The effects of overgeneralization on public policy

Jonathan Baron[1]

University of Pennsylvania

December 20, 2000

[1]This work was supported by National Science Foundation Grant SES 9876469. Send correspondence to Jonathan Baron, Department of Psychology, University of Pennsylvania, 3815 Walnut St., Philadelphia, PA 19104–6196, or (e-mail) baron@cattell.psych.upenn.edu.

Introduction

Let me begin with a few examples from psychology. When children learn the meanings of words such as "doggie," they often apply the term to cats or hamsters. A little later, when faced with arrays like the following,

```
   A     A     A     A

 B           B           B


 C C C C C C C C C C C C
D  D  D  D  D  D  D  D  D
```

they say that there are more Ds than Cs, responding to the length of the two rows rather than the number. While they correctly acknowledge that the D row is longer than the C row, and that the the A row has more letters than the B row, they mistakenly say that the A row is longer than the B row. In essence, they conflate length and number, failing to distinguish them, much as younger children fail to distinguish cats and dogs. When they learn that the B row is in fact longer than the A row, they often come to assert that it also has more letters (Lawson, Baron, & Siegel, 1974; Baron, Lawson, & Siegel, 1975).

When they are somewhat older, they learn to compute the area of a parallelogram by multiplying the base times the height. Some children then apply the same formula to a trapezoid (Wertheimer, 1959). All three of these examples can be seen as cases of overgeneralization. A rule is learned which works in one case because it is correlated with the correct rule. It is overgenerlized to a case where the correlation is broken. Of course, every case of overgeneralization is accompanied by a case of undergeneralization. If an incorrect rule is applied then a correct rule is not applied, perhaps because it isn't known at all. It might be more appropriate to call these cases of misgeneralization rather than overgeneralization, but it is the latter that we see.

Yet another way of describing the situation is that people have too few rules. In the case of dogs and cats, length and number, or parallelograms and trapezoids, people use one rule when they should use two. (Later I shall argue that, in some cases, the problem is exactly the opposite from what it seems to be, that is,

that people have many rules when one would do.)

The parallelogram/trapezoid case is in some ways the clearest. This is because the application of the formula to the trapezoid is necessarily wrong. We could choose to apply the same word to cats and dogs, but we would have to end the concept of area as we know it in order to say that the area of a trapezoid is the base times the height. We understand the *purpose* of the concept of area. For example, area should predict the amount of paint that is needed to cover something. Although we have a clear standard in this case, the psychological mechanism of the mistake might be much the same as in the cases of length/number and dog/cat.

I want to argue here that the same kinds of psychological mechanisms affect our intuitions about public policy. By intuitions, I mean our judgments of better and worse, of the sort that come to mind without any analysis. These are our initial reactions to proposals about policy, from tax cuts to oil drilling in Alaska.

Moreover, these intuitions can be studied in psychology experiments, and their effect can be seen in the policy world itself. When the experiments show phenomena that parallel what we observe in the real world of policy, we gain confidence that the same psychological mechanisms are involved. We can never be certain, but psychology experiments can provide us with additional evidence, perhaps enough to take action.

I shall give three examples: omission bias, protected values, and the morality-as-self-interest illusion. Omission bias is the bias toward greater toleration of sins of omission than sins of comission, even if the consequences of omission are worse than those of action. Examples from policy range from the prohibition of active euthanasia to the slow approval of new drugs to the neglect of the world's poor. This bias results from a rule against harm caused by action, a rule that is usually correlated with avoiding bad outcomes.

Protected values are values that people see as absolute, not to be sacrificed no matter how great the benefit. These vary considerably, but they range from environmental protection to religious prohibitions. These can be understood as analogous to poorly formed concepts, like the young child's concept of "doggie."

The morality-as-self-interest illusion results from a failure to distinguish morality and self-interest, much like the failure to distinguish length and number. Usually, the two are correlated, but, when they are

not, people have trouble understanding the conflict between the two.

Omission bias

Spranca, Minsk, and Baron (1991) found a bias toward omissions in situations that were more obviously moral than those discussed so far. In one scenario, John, the best tennis player at a club, wound up playing the final of the club's tournament against Ivan Lendl (then ranked first in the world). John knew that Ivan was allergic to cayenne pepper and that the salad dressing in the club restaurant contained it. When John went to dinner with Ivan the night before the final, he planned to recommend the house dressing to Ivan, hoping that Ivan would get a bit sick and lose the match. In one ending to the story, John recommended the dressing. In the other, Ivan ordered the dressing himself just before John was about to recommend it, and John, of course, said nothing. When asked whether John's behavior is worse in one ending or the other, about a third of the subjects said that it was worse when he acted. These subjects tended to say that John did not cause Ivan's illness by saying nothing. (In reply, it might be said that the relevant sense of "cause" here concerns whether John had control over what Ivan ate, which he did.)

Ritov and Baron (1990) examined a set of hypothetical vaccination decisions modeled after such cases. In one experiment, subjects were told to imagine that their child had a 10 out of 10,000 chance of death from a flu 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 vaccinate their child. Most subjects answered well below 9 per 10,000. Of the subjects who showed this kind of reluctance, the mean tolerable risk was about 5 out of 10,000, half the risk of the illness itself. The results were also found when subjects were asked to take the position of a policy maker deciding for large numbers of children. When subjects were asked for justifications, some said that they would be responsible for any deaths caused by the vaccine, but they would not be (as) responsible for deaths caused by failure to vaccinate.

In sum, "omission bias" is the tendency to judge acts that are harmful (relative to the alternative option) as worse than omissions that are equally harmful (relative to the alternative) or even more harmful (as in the vaccination case) (Baron and Ritov, 1994). In any given case, some people display this bias and others

do not.

Omission bias is related to issues of public controversy, such as whether active euthanasia should be allowed. Most countries (and most states of the United States) now allow passive euthanasia, the withholding of even standard medical treatment for those who are judged to be no worse off dead than alive, but active euthanasia is almost everywhere banned even for those who wish to die. Opponents of active euthanasia can, of course, find other arguments against it than the fact that it is "active." But it is possible that these arguments would not be seen as so compelling if the distinction between acts and omissions were not made.

Another example is the vaccine for pertussis, a bacterial infection that causes whooping cough and that can cause infants and toddlers. Pertussis is the "P" in "DPT," a commonly used combination vaccine that immunizes babies against diphtheria and tetanus as well. Some governments have not required pertussis vaccination in any form because it causes serious side effects and perhaps even death in a tiny fraction of a percent of children vaccinated. The officials responsible feel that the harm from requiring the vaccine is worse than the harm from doing nothing, even if the latter leads to more disease and death.

Before the DPT vaccine, about 7,000 children died each year from whooping cough caused by pertussis infection. The death rate from whooping cough is now less than 100 per year. (Many children are still not fully vaccinated.) Despite this record of success, people do not like the idea of causing a disease with a vaccine. When a few cases of brain damage apparently caused by DPT vaccine were reported in England and Japan in the mid 1970s, requirements for vaccination lapsed and many children were not vaccinated. In Great Britain, rates of vaccination fell from 79% in 1971 to 37% in 1974. A two-year epidemic that followed killed 36 people. The epidemic ended when vaccination rates increased. Similar epidemics occurred in Japan, also involving 30 to 40 deaths per year. (Joint Committee on Vaccination and Immunization, 1981; Smith, 1988). We have found that mothers who resist DPT vaccination for their children show a lower mean tolerable risk than mothers who accept the vaccination (Meszaros et al., 1992; Asch et al., 1993).

Polio vaccine is similar. The Sabin polio vaccine can cause polio. The Salk vaccine can fail to prevent it. The total risk of polio was higher with the Salk vaccine, but some (for example, Deber and Goel, 1990)

argued against the Sabin vaccine because it causes polio through the action of vaccinating. (By 1999, polio had largely disappeared from developed countries, so the situation has changed.)

Omission bias could also justify a lack of concern with the problems of others (Singer, 1993). For example, much of the world's population lives in dire poverty today and into the foreseeable future. People — even people who take an interest in social issues — often think that they are not responsible for this poverty and need do nothing about it. It can be argued, however, that with a little effort we can think of all sorts of things we can do that will help the situation immensely at very low cost to ourselves, such as supporting beneficial policies. Failure to do these things can be seen as a harm, but many people do not see it that way. More generally, omission bias helps people believe that they are completely moral if they obey a list of prohibitions while otherwise pursuing their narrow self-interest.

Omission bias is somewhat labile. It can be reduced by the following kind of instructions (Baron, 1992), which are illustrated for the vaccine case just described:

> The questions you have just answered concern the distinction between actions and omissions. We would like you now to reconsider your answers. First, read this page and think about it. Then answer the questions again on the next page. (Do not go back and change your original answers.) Feel free to comment as well.
>
> When we make a decision that affects mainly other people, we should try to look at it from their point of view. What matters to them, in these two cases, is whether they live or die.
>
> In the vaccination case, what matters is the probability of death. If you were the child, and if you could understand the situation, you would certainly prefer the lower probability of death. It would not matter to you how the probability came about.
>
> In cases like these, you have the choice, and your choice affects what happens. It does not matter what would happen if you were not there. You are there. You must compare the effect of one option with the effect of the other. Whichever option you choose, you had a choice of taking the other option.
>
> If the main effect of your choice is on others, shouldn't you choose the option that is least bad for them?

Of interest, this instruction influenced not only what people thought they should do but also whether they thought they would feel guiltier after vaccinating or not, if the bad outcome happened. People's expected emotions are correlated with their moral opinions, but the opinions seem to drive the expectations.

Protected values

People think that some of their values are protected from tradeoffs with other values (Baron & Spranca, 1996; Tetlock et al., 1996). Many of these values concern natural resources such as species and pristine ecosystems. People with protected values (PVs) for these things do not think they should be sacrificed for any compensating benefit, no matter how small the sacrifice or how large the benefit. The term "value" here is use to mean "utility," that is, the measure of desirability that decisions are meant to increase (Baron, 2000). In an economic sense, when values are protected, the marginal rate at which one good can substituted for another is infinite. For example, no amount of money can substitute for a type of environmental decline.

PVs concern rules about action, irrespective of their consequences, rather than consequences themselves. What counts as a type of action (e.g., lying) may be defined *partly* in terms of its consequence (false belief) or intended consequence, but the badness of the action is not just that of its consequences, so it has value of its own. Omission bias is greater when PVs are involved (Ritov & Baron, 1999). For example, when peole have a PV for species, they are even less willing to cause the destruction of one species in order to save even more species from extinction. Thus, PVs apply to acts primarily, as opposed to omissions.

People think that their PVs should be honored even when their violation has no consequence at all. People who have PVs for forests, for example, say that they should not buy stock in a company that destroys forests, even if their purchase would not affect the share price and would not affect anyone else's behavior with respect to the forests. This is an "agent relative" obligation, a rule for the person holding the value that applies to his own choices but not (as much) to his obligations with respect to others' choices. So it is better for him not to buy the stock, even if his not buying it means that someone else will buy it.

PVs are at least somewhat insensitive to quantity. People who hold a PV for forests, tend to say that it is just as bad to destroy a large forest as a small one (Ritov & Baron, 1998). They say this more often than

they say the same thing for violations of non-protected values (NPVs).

PVs tend to apply to acts, not omissions, and to tradeoffs with gains in other values, not losses (Ritov & Baron, 1999). People with PVs show more omission bias: they are less willing than others to destroy one forest to prevent destruction of other forests, holding the destruction constant. It is not the destruction that matters to them so much as the destroying, the act. And PVs are somewhat limited to tradeoffs with gains. Some people are unwilling to destroy a forest in order to promote "economic" gain, but they are willing to destroy it in order to prevent a loss. Of course, what counts as a gain or loss is relative and easily manipulated (e.g., Shafir et al., 1998). Such differences in description are relevant to classifications of acts even though they are not relevant to evaluation of consequences. In sum, protected values seem to be deontological. They go against maximizing consequences.

Several researchers have noted that PVs cause problems for quantitative elicitation of values, as is done in cost-benefit analysis or decision analysis (Baron, 1997; Bazerman et al., 1999). Methods used in such elicitations include contingent valuation, decision analysis, and conjoint analysis (Baron, 1997). In principle, such methods permit optimal allocation of scarce resources.

But application of these methods often requires measurement of values — in terms of money or utility. When some value has infinite utility for some person, it swamps everything else and makes this sort of analysis impractical. In practical terms, we cannot spend all our resources on protecting the environment, saving human lives, protecting human rights, or any one thing. We must make tradeoffs. Suppose the measure of value is willingness to pay. If we try to find the average willingness to pay more taxes (say) in order to save a forest, and if some people say that the forest has infinite value, the average will be infinite, regardless of what others say. Increased expenditures on one good will involve sacrifices of other goods, such as human life and health, and other people might claim absolute values for these. If that happens, no decision is possible based on the measures of values. Even if only one side of the equation has an absolute value, we could almost never honor it in practice. The two problems are thus: 1., the possibility that one person could dominate a decision by expressing an absolute value, and, 2., the possibility that people with conflicting absolute values can make a decision impossible. The problems can happen with any measure of value that allows expression of PVs.

Some writers see the problem of refusal to make tradeoffs as the basis of a philosophical objection to cost-benefit analysis. Anderson (1993), for example, criticizes cost-benefit analysis because it "ignores the possibility that goods such as endangered species may be specially valued as unique and irreplaceable higher goods. The distinction between higher and lower goods, which supports norms that prohibit certain tradeoffs between them, plays no part in the analysis" (pp. 193–194). Other writers speak of "protest responses" to questions about economic values of resources such as the dollar value of environmental preservation (Mitchell & Carson, pp. 32–34). When respondents refuse to answer questions about the dollar value of a natural resource such as a forest, this is seen as a methodological problem in survey design, to be overcome by such techniques as asking people yes-no questions about whether or not they would accept a certain amount of money.

PVs also cause problems for negotiation. Governments often try to settle environmental (and other) disputes through negotiation among interested parties or their representatives. When parties to the disputes have protected values, negotiation gets stuck. The task of a mediator is to look for possible compromises, and some of these will involve the sacrifice of PVs. In this case, of course, false expressions of PVs are often used as negotiating ploys, but sometimes these expressions may stem from the same sources as they do in responding to surveys about values. These seem not to be a matter of mere posturing (Baron & Spranca, 1996, Experiment 1).

Notice that the issue here is not behavior. Surely, people who endorse PVs violate them in their behavior, but these violations do not imply that the values are irrelevant for social policy. People may want public decisions to be based on the values they hold on reflection, whatever they do in their behavior. When people learn that they have violated some value they hold, they may regret their action rather than revising the value.

In saying this, I express a concept of values (or utilities) as criteria for evaluating states of affairs. Values are reflectively endorsed. They are the result of thought, and are, in this sense, "constructed," in much the way that concepts are the result of reflection. Values are not simply desires, and very young children might properly be said to have no values at all, in the sense at issue. Values are important because they are our best judgments of the goodness of outcomes. Insofar as governments or other organizations

seek to produce good outcomes, the ultimate judgments of those outcomes are values.

If values are seen as constructed, like concepts, we can ask whether they are constructed well, just as we can ask whether concepts are formed well. It is possible that PVs result from the same kind of unreflectiveness that leads to overgenerlized concepts of the sort I have been discussing. People may agree with the claim that "all apples are red" without pausing to consider counterexamples. Likewise, they may endorse the statement that "no benefit is worth the sacrifice of a pristine rainforest" without thinking much about possible benefits (a cure for cancer or malaria?). Or, when people say that they would never trade off life for money, they may fail to think of extreme cases, such as crossing the street (hence risking loss of life from being hit by a car) to pick up a large check, or failing to increase the health-care budget enough to vaccinate every child or screen everyone for colon cancer.

Such unreflective overgeneralizations provide one possible avenue for challenging PVs in order to make compromise and tradeoffs possible. If PVs are unreflective in this way, then PVs should yield to simple challenges.

Baron and Leshner (2000) reported several experiments address the possibility that PVs are unreflective overgeneralizations. Subjects answered questions about whether they would regard certain outcomes, such as "electing a politician who has made racist comments," as so much against their values that no benefit would be sufficient to justify actions that caused such outcomes. Then when values were protected in this way, we challenged them by asking the subjects to think of counterexamples. PVs do sometimes respond to such challenges.

The fourth experiment also found that the effects of counterexamples can transfer to measures of omission bias (Ritov & Baron, 1990), the bias toward harm caused by omissions when that is pitted against harm caused by acts.

Another possibility is that people have not thought much about what happens when PVs conflict with each other. In Experiment 5, we asked people specifically about such conflicts, which people seem to find unproblematic. They are thus willing to trade off PVs when they conflict with other PVs.

The last two experiments ask what happens when harm (that goes against a PV) is probabilistic, and when it varies in amount. If (as we find), PVs are not honored when the probability and magnitude of harm

is low enough, this suggests a way in which we can measure tradeoffs. It may also suggest that PVs are, in a sense, unreflective. When people say that a value is absolute, they seem to have in mind a violation of a certain magnitude and probability.

In sum, PVs often yield to challenges. People who hold them can sometimes think of counterexamples. They will accept actions that violate PVs if the probability or amount of the harm is small relative to the probability and magnitude of benefit. Although people claim that amount of the violation of a PV doesn't matter (Baron & Spranca, 1997), PVs are less likely to be invoked if the amount of the violation is small or if it is improbable. This fact allows measurement of tradeoffs.

Our results suggest that PVs are strong opinions, weakly held. They are strong in the sense that they express infinite tradeoffs. Holders of these values assert that they are so important that they should not be traded off for anything. This assertion yields to a variety of challenges. After yielding, of course, the value may still be strong in the sense that a large amount of benefit is required to sacrifice the value.

The results are of greatest relevance to elicitation of values for public policy through the use of surveys. The results suggest that expressions of infinite tradeoffs need not be accepted at face value and that respondents will change their expressions on probing. It remains to be determined whether probing will *always* suffice to elicit usable responses. Additional probes, other than those used here, may be needed. For example, it may be helpful to focus respondents on consequences, independently of the actions that produced them. (They could, for example, imagine that the consequences were not intended, or caused by natural events.) The most general conclusion of the present studies is that we need to ask these questions before giving up on value elicitation for public policy.

The results may also be relevant to negotiation. Negotiation, however, differs from the present context in that exaggeration is almost expected at the outset, so that expressions of "nonnegotiable demands" may be even more subject to change than the PVs in our experiments. Still, some of these expressions may not be exaggeration. They may be initially honest expressions but they may still yield to further probing.

The morality-as-self-interest illusion

In an ordinary social dilemma (Dawes, 1980), each member of a group has two options. The cooperation option hurts the decision maker but helps the other members of the group. The defection option helps the decision maker but does not help the others. The total benefit of cooperation is greater than the loss to the cooperator, so it is best for all if each person cooperates.

In laboratory experiments as in real life, people often cooperate even though they must sacrifice their self-interest to do so. A large literature has explored the many reasons for such cooperation. These include altruism, reciprocity (the desire to cooperate when others are cooperating) and various illusions, such as the voter's illusion (Quattrone and Tversky, 1984). In that illusion, people behave as if they thought their behavior would influence others, even though they know only that they and others are subject to common influence.

At issue here is a second type of illusion that causes cooperation, the "illusion of morality as self-interest" (Baron, 1997). People seem to deny the existence of the conflict between self and others, the conflict that defines a social dilemma. Because morality and self-interest are usually correlated, just as length and number are correlated, people tend to overgeneralize and act as though the two are correlated even when they are not.

In a social dilemma, people try to reduce the apparent self-other conflict by convincing themselves that it doesn't exist. They may do this by telling themselves that "cooperation doesn't do any good anyway, so I do not need to sacrifice my self-interest." They may also do the opposite, and convince themselves that cooperation is in their self-interest after all. They may focus on the slight self-interested benefit that accrues to them indirectly from their own cooperation and ignore the fact that this benefit is less than the cost of cooperating. (If it were not less than the cost, then we would not have a social dilemma after all. The necessity for self-sacrifice for the good of others is a defining property of social dilemmas.)

In one study subjects read the following scenario (Baron, 1997): "Suppose you are an ocean fisherman. The kind of fish you catch is declining from over-fishing. There are 1,000 fishermen like you who catch it. The decline will slow down if the fishermen fish less. But, of course, you will lose money if you cut back. Nobody knows how much fish you catch. .... If nobody cuts back, then everyone can keep fishing at

roughly their current rate for 2 years and then they will have to stop. Every 100 people who cut back to 50% of their current catch will extend this time for about a year. Thus, if 100 people (out of 1000) cut back this much, fishing can continue for 3 years, and if 200 people cut back, it can continue for 4 years. (It is not expected that many more than this will cut back voluntarily.)"

Subjects were asked whether they would cut back and whether cutting back would "increase your income from fishing over the next few years?" Many of the subjects (29%) saw cooperation as helping them financially in the long run, although it clearly did not. The most common argument referred to the long term versus short term distinction. Examples were: "If I cut back now, I would not have to face the fear of running out of fish to sell and I could keep staying in business for a longer time and everyone else would also benefit." "Because if we cut back, there will be more fish for 4 years of 3 years so it would be a greater profit." The second response was typical of many that explicitly changed the question so that it was about the group rather than the individual.

The self-interest illusion can encourage cooperation, and this is a good thing when cooperation should be encouraged. However, it can also encourage cooperation that benefits one's group at the expense of outsiders. People who sacrifice on behalf of others like themselves may be more prone to the self-interest illusion, because they see the benefits as going to people who are like themselves in some salient way. They think, roughly, "My cooperation helps people who are X. I am X. Therefore it helps me." This kind of reasoning is easier to engage in when X represents a particular group than when it represents people in general.

Of course, this sort of reasoning contains a germ of truth. If we hold constant the gross cost of cooperation to the individual and the total benefit of cooperation to the group, then the net cost of cooperation (gross cost minus the cooperator's share of the group benefit) is smaller when the group is smaller. Also, in real life, people can influence each other more when the group is smaller. The illusion goes beyond this germ of truth, however.

The tendency of people to favor a group that includes them, at the expense of outsiders and even at the expense of their own self-interest, has been called parochialism (Schwartz-Shea & Simmons, 1991).

An experiment by Bornstein and Ben-Yossef (1994) shows a parochialism effect. Subjects came in

groups of 6 and were assigned at random to a red group and a green group, with 3 in each group. Each subject started with 5 Israeli Shekels (IS; about $2). If the subject contributed this endowment, each member of the subject's group would get 3 IS (including the subject). This amounts to a net loss of 2 for the subject but a total gain of 4 for the group. However, the contribution would also cause each member of the *other* group to *lose* 3 IS. Thus, taking both groups into account, the gains for one group matched the losses to the other, except that the contributor lost the 5 IS. The effect of this 5 IS loss was simply to move goods from the other group to the subject's group. Still the average rate of contribution was 55%, and this was substantially higher than the rate of contribution in control conditions in which the contribution did not affect the other group (27%). Of course, the control condition was a real social dilemma in which the net benefit of the contribution was truly positive.

It seems that subjects were willing to sacrifice more for the sake of winning a competition than for the sake of increasing the amount their own group received. Similar results have been found by others (Schwartz-Shea and Simmons, 1990, 1991 — I emphasize the Bornstein and Ben-Yossef study because it is the basis for my own studies). Notice that the parochialism effect is found despite the fact that an overall analysis of costs and benefits would point strongly toward the opposite result. Specifically, cooperation is truly beneficial, overall, in the one-group condition, and truly harmful in the two-group condition, because the contribution is lost and there is no net gain for others.

This kind of experiment might be a model for cases of real-world conflict, in which people sacrifice their own self-interest to help their group at the expense of some other group. We see this in international, ethnic, and religious conflict, when people even put their lives on the line for the sake of their group, and at the expense of another group. We also see it in strikes and in attempts to influence government policy in favor of one's own group at the expense of other groups. What is interesting about these cases is that we can look at the behavior from three points of view: the individual, the group, and everyone (the world). Political action in favor of one's group is beneficial for the group but (in these cases) both costly to the individual and to the world.

Parochialism underlies the concept of "rent seeking" (Krueger, 1974), the idea that groups organize to promote their group interests against the interests of others, in a game that would be zero sum except for

the effort expended in competition itself. “Public choice theory” and “rational choice theory” have incorporated the idea of rent seeking to explain the function of democratic governments through the idea that people pursue their rational self-interest (Brennan and Buchanan, 1985; Green and Shapiro, 1994). Often hidden in such explanations, however, is the assumption that people go beyond their self-interest in order to act on behalf of their rent-seeking group (as pointed out by Brennan and Lomasky, 1993). If rent-seeking is as widespread as it seems to be, then we must explain why people are so willing to sacrifice on behalf of rent-seeking groups, and apparently so much less willing to sacrifice on behalf of larger, more inclusive, groups. Although explanations abound, one of them may be that the self-interest illusion applies more to groups with which people identify, because people find it easier to confuse their own interest with that of others who are similar to them than with that of others who are less similar. If so, then dis-illisioning people about the self-interest illusion could reduce their desire to favor such groups, without substantially reducing their desire to cooperate for the good of all.

I did an experiment following the design of Bornstein and Ben-Yossef (1994) in comparing cooperation within a single group with cooperation within a group when that group’s gain is another group’s loss (the two-group condition) (Baron, 2000). The main addition is that subjects answer questions about their self-interest, in order to test the hypothesis that the self-interest illusion is greater in the two-group condition.

The experiment was done on the World Wide Web. Subjects ($N = 84$) were assigned to groups according to the time at which they began the study; one group at a time. Because each subjects had essentially no idea who the other subjects were, or even what country they lived in, the usual group identification that results from (for example) being students at the same university was absent.

The payoffs corresponded to small amounts of real money. Each subject participated in several games, and one game was picked at random for payment. The pay in question was added to the minimum payment and sent to the subject along with payment for other studies completed within the same month. Discussion of World Wide Web research in general is found in Birnbaum (2000).

One of the critical conditions, with two groups, began as follows:

> This game has two groups. Each group has three subjects. Each member of your group will

receive a bonus based on the number of your group members who contribute and on the number of the other group members who contriubte.

The endowment is $1.50 [or $2.00 or $2.50].

The bonuses will be distributed as follows:

Contributors in your group	Contributors in other group	Bonus for each in your group	Bonus for each in other group
3	0	$6	$0
2	0	$5	$1
1	0	$4	$2
0	0	$3	$3
3	1	$5	$1
2	1	$4	$2
1	1	$3	$3
0	1	$2	$4
3	2	$4	$2
2	2	$3	$3
1	2	$2	$4
0	2	$1	$5
3	3	$3	$3
2	3	$2	$4
1	3	$1	$5
0	3	$0	$6

This condition (and another two-group condition without the tabular presentation) were compared to one-group conditions in which the minimum payoff to contributors was $1 or $2, this approximating the payoff in the two-group condition with a moderate number of contributors in the other group.

All items ended with the following statement and questions (with additional spaces for comments, and a test question to insure attention):

Do you contribute your endowment now? (y=yes, n=no)

Is it your personal self-interest to contribute your endowment? (y=yes, n=no, d=don’t know or not sure)

Do you expect more money if you contribute your endowment than if you do not? (y=yes, n=no, d=don’t know or not sure)

Subjects did contribute more in the two-group condition than in the one-group condition (82% vs. 73%), replicating the parochialism effect. More importantly, the parochialism effect for contributing was highly correlated across subjects with the parochialism effects for Self-interest and More-money ($r = .75$ for each correlation, $p = .0000$). In other words, those subjects who showed a greater parochialism effect for contributing showed a greater self-interest illusion when the gain for their group was a loss for the other group.

However subjects understood the Self-interest question, their response to the More-money question was, on its face, an error in arithmetic. What happens when we try to correct this error by forcing subjects to do the arithmetic? In particular, would dis-illusioning subjects about the self-interest illusion make them less parochial? A second experiment explored this question, using a hypothetical scenario (with much larger amounts of money, and four-member groups) given to 67 subjects on the web. The questionnaire, called, "Investments," began:

> Imagine that you are a member of a group of 4 business partners. You and each of the other members of your group must decide whether to contribute money to a common pool, for a project that will yield benefits to everyone in your group, including you. This means that you get some of your contribution back, but never all of it. The rules are the same for everyone. The benefits depend on the number of contributors. Contributions and benefits are described in millions of dollars. (Remember, this is hypothetical!)
>
> In half of the cases, there are two groups, each with 4 partners. The two groups are competing for the benefits. The group with more contributors gets more of the benefits. The members of the two groups are similar, and it is just chance that the groups formed the way they did.
>
> You will make 24 choices. Before each choice, you will answer one or more test questions. After you indicate whether you think you would contribute or not, you will be asked a series of questions about your decision. Please consider the specific scenario in answering these questions.

Thus, there were three groups of eight items, each group containing four one-group and four two-group items. The first group was a pretest group, then the two others were interleaved, one "trained" — i.e., with

test questions that required calculation — and the other "untrained." Within each group of four, the required contribution (endowment) was $1 million, $3 million, $5 million, or $7 million.

The critical test question for the one-group condition in the training condition was: "What is the net effect of your contribution on your group, taking into account your contribution and the benefits to you and others?" In the two-group condition, a second training question was asked: "What is the net effect of your contribution on both groups, taking into account your contribution and all the gains and other losses in both groups, including yours?" The remaining questions were similar to those in the first experiment.

Only 15 of the 67 subjects showed a parochialism effect. However, these subjects also showed a parochialism effect for More-money, as in the first study. In sum, once again, some subjects do contribute more to their own group when its gain is the out-group's loss, and these subjects show a self-interest illusion.

Most importantly, the training (in the form of the test questions) reduced the self-interest illusion in the two-group condition, and training reduced the size of the parochialism effect. In particular, comparing respectively, the illusion (More-money) was 2.70 for trained and 2.87 for untrained, on a four-point scale, and the Would-contribute question responses were 2.89 and 2.73, respectively. The effect of training was specific to the two-group condition, so the training also reduced the parochialism effect. (Although these differences seem small, they were highly significant statistically. Their small size is in part the result of the absence of a parochialism effect in many subjects.)

The results on the whole suggest that one determinant of the parochialism effect is that the self-interest illusion is greater when an in-group is in competition with an out-group. The contrast between the two groups increases the perceived similarity between the decision maker and the other group members, thus increasing the tendency to think that "anything that helps the group helps me, because I am like them." The experiments here did not test this explanation directly. Another possible explanation — not inconsistent with the similarity explanation — is that the two-group cases evoke a schema or heuristic for competition. Such a heuristic or schema could be learned, as the result of experience with group competition, and it could be favored by species-specific predispositions that evolved over time.

Whatever the explanation, this effect is somewhat labile. As suggested by Singer (1982), it may be

possible, through reason, to understand the arbitrariness or group boundaries. The more that people think of boundaries as arbitrary, the more they can direct their non-selfish concern at the greater good rather than the parochial interests of their group.

We might think of actions as potentially affecting the self, the group, and the world. (The "group" and "world" may be defined differently in different situations.) In the situations of interest here, some action helps the group but hurts both the self and the world. Other actions might hurt the self and help the world, and still others might help the self only. One question for future research is what sorts of interventions might reduce parochialism without seriously harming altruistic behavior towards the world.

Concluding comments

The three sets of studies I have described are very different kinds of illustrations of how overgeneralized principles can affect public outcomes. The have several features in common, however.

One is that they all can be seen as resulting from a kind of overgeneralization, a misapplication of a principle that is usually reliable to cases where it causes trouble by producing non-optimal consequences. I have said that we have too few principles, but, in another sense, we might have too many. The effects I have described here would, for most part, not exist if people applied a single principle: try to achieve the best outcomes.

Second, most of these examples involve some commitment to the principles in question. The literature on cognitive biases often treats these biases as if they were inconsistent with reflective judgments. This is true of some biases, but not all. Methodologically, the difference is seen in the use of within-subject vs. between-subject designs. Some biases show up in between-subject designs only, in which two items given to different groups of subjects show inconsistent answers. When the same items are given to the same subjects, right next to each other, the inconsistency disappears. In other cases, the inconsistency remains (Frisch, 1993). The biases I have been discussing do not even lead to any obvious internal inconsistency (although, in some cases, it is possible to devise experiments in which such inconsistency is found). If they are biases at all it is mainly because they are inconsistent with an external, consequentialist, normative

standard.[1] They are found not only in laboratory experiments but also in the writings of philosophers and social critics.

I argue that these biases, as I call them, are still of interest, for two reasons. First, I believe that normative standard can be defended (Baron, 2000). So the answers are, in a sense, deficient. But, of course, not everyone is willing to slog through the somewhat ponderous defenses of consequentialist theory, and (thus, I think) not everyone accepts it as normative. In fact, it is downright unpopular.

The second reason for taking these results seriously is of greater interest even to these doubters. It is that we have some reason to think that the kinds of thinking I have described actually play a causal role in determining policy outcomes. Each policy outcome — such as intractability in negotiation, rent seeking, or neglect of those we might easily help — may occur more often, or to a greater degree, because of its support in people's intuition.

I cannot prove this, but the parallel between intuitions and outcomes is found in many cases, many more than I have described here. (See Baron, 1998, and the last chapter of Baron, 2000, for many other examples.) The general method here, by the way, is to move back and forth between observation of the outcomes themselves and the laboratory studies.

In sum, it seems likely that part of the reason for non-optimal outcomes is that they are supported by non-consequentialist judgments. If we are disturbed about the outcomes, we might consider changing the judgments that support them. This argument, of course, will appeal only to those who do not take these judgments as so necessary as to require the sacrifice of human welfare.

Another common theme running through these studies gives me reason for optimism. It is that the kind of judgments I have described are labile. They vary considerably from person to person, and from situation to situation. I have shown in each case how they may be affected by simple arguments. Of course, these arguments don't work for every person on every issue. Still, the effort to make judgments "better" from a consequentialist point of view might be a cost-effective way to improve the state of the world, given that many other methods have reached the limit of their effectiveness.

---

[1]The term "normative" in the psychology of judgments and decisions has a very different meaning from its use in many other social sciences. It refers to the standard by which we evaluate judgments and decisions.

References

Anderson, E. (1993). *Value in ethics and economics*. Cambridge, MA: Harvard University Press.

Asch, D., Baron, J., Hershey, J. C., Kunreuther, H., Meszaros, J., Ritov, I., & Spranca, M. (1994). Determinants of resistance to pertussis vaccination. *Medical Decision Making, 14*, 118–123.

Baron, J. (1973). Semantic components and conceptual development. *Cognition, 2*, 189–207.

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

Baron, J. (1997). The illusion of morality as self-interest: a reason to cooperate in social dilemmas. *Psychological Science, 8*, 330–335.

Baron, J. (1997). Biases in the quantitative measurement of values for public decisions. *Psychological Bulletin, 122*, 72–88.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. New York: Oxford University Press.

Baron, J. (2000). *Thinking and deciding* (3d ed.). New York: Cambridge University Press.

Baron, J. (2000). Confusion of group-interest and self-interest in parochial cooperation on behalf of a group.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values. *Journal of Experimental Psychology: Applied, 6*, 183–194.

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

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1–16.

Baron, J., Lawson, G., & Siegel, L. S. (1975). Effects of training and set size on children's judgments of number and length. *Developmental Psychology, 11*, 583–588.

Bazerman, M. H., Moore, D. A., & Gillespie, J. J. (1999). The human mind as a barrier to wiser environmental agreements. *American Behavioral Scientist, 42*, 1277–1300.

Birnbaum, M. H. (Ed.) (2000). *Psychological Experiments on the Internet*. New York: Academic Press.

Bornstein, G., & Ben-Yossef, M. (1994). Cooperation in intergroup and single-group social dilemmas. *Journal of Experimental Social Psychology, 30*, 52–67.

Brennan, G. & Buchanan, J. M. (1985). *The reason of rules: Constitutional political economy.* Cambridge: Cambridge University Press.

Brennan, G., & Lomasky, L. (1993). *Democracy and decision: The pure theory of electoral politics.* Cambridge: Cambridge University Press.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes, 54*, 399–429.

Green, D. P. & Shapiro, I. (1994). *Pathologies of rational choice theory: A critique of applications in political science.* New Haven: Yale University Press.

Hershey, J. C., & Johnson, E. (1990). How to choose on autoinsurance. *Philadelphia Inquirer*, June 24, 1990, 8A.

Lawson, G., Baron, J., & Siegel, L. S. (1974). The role of length and number cues in children's quantitative judgments. *Child Development, 45*, 731–736.

Meszaros, J. R., Asch, D. A., Baron, J., Hershey, J. C., Kunreuther, H., & Schwartz-Buzaglo, J. (1996). Cognitive processes and the decisions of some parents to forego pertussis vaccination for their children. *Journal of Clinical Epidemiology, 49*, 697–703.

Mitchell, R. C., & Carson, R. T. (1989). *Using surveys to value public goods: The contingent valuation method.* Washington: Resources for the Future.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and the voter's illusion. *Journal of Personality and Social Psychology, 46*, 237–248.

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

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79*, 79–94.

Schwartz-Shea, P., & Simmons, R. T. (1990). The layered prisoners' dilemma: ingroup vs.

macro-efficiency. *Public Choice, 65*, 61–83.

Schwartz-Shea, P., & Simmons, R. T. (1991). Egoism, parochialism, and universalism. *Rationality and Society, 3*, 106–132.

Singer, P. (1982). *The expanding circle: Ethics and sociobiology*. New York: Farrar, Strauss & Giroux.

Singer, P. (1993). *Practical ethics* (2nd ed.). Cambridge: Cambridge University Press.

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

Tetlock, P. E., Lerner, J. & Peterson, R. (1996). Revising the value pluralism model: Incorporating social content and context postulates. In C. Seligman, J. Olson, & M. Zanna (Eds.), *the psychology of values: The Ontario symposium, Volume 8*. Hillsdale, NJ: Erlbaum.

Wertheimer, M. (1959). *Productive thinking* (rev. ed.). New York: Harper & Row (Original work published 1945)