the means were larger than those here, perhaps because the overall rate was not provided. The overall rate may have encouraged subjects to think in a utilitarian manner.

At the end of the questionnaire, subjects were asked 'Do you think it is possible to know anything about which disease each patient had before beginning the treatment?' Twenty-nine subjects answered 'yes' (34 'no', 10 no answer); some of these subjects remarked that future research might find a way, so they did not necessarily misunderstand the instructions. The magnitude of the two effects found, however, did not depend significantly on the answer to this question ($\tau = 0.03$ for 50–50 versus 70–30, -0.07 for same-groups versus different-groups). Nor did the magnitude depend on order.

Subjects were also asked for comments. Those few subjects who provided reasons for distinguishing treatments with the same overall characteristics typically said that they based their decisions on the (cure or death) rates for each group separately. They did not say how they combined these two judgments, but it seems plausible that they gave more weight to the group that was worse off.

EXPERIMENT 3: VACCINATIONS

Vaccinations sometimes cause side effects as bad as the disease they prevent. Subjects evaluated hypothetical vaccines in which the benefits and harmful side effects are distributed differently to two groups of children. Again, no one can tell which group a child is in. Otherwise, these cases are similar to those used by Ritov and Baron (1990), described earlier. If subjects apply the do-no-harm heuristic to groups, they will be more reluctant to vaccinate when the benefits and harms occur in different groups, even though the distribution across individuals, and the probabilities faced by each individual, are unaffected by the distribution of harms and benefits across groups.

Method

This study involved four items (here identified as 1–4) added to the end of a questionnaire following up the results of Ritov and Baron (1990). Fifty-three subjects were told 'Research has discovered two types of children, group A and group B. Groups A and B are of equal size. These groups can differ in their susceptibility to death from the disease and from the vaccine. The test to determine which group your child is in cannot be given to living children, so nobody can find out which group a given child is in. Group membership is not related to sex or any other observable characteristic.' The cases were then presented in the form:

For group A:
Twelve out of 10,000 unvaccinated children will die from the disease in the coming year.
Three out of 10,000 vaccinated children will die from the vaccine.
For group B:
Six out of 10,000 unvaccinated children will die from the disease in the coming year.
Nine out of 10,000 vaccinated children will die from the vaccine.
What is the probability that you would vaccinate your child?

Exhibit 3 shows the figures for the cases used. The cases were repeated from the viewpoint of a policy maker who had to decide whether to make the vaccine compulsory. The results for the two series did not differ significantly and were therefore combined. Note that all cases are the same in the total risk reduction from vaccination. (Other cases in the questionnaire did not reduce risk by this amount.)

Exhibit 3. Outcomes and subjects' probability of vaccinating [p(vac)] for the cases in Experiment 3 and the replication

Case	Group A		Group B		p(vac)	replication
	unvac	vac	unvac	vac		
1	12	3	6	9	0.580	0.667
2	9	3	9	9	0.608	0.725
3	6	3	12	9	0.670	0.723
4	9	6	9	6	0.713	0.730

The study was replicated without most of the additional cases, and with the order of cases counterbalanced (19 in the original order, 17 in the reverse order).

Results

Subjects were reluctant to harm one of the groups to help the other. Exhibit 3 shows the mean probabilities that subjects provided for the original study and the replication. An analysis of variance showed that the cases differed ($p = 0.000$) in both the original and the replication. In the original study, all differences were significant ($p = 0.05$) except that between cases 1 and 2. Subjects were most reluctant to vaccinate in case 1, where one group is harmed in order to help the other, even though the benefit is greater than the harm, and in case 2, where only one group was helped by the vaccination. Subjects were most willing in case 4, where both groups are equal not only in the benefit of vaccination but also in the final level of risk. In the replication, case 1 differed from all other cases, which did not differ from each other. Here, subjects were more reluctant to vaccinate only when one of the groups was harmed by the vaccination. Both studies agree that subjects are sensitive to group differences even when the overall benefit is the same and when group membership is unknown. Comparing cases 1 and 3 (which are matched in both initial and final distribution of risk), 28 (out of 53) subjects in the original study were more reluctant to vaccinate in case 1 (where one group is harmed), and 8 were more reluctant in case 3; in the replication, the respective numbers were 14 (out of 36) versus 5.

EXPERIMENT 4: VACCINATION WITH SUMMARY STATISTICS

Experiment 4 provided summary statistics for both groups combined as well as statistics for individual groups.

Method

Thirty-seven subjects completed a questionnaire like that used in Experiment 3. Cases, shown in Exhibit 4, were presented in the following format:

Death rate	Group A	Group B	Both groups
Vaccinated	0/10,000	12/10,000	6/10,000
Unvaccinated	9/10,000	9/10,000	9/10,000

Sixteen subjects completed the items in the order indicated by case numbers in Exhibit 4. The cases were first presented in the parent's perspective, then the policy maker's. Twenty-one subjects did the cases in the reverse order.

Exhibit 4. Outcomes and subjects' mean probabilities of vaccinating for the cases in Experiment 4

Case	Group A		Group B		p(vac)
	unvac	vac	unvac	vac	
2	12	3	12	15	0.432
4	12	3	6	9	0.541
1	9	0	9	12	0.565
3	6	3	12	9	0.736
5	9	6	9	6	0.762

Results

Again, subjects were reluctant to harm one group in order to help another. Exhibit 4 shows the mean probabilities of vaccinating. Order of presentation had no effect, nor did perspective (parent versus policy maker), so the results are combined. A Friedman nonparametric 'analysis of variance' showed that the five cases differed ($p = 0.000$). Sign tests were used for comparisons of individual cases because (unlike Experiment 4), many subjects gave values of 1 or 0 for probability of vaccinating, so quantitative differences were unlikely to be meaningful. (Results were essentially the same with Wilcoxon tests.) In essence, probability of vaccination was lower in cases in which one of the groups was harmed, cases 1, 2, and 4; all differences between these cases and cases 3 and 5 were significant at $p = 0.003$ or better. No other comparisons were significant except for cases 1 and 2 ($p = 0.05$), which suggests that subjects might value reducing risk to zero in one of the groups (as found by Ritov *et al.*, 1993). (Subjects were also asked whether group membership could be detected. Only five said it could. These five did not differ from others in the magnitude of the effect.)

A few subjects wrote comments explaining their reasons for distinctions among the cases. These reasons were consistent with the 'do-no-harm' principle. For example, ' ... I wouldn't [make the vaccine compulsory in cases 1, 2, and 4] because the vaccine can be a greater risk to some than the disease, and, while the vaccine could help *some*, I couldn't *force* it on anyone.'

CONCLUSION

Many subjects were reluctant to harm one group to benefit another group more. This reluctance cannot be explained by any utilitarian considerations. The benefit is clearly greater than the harm in the AIDS and vaccination cases, and in the income study (Experiment 1) subjects acknowledge that the benefit is greater than the harm. Subjective responses to harm (in those affected) cannot account for the effects in at least the AIDS and vaccination experiments, because group membership is unknown.

The idea of harming no identifiable group is, of course, a good heuristic in political contexts, where consent of all groups is required. But that is not the case here. In the AIDS and vaccine cases, consent is impossible because the members do not know who they are. The very existence of groups can be seen as an artificial way of dividing up those affected. The results can be seen as a kind of framing effect, caused by unreflective overgeneralization of rules.

Such overgeneralization seems 'unreflective' because the rules that people use cannot apparently be

justified as serving any purpose. They seem, instead, to be pure intuitions, thought to be right simply because they feel right. This may also be true of some cases in which people are attached to principles of group fairness in the real world. Of course, this account of these intuitions may be wrong. They may be justifiable in terms of some deeper, but still nonutilitarian, purpose. I have tried to shift the burden of proof to those who believe that such purposes exist (Baron, 1993c).

Most subjects in fact do not have these intuitions. In a within-subject design, which encourages subjects to compare their responses to different cases, most responses are consistent with utilitarianism. (In other cases, subjects may not take the opportunity to compare responses, so there is also some random error in both directions.) The do-no-harm principle is therefore not an inevitable 'cognitive illusion', but, rather, a heuristic rule used some of the time by some people. If this principle is applied even when it cannot be justified in terms of outcomes, then people may be capable of learning other rules that cannot be justified. Such a capacity may be a cause for concern.

ACKNOWLEDGEMENTS

REFERENCES

Asch, D. A., Patton, J. P., Hershey, J. C. and Mennuti, M. T. 'Reporting the results of cystic fibrosis carrier screening', *American Journal of Obstetrics and Gynecology*, **168** (1993), 1–6.

Azzi, A. E. 'Procedural justice and the allocation of power in intergroup relations: Studies in the United States and South Africa', *Personality and Social Psychology Bulletin*, **18** (1992), 736–747.

Bar-Hillel, M. and Yaari, M. 'Judgments of distributive justice', in Mellers, B. A. and Baron, J. (eds), *Psychological Perspectives on Justice: Theory and applications*, pp. 56–84, New York: Cambridge University Press, 1993.

Baron, J. 'The effect of normative beliefs on anticipated emotions', *Journal of Personality and Social Psychology*, **63** (1992), 320–330.

Baron, J. 'Postscript', in Mellers, B. A. and Baron, J. (eds), *Psychological Perspectives on Justice: Theory and Applications*, pp. 315–330, New York: Cambridge University Press, 1993a.

Baron, J. 'Heuristics and biases in equity judgments: a utilitarian approach', in Mellers, B. A. and Baron, J. (eds), *Psychological Perspectives on Justice: Theory and Applications*, pp. 109–137, New York: Cambridge University Press, 1993b.

Baron, J. *Morality and Rational Choice*, Dordrecht: Kluwer, 1993c.

Baron, J. *Thinking and Deciding*, 2nd edition, New York: Cambridge University Press, 1994a.

Baron, J. 'Nonconsequentialist decisions (with commentary and reply)', *Behavioral and Brain Sciences*, **17** (1994b) 1–42.

Baron, J. and Jurney, J. 'Norms against voting for coerced reform', *Journal of Personality and Social Psychology*, **64** (1993), 347–355.

Baron, J. and Ritov, I. 'Intuitions about penalties and compensation in the context of tort law', *Journal of Risk and Uncertainty*, **7** (1993), 17–33.

Baron, J. and Ritov, I. 'Reference points and omission bias', *Organizational Behavior and Human Decision Processes*, **59** (1994), 475–498.

Bazerman, M. H., Loewenstein, G. F. and White, S. B. 'Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives', *Administrative Science Quarterly*, **37** (1992), 220–240.

Deutsch, M. 'Equity, equity, and need: What determines which value will be used as the basis of distributive justice?' *Journal of Social Issues*, **31** (1975), 137–149.

Elster, J. 'Justice and the allocation of scarce resources', In Mellers, B. A. and Baron, J. (eds), *Psychological Perspectives on Justice: Theory and Applications*, pp. 259–278, New York: Cambridge University Press, 1993.

Kahneman, D., Knetsch, J. L. and Thaler, R. 'Fairness as a constraint on profit seeking: Entitlements in the market', *American Economic Review*, **76** (1986), 728–741.

Kahneman, D. and Tversky, A. 'Prospect theory: An analysis of decisions under risk', *Econometrica*, **47** (1979), 263–291.

Railton, P. 'Broadening the base for bringing cognitive psychology to bear on ethics', *Behavioral and Brain Sciences*, **17** (1994), 27–28.

Ritov, I. and Baron, J. 'Reluctance to vaccinate: omission bias and ambiguity', *Journal of Behavioral Decision Making*, **3** (1990), 263–277.

Ritov, I. and Baron, J. 'Status-quo and omission bias', *Journal of Risk and Uncertainty*, **5** (1992), 49–61.

Ritov, I., Baron, J. and Hershey, J. C. 'Framing effects in the evaluation of multiple risk reduction', *Journal of Risk and Uncertainty*, **6** (1993), 145–159.

Samuelson, W. and Zeckhauser, R. 'Status-quo bias in decision making', *Journal of Risk and Uncertainty*, **1** (1988), 7–59.

Spranca, M., Minsk, E. and Baron, J. 'Omission and commission in judgment and choice', *Journal of Experimental Social Psychology*, **27** (1991), 76–105.

Tetlock, P. E. 'The consequences of taking consequentialism seriously', *Behavioral and Brain Sciences*, **17** (1994), 27–28.

Author's biography:
Jonathan Baron is Professor of Psychology at the University of Pennsylvania. He is interested in the relations among decision making, moral judgement, and public issues.

Author's address:
Jonathan Baron, Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, PA 19104–6196, USA (e-mail baron@cattell.psych.upenn.edu).