NOÛS

NOÛS 00:0 (2017) 1–26
doi: 10.1111/nous.12227

The Asian Disease Problem and the Ethical Implications Of Prospect Theory

Sandra Dreisbach
UC Santa Cruz

Daniel Guevara
UC Santa Cruz

Abstract

We discuss the bearing of Daniel Kahneman and Amos Tversky's Prospect Theory on some central issues in ethics. It has been argued that the theory provides a better explanation of our intuitive responses to some important ethical decision cases—like some famous cases put by Philippa Foot and others—than traditional and widely acknowledged ethical principles do. In this way, Prospect Theory contributes to the new wave of skepticism, emanating from the social sciences, about the role of intuitive judgments in ethical theory and philosophy more generally. We focus on Kahneman and Tversky's famous Asian Disease Problem. We show that the case fails to support Prospect Theory over traditional ethical theory as an explanation of the most common intuitive responses to the case, and, moreover, fails as an account of the most common intuitive responses to Foot's famous trolley case and related cases. We maintain that careful critical attention to all these cases shows that Prospect Theory has not made a successful incursion into ethics, whatever it may have established about non-ethical decision-making.

§1. Prospect Theory is a remarkably novel and ambitious theory purporting to explain why we choose the way we do in situations of the kind commonly studied in the social sciences and philosophy. It became famous as a compelling alternative to Expected Utility Theory, when Daniel Kahneman won the Nobel Prize in economics for developing it with Amos Tversky (who had passed). Before long, it was also thought to challenge traditional ethical theories. For example, in one of the first and still most important applications of Prospect Theory to ethics,[1] Tamara Horowitz argues that the theory provides a better explanation of our intuitive responses to some well-known ethical decision cases—including the now wildly famous Trolley cases put by Philippa Foot and others[2]—than any of the celebrated, traditional ethical principles do, in particular the Principle of Doing and Allowing.[3] As we will see, Prospect Theory thereby contributes to the new wave of skepticism[4] emanating from the social sciences about the role of intuitive judgments in ethical theory and in philosophy more generally. It calls into question the authority and relevance of those judgments, and thus also the authority and relevance of the most widely discussed principles and distinctions which philosophers have developed to account for them.

We focus our attention on Kahneman and Tversky's Asian Disease Problem,[5] by far the most important case adduced in their theory's favor against traditional accounts of our ethical intuitions as flowing from reasoned ethical principles or distinctions. What surveyed responses to the case seem to demonstrate—rather decisively in Kahneman's influential opinion—is that our ethical intuitions are hopelessly inconsistent or arbitrary, and thus not guided by reasoned principles or values, in the case. What's more, the same arbitrariness or inconsistency seems to occur in many other cases of a similar structure, as has been amply confirmed in non-ethical domains, such as the domain of money (which is in part what lead to the Nobel Prize). The contention is that Prospect Theory provides the best account of all the relevant intuitions, and certainly a better account of the ethical ones than traditional ethical principles do, for the simple reason that ethical principles cannot be made to fit entirely arbitrary or inconsistent intuitions.

Our main purpose is to show that, on the contrary, the most common intuitive responses to the Asian Disease Problem fit nicely with the idea that they were influenced by consistent, principled, and fairly standard non-consequentialist ethical considerations, including considerations in the neighborhood of Doing and Allowing. Moreover, such considerations actually account for the common intuitions in the Asian Disease case better than Prospect Theory does, especially in light of the equally common intuitive responses to the earlier cases introduced by Philippa Foot and others.

The reason that we focus on the Asian Disease Problem is that no other ethical decision problem adduced by prospect theorists enjoys anything close to the same status as that one does. The empirical results that Kahneman and Tversky obtained from the case have been widely confirmed by them, and many other researchers, for more than three decades and on an impressive variety of respondents. So we do not dispute the results. We only dispute the implications being drawn from them about our ethical judgments. For, although the great majority of discussions of the Asian Disease Problem have been concerned with its implications for *non-ethical* domains, it is now also cited by distinguished moral philosophers as compelling evidence for Prospect Theory's skeptical account of our moral intuitions and the principles traditionally thought to explain and influence them.[6]

There have been a few other celebrated cases—such as the Snow Shovel Case and Schelling's Tax Case—adduced in favor of Prospect Theory's skeptical account of our ethical judgments and principles. But we believe that it has already been shown, mainly by F.M. Kamm, that these other cases cannot be used as evidence for Prospect Theory either. Kamm's results have been unduly neglected, so we summarize them and relate them to our own.[7] We conclude that Prospect Theory has not been successful in accounting for ethical decision-making, whatever it may have shown about non-ethical decision-making where, for example, only a modest sum of money, or the like, is at stake (rather than innocent human lives, or basic rights, say). The Richer Decision problem, discussed in a moment, exemplifies what we are calling the non-ethical domain. As we will see, this problem raises the question of whether to gamble a little windfall or not. Now, anything can be an ethical question in the right context, even the question of whether to gamble a

modest sum (an innocent person's life might hang on it somehow, for example). But we take it for granted that the contrast between such a case and the Asian Disease case (where the explicit concern *is* innocent human life) exemplifies the contrast we draw when we observe that the great majority of cases in support of Prospect Theory fall within non-ethical, economic domains, like the domain of money. We do not dispute the theory's claims in these non-ethical domains.

One more remark by way of introduction. Some may wonder whether we are from the start letting a familiar mistake slip in to our discussion. Prospect Theory is taken by its proponents to be an empirical-psychological theory about how our decisions are *caused*, whereas ethical theories are normative theories about how our decisions *ought to* go, as opposed to how they do *in fact* go. It might seem like mixing apples and oranges for us to try to compare the one type of theory with the other. We address this concern later, in §4, when we are in a better position to appreciate the full skeptical impact of Prospect Theory, in particular, the doubt it casts on the common assumption that our judgments are being influenced and guided by moral values and principles in the cases under scrutiny.

§2. In this section we describe the Asian Disease Problem and Prospect Theory's use of it. We deploy our argument against Prospect Theory's use of it, in the next section. That argument will defend standard *non*-consequentialist principles or values, in particular, Doing and Allowing (or something close to it) against Prospect Theory's attempt to debunk and supplant it. But we should emphasize that it is *not* our purpose, nor is it necessary for us, to adjudicate the arguments for or against non-consequentialism generally: i.e., for or against the view that there are moral constraints on what we are allowed to do when promoting even the best consequences. Rather, our concern is to show that whatever the final truth is about non-consequentialism, there is a perfectly reasonable application of it in accounting for the most common responses to the Asian Disease problem.

We understand that this leaves open the question of whether our own account will hold up under further empirical testing. As noted, we are in the middle of a large and growing wave of empirical work calling into question the reliability and authority of our intuitive ethical responses, and the traditional ethical accounts of what might be motivating them, in these and related famous cases. Aside from the phenomena we're about to discuss in this paper, there are ordering effects, and many other empirical effects to cope with. Not to mention the older and larger wave of *conceptual* challenges to cope with, generated by so-called Trolley problems, and the like, that seem capable of causing intractable difficulties for any ethical sensibility or theory that tries to account for our intuitive ethical judgments across the various cases. We will have something to say about all this eventually, after we have shown that our account holds up just fine in the Asian Disease case.

We turn now to that case.

Kahneman and Tversky presented a group of subjects (Group 1) with the following scenario and prompt, in a scientifically conducted survey:[8]

Asian Disease Problem

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
>
> Which of the two Programs would you favor?

For a second group of subjects (Group 2), instead of Programs A and B, the following alternative Programs C and D were given (all else the same):

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

The abundantly confirmed results of this survey indicate that the difference in wording ("will be saved," "will die") induces a large majority of Group 1 to select Program A, and a large majority of Group 2 to select Program D. All the programs have the same expected utility, so there must be something about the wording distinct from expected utility that guides and explains the choices. Prospect Theory has become famous in part because it seems to many to have the best (or even *only*) explanation of why the subjects choose the way they do in this decision case and similarly structured cases. The theory explains the choices as the effects of *framing*, i.e. the effects of describing essentially equivalent decisions in *in*significantly different ways. In order to see how this works, we first need to understand how Prospect Theory explains what happens when the framing effects occur. Specifically, we need to understand its explanation of how framing induces one group to pick Program A over Program B, and the other group to pick D over C. Kahneman puts it this way in his succinct, recent summary of the theory: "Decision makers tend to prefer the sure thing over the gamble ... when the outcomes are good [lives saved]. They tend to reject the sure thing and accept the gamble when both outcomes are negative [lives lost]."[9] (What he means, of course, is when the outcomes are *framed* as good or negative.)

Now a tendency to prefer a sure gain when its expected utility is equal to that of a gamble (or even when it is *less than* the expected utility of a gamble) was already well established in the literature that studied similar decision problems in the domain of money. Kahneman and Tversky's Asian Disease Problem was one of the first to show that this tendency also seems to exist in these more or less explicitly ethical contexts (life and death), and, more to the point, in a way that appears to be described and predicted best by their theory, rather than by friendly adjustments to Expected Utility Theory. We are not concerned in this paper with possible responses in defense of Expected Utility Theory (nor, again, with Prospects Theory's claims

in non-ethical domains), but only with the implications of the theory for traditional ethical theories, more particularly, non-consequentialist theories.

According to Prospect Theory, the subjects in Group 1 are affected by the wording "will be saved" to think of the people threatened by the disease as lost, i.e. as dead or as good as dead. By contrast, the subjects in Group 2 think of them as alive and well, because the decision problem is framed with the wording, "will die." Then, the differences in wording dispose the subjects in the two groups to draw different *baselines*, or *neutral reference points*, according to which outcomes are assessed as gains (e.g., in Group 1, lives saved from death) or losses (deaths of the alive and well, in Group 2). It is this introduction of a neutral reference point, or baseline, that's crucial to distinguishing Prospect Theory from Expected Utility Theory, by the way.

Many examples from the much-studied domain of money illustrate and confirm this sort of framing effect in cases with a structure like the Asian Disease case. Consider, for example: the Richer Decision Problem (as we call it).[10]

Richer Decision Problem

Group I
Participants are told they are given $300. Then asked to consider the following two options:

Option One: Gain $100

Option Two: 50% Gain Nothing, 50% Gain $200

Group II
Participants are told they are given $500. Then asked to consider the following two options:

Option One: Lose $100

Option Two: 50% Lose Nothing, 50% Lose $200

There is a lot of evidence that most people in Group I will tend to choose the first option in this decision problem, most in Group II the second option. As we've noted, the Asian Disease Problem has a similar structure. And in both cases, where there are only gains to be considered relative to the baseline, people tend to pick the sure thing. Where losses are considered relative to the baseline, they tend to gamble in order to try to avoid loss. And this is just as Prospect Theory maintains and predicts: for, according to the theory, we are more averse to losses than we are interested in gains, and we will therefore tend to embrace risk in order to avoid the losses, but will tend to be risk averse in securing gains. And, again, losses and gains are always assessed relative to a neutral reference point or baseline. In the Richer Decision Problem the initial $300 or $500 up front disposes respondents to set the baseline at $300 and $500 respectively, and gains and losses accordingly from there. In the Asian case, the theory holds that the wording "will be saved" disposes the respondents in Group 1 to draw the baseline at 600 dead, or as good

as dead. Then Group 1 tends to go for the sure thing, because it sees itself as considering gains only. Group 2 is induced by the "will die" wording to draw the baseline at 600 alive and well. So, Group 2 tends to gamble because it sees itself as considering only losses and no gains. Moreover, and most important, it seems that the ethical principles that might be thought to inform and justify the intuitive responses could get no authority from the intuitive responses themselves, given their pedigree: namely a psychological mechanism that selects sure things, or else gambles, in accordance with gains or losses as determined by the baseline, which itself is determined by what appears to be an arbitrary effect of framing. Ethical principles would seem at best to be post hoc rationalizations of judgments that are driven by non-ethical psychological mechanisms, and manipulated by ethically arbitrary variations of framing.

§3. We begin our argument against this skeptical-psychological account of our choices in the Asian Disease Problem by assuming that much of the thinking that leads to the selections of the programs is *un*conscious, or at least implicit and not entirely conscious. Here we follow Prospect Theory, for it does not claim or suppose that the framing effects, the setting of baselines, or the aversions and attractions to risk, are entirely *conscious* mental phenomena. On the contrary, the phenomena are largely taken to be *un*conscious. As Kahneman reports, evidence for this lies in the fact that when subjects are confronted with the apparent inconsistency of their choices in the Asian Disease Problem, they usually respond with puzzlement and embarrassed silence.[11]

It's interesting, by the way, that despite the appearance of an inconsistency respondents nevertheless wish to stick with their choices. Though dumbfounded, it seems that they sense that there must be a reasonable explanation. We believe one advantage of our own account is that it vindicates their sense by articulating a reasonable and consistent narrative for their choices. However, the immediate point is that *whatever* is influencing the subjects in Groups 1 & 2 to make the choices they do, it is mostly unconscious or implicit, and not easy to articulate or make conscious.

Furthermore, we accept that the choices are affected by the changes in wording, and to that extent, we accept the so-called framing effects, which are in any case empirically well confirmed. However, it should become clear that we do not think of them as arbitrary or insignificant changes or effects; in fact the claim that they are is a fundamental error of the prospect theoretic account, in our view.

We also assume for the sake of the argument Prospect Theory's account of how we draw baselines in the Asian case. But in §5, below, we retract some of what we have assumed here, for reasons that should also become clear.

So much for general remarks. We turn now to the first part of our main argument.

The first thing we wish to show is that, contrary to what Kahneman and others claim,[12] there is nothing about the pattern of selections that rules out, or even speaks against, the influence of consistent and fairly reasonable ethical considerations, in particular, non-consequentialist ones. So ethicists who take it as a working assumption that some such considerations are what motivate and explain the

selections have no reason to retract that assumption when faced with the results of the Asian Disease Problem. On the contrary, much speaks against the prospect theoretic account, as we will show eventually.

We will concentrate first on the *second* group's (Group 2's) selections: the choice of D over C. The result is consistent with the assumption that basic non-consequentialist ethical considerations, including one closely related to Doing and Allowing, are what guide and explain the selection.

The exact scientific estimate is that if Program C is implemented, then 400 people—all presumably innocent—will die. There is, in general, a strong ethical presumption against implementing a program that ensures that alive and well innocent people will die. Of course, non-consequentialists share this presumption, but also a further one against the use of harmful *means*; they would not be concerned just with the number of alive and dead at the end of the day (or with the expected utilities of D's outcomes). They would tend to be concerned about any thought that the alive and well innocents will die because of something lethal about the intervention itself. We can imagine various harmful means that the "will die" language might (at least implicitly) suggest: to fix ideas imagine that C is an experimental vaccine whose side-effects will immediately bring about the death of most, even while it works as hoped for on the rest. It may be unusual for a program of vaccinations to have such a side-effect, but the disease is said to be unusual, and so might require unusual measures. A non-consequentialist cannot take the attitude that everyone was going to die anyway.

Moreover, the subjects are presented with an alternative in Program D, which only puts the alive and well innocents at *risk* of death and thus allows chance to decide their fate, with a respectable chance of none dying. With nothing more to go on, it would be reasonable for someone with non-consequentialist sensibilities to opt for D rather than C, on the basis of something in the neighborhood of Doing and Allowing: Allow chance to decide (with a respectable chance of no one dying), rather than implement a program that an exact scientific estimate says will ensure that most alive and well die—means unspecified.

Before we take up objections to this account, let us emphasize that we do *not* maintain that the respondents go through a line of thinking that *explicitly* considers such things. Nor are we flatly asserting that such thinking is in fact what *implicitly* motivates the responses; we are only defending the working assumption that it does. Nor (much less) are we asserting that the thinking we narrate is impeccable or unassailable. We maintain only that it is reasonable and plausibly non-consequentialist, and that the empirical results of the Asian Disease Problem do not conflict so far in any way with it.

Then, let us take up objections to the account so far, before moving on to the choice of A over B.

It might be observed that there are non-consequentialist reasons for *avoiding* D too. For, just as there is a strong non-consequentialist presumption against doing something that ensures the deaths of innocents, through possibly harmful means, there is also a strong presumption against putting everyone at serious risk of death, through possibly harmful means. Both presumptions are related

to respect for the intrinsic value of a person's life or well-being—a crucial and standard non-consequentialist constraint on anything we do. However, our own non-consequentialist sense is that the presumption against doing what ensures the deaths of alive and well innocents prevails, since doing that (through possibly harmful means) is worse than doing what, as in D, has a 1/3 chance of bringing it about that *no one* dies (through possibly harmful means, or otherwise). We understand that D also involves a 2/3 chance that *all* 600 will die, through possibly harmful means: let us keep the cases similar and imagine, for example, an all-or-nothing, alternative experimental vaccine which has a 1/3 chance of working right, and bringing it about that no one dies, or else a 2/3 chance of bringing it about that everybody dies immediately through side effects. Even so, the alternative is to actively take measures that ensure the death of 400, through possibly harmful means. Might the numbers, i.e. the extra 200 who likely die in D, count in favor of implementing C instead? For example, might it be that C is a better way of showing respect for the rights or the intrinsic value of persons, than introducing a 2/3 chance that 600 die, with corresponding 1/3 chance that no one dies?

What would be the non-consequentialist reasoning in that? Is it that 200 for sure do not die?

Well, this is the only positive feature C has: namely that it ensures that 200 of 600 do *not* die (or, less positively, at least it does *not* ensure that they *will* die, since nothing is said about them one way or the other). But C does this only by also ensuring that most of the 600 *do* die. Perhaps there are non-consequentialist reasons to prefer that to a real possibility of no one dying. We doubt it, but do not wish to assert flatly that there are no such reasons. And even if there are, the point is that we have articulated plausible and familiar non-consequentialist considerations that favor the implementation of D over C. A respondent sensitive to such considerations will tend to think it better to let chance decide, all things equal. The reasoning is not unassailable, but we are dealing with respondents who are ordinary people, not sophisticated philosophers attuned to all the complicated casuistry around such cases.

Someone might point out that if the deadly results of C's implementation are random, then the same chances obtain as in D. Under either program there is a 2/3 chance that any given person will die, a 1/3 she won't. Our reply is that, even so, C would *not* in that case show everyone the same sort of equal respect as D. For the randomness does not change the fact that C's implementation ensures the death of most of the innocents. So, while it is true that chance would decide *who* dies in particular in both C and D (with each alive and well innocent person standing under a 2/3 risk of death in both), only D holds out at the same time a real possibility of no one dying. Randomly or not, 400 otherwise alive and well innocent people will die for sure under C (again, means unspecified), when D offers a real possibility that none die.

Finally, there might be an objection to our construal of the "will die" language as unconsciously or implicitly suggesting possible harm in the implementation of the program. At the risk of being dogmatic, we see no difficulty with this construal, and with our abandoning the prospect theoretic claim that "will die" and "will

be saved" are morally *in*significant variations in wording. The claim that they are morally insignificant strikes us as implausible on its face; moreover, in §5 we will demonstrate advantages to our construal.

We conclude that a fairly reasonable, non-consequentialist line of thinking, in the neighborhood of Doing and Allowing, favors D over C. Moreover, nothing about the results of the Asian Disease case counts against the idea that such a line of thinking is what guides and explains the selection.

Later (§5), we will raise further objections to Prospect Theory's claims about the framing effects, in particular about the claim that the framing induces us to set baselines like "as good as dead" or "alive and well." But we agree that there is something about the "will die" language that induces respondents to select D rather than C, and something about the "will be saved" language that induces them to change their selections by selecting C's analogue rather than D's (the sure thing over the risk). This much we accept as a firm result of the Asian Disease Problem. But as just noted, we don't assume or maintain that the difference in language is arbitrary or insignificant. What's distinctive of, and essential to, our own account of the selections of Group 2 is our idea that the "will die" language induces the respondents to think (implicitly or unconsciously) in terms of the non-consequentialist ethics of *harm*, and thus in terms of steering clear of lethal harm as much as possible—where it's understood that harm can come from the means to an end. The description of the case leaves it entirely open how the interventions against the disease bring about their outcomes, and harmful or unjust means are easy to imagine, even if never consciously brought to mind. This induces respondents to favor Program D, which holds out a respectable chance that there will be no lethal harm. Or, as always and more carefully, nothing counts against this interpretation of the data.

Let us turn, then, to our non-consequentialist account of the selections of Group 1. The challenge here is to remain reasonable and consistent, in light of our account of Group 2's selections. This is commonly thought to be an impossible challenge because it is thought that the options put to each group are essentially equivalent, with only inconsequential variations in how the choices are framed.

We take it back later, but for now we continue to assume with Prospect Theory that because of the "will be saved" language, the neutral reference point, or baseline, is 600 dead, or as good as dead. So the respondents make their decisions in the frame of mind that thinks of the harm as already having been done, or as being as good as done. Here is where we find an opportunity to continue our narrative in a consistent and principled way. It seems reasonable to suppose that the "will be saved" language induces respondents to think of Programs A and B as involving the distribution of *benefits*, rather than potential harms. If so, the non-consequentialist reasoning that we have proposed for Group 2's selections is no longer plausibly applicable. There is no longer any suggestion of harm. More specifically, the reasoning that leads to the rejection of C (A's analog) above is *certainly* not applicable since, given the effects of the change in wording, there is no longer any question of doing something that ensures the death of 400 (or any other number of) alive and well innocents. Rather, what's ensured by A is the *saving* of 200 all-but-dead innocents, with no ethical

downside in not ensuring the same for any of the remaining 400, since no one is presented with a way of doing *that*. So far as this goes, then, A seems to be a fine selection on our narrative. But, of course, the question for our account is whether A is better than B on a non-consequentialist view. It seems so. B has the obvious downside, on any view, that we gamble with a great benefit (the odds significantly against us) rather than use our power to bestow it for sure. This is a serious and, more to the point, *ethical* downside because of the nature of the benefit (life or salvation from death).

So the differential responses to the change in wording may be reasonable, if the "will save" language suggests possible benefits, rather than harms, and the "will die" language suggests possible harms.

It might be objected that, in terms of the framing effects—and thus according to losses and gains relative to the baseline—both programs, A and B, involve only gains or, at least, no losses/no gains. So *not* bestowing the benefit is *no* loss or downside, in those terms. There is therefore no downside in losing the gamble, given the neutrality of the baseline. However, the response ignores the obvious ethical absurdity in the idea that the no gain/no loss outcome is a *neutral* outcome.[13] This is clear when we consider the option of doing nothing at all, not implementing either program. Obviously, it would be terrible to do nothing. So this counts against the alternative, i.e., B over A, since B makes it pretty likely that we'd get the terrible "neutral" outcome that results from doing nothing, squandering the opportunity to secure a great benefit for some.

But what of those in A who, for all we know, lose out on the benefit? Well, as noted, there is no way of making sure the benefit extends to them, neither A nor B offers us that pleasant outcome.

Still, B at least involves the *possibility* of the benefit reaching the rest too. Could extending that possibility to *all* be a kind of egalitarianism, a proper way of showing equal respect for them?

As before, we do not wish to maintain that some such thinking is clearly wrong, nor to deny that it is non-consequentialist. Considerations of equal respect are paradigmatically non-consequentialist. Furthermore, we have emphasized that we do not see ourselves as proposing unassailable reasoning in our narrative—including reasoning that would rule out of hand non-consequentialist narratives pulling the other way. It's understandable for anyone to be attracted to the 1/3 chance of saving all from the jaws of death. All are presumed to be equally worthy of the benefit, and it is painful to opt for A over B, when (for all we know) that will result in most *not* receiving the benefit. Perhaps the minority selections of the subjects in Group I (i.e. of B over A) could be given a non-consequentialist explanation along some such lines (which would suit us fine, since it would show that there are reasonable ethical explanations for *all* of the selections). Still, it is a questionable egalitarianism that opts for B over A, since B makes it so likely that the benefit will fail to be bestowed at all. Also, if we assume that Program A can distribute the benefit to the 200 randomly, it could also extend to all 600 an equal 1/3 chance of receiving the benefit. In this way there would be no unfairness or inequity in the treatment of the other 400. Either way, it smacks of spiteful self-centeredness for

them to complain (from the grave, as it were) that we did not try to save all, given that probably nobody would have been saved if we'd tried that.

In any case, on our account, the so-called framing effects are neither arbitrary nor inconsistent, rather they change the ethical stakes and priorities as we move from one group of respondents to the other. The "will die" language makes Group 2 unconsciously or implicitly think of possible harm in the means adopted by the selected programs; "the will be saved" language does not, but induces Group 1 to think instead in terms of the distribution of possible benefits. Then, Group 1 thinks in terms of the non-consequentialist *ethics of benefits*, Group 2 in terms of the non-consequentialist *ethics of harm*, where it's understood that harm can come from the means adopted, independently of any harm or badness in the outcome. Moreover, despite the differences in the ethical stakes and priorities, the basic distinction in Doing and Allowing still helps make sense of the subjects' selections throughout. This can be seen through a common illustration of the distinction. The illustration involves contrasting two cases, cases we'll return to in §5.

Case (1): Suppose one innocent is in danger of drowning in spot A, five in danger of drowning in spot B. You can make it in time to rescue the one at A, or the five at B, but not both. This case is commonly contrasted with the following alternative: Case (2): Suppose five innocents are stuck on a hillside that is about to collapse and kill them, while one other is stuck in the middle of the only road to the 5. You can make it in time to save the five, but only if you run roughshod over the one, killing him in the process. If you save him instead, the others will die in the mean time.

If you think that heading out for the five at B is OK, but that running roughshod over the one stuck in the road is not, then you seem to accept that there are harms (the death of an innocent) we are allowed to let happen, but that we cannot actively bring about, notwithstanding the same number harmed or unharmed at end of the day. The Doing and Allowing distinction is of course controversial, questioned by consequentialists long before the sort of criticism prospect theorists raised against it. And even among non-consequentialists its application is often difficult and controversial, especially in matters of life and death. But the cases contain the non-consequentialist heart of the matter for our purposes: if in one scenario you are just concerned with saving people, you of course ought to save as many as you can. If in another scenario you're concerned about actively introducing some great harm in the process, then it's *not* OK to simply save as many as you can save in the first scenario. There are ethical side constraints on you to avoid doing the harm, whatever the overall benefits. Perhaps the side constraints are not absolute, but they matter a lot.

For all the complications introduced by the probabilities, and other dis-analogies with the Asian Disease case, this sensibility could be what guides the selections in the case, as we have narrated.

We understand that our argument has not shown that the selections are in fact a result of the ethical thinking we've narrated, and perhaps we have strained the patience of the reader who thinks we should simply cut to the chase, and submit our speculations to further empirical tests. We will say more about that later, but

perhaps the following provides some indication of why we show so little interest in turning immediately to such tests: If we are correct, then it's a fact that one of the most studied, well-confirmed empirical tests, constructed by two of the best in the business, and widely accepted as formidable if not decisive evidence for the hopelessness of our ethical intuitions in the case (and similar cases), turns out to be entirely consistent with a relatively simple and reasonable ethical account of them. As we will see in the next section, the same is true of other putatively formidable and widely discussed surveys intended to debunk our ethical intuitions. Our view is that, rather than adding to the proliferation of surveys (most of which will not stand the test of time), our efforts ought to be directed for a while at careful reflection on the best ones, and in particular those which have otherwise well-confirmed, powerful and elegant theories that seem capable of accounting for the results, like Prospect Theory. This is what we have been trying to do here.

Before wrapping this section up, a remark about simplicity. Our presentation has been complicated by the concern to anticipate objections from a professional philosophical audience. But we think that, once those objections are answered and set aside, our narrative is evidently relatively simple, and thus attributable to the average person, who cannot be expected to be attuned to the complications raised by professionals. Anyone familiar with cases and principles in this area knows how difficult the casuistry can get, and lives in fear (or glee) in anticipation of new cases that make trouble for initially promising ethical accounts of our intuitive judgments and guiding principles. All we mean to indicate is that our narrative is a perfectly reasonable basis for the selections, especially if it reflects the intuitive ethical thinking of ordinary people with probably no special training in or talent for the casuistry.

In stark contrast to our narrative, Prospect Theory hypothesizes that the selections are in fact the result of a psychological law[14] that causes us to make the selections according to the psychological mechanisms described by the theory. So that, as a matter of psychological fact, in the Asian Disease case, as in many others with the same basic structure, we draw a neutral reference point susceptible to the arbitrary effects of framing, and then react to gains and losses from there, as the theory describes and predicts we will, i.e. with a much greater aversion to loss than an inclination to gain or to breaking even. That's what, according to the theory, is at work behind the arbitrary and incoherent selections in the Asian Disease case, and many other cases of a similar structure.

But of course we dispute this, and in particular that Prospect Theory has discovered a hypothesis or law that can be generalized into the domain of ethics. *Maybe* the similarly structured *non*-ethical cases (concerning the loss or gain of some money) provide *some* basis for thinking that the theory will hold in ethical domains as well. But it must be tested in those significantly different domains (where basic rights, and life and death, are at issue) before we can have any real confidence that it does hold, as of course Kahneman and Tversky understood in constructing the Asian Disease case. We have yet to deploy our argument showing why we think the theory fails to hold in the ethical domain—at the moment the theory is compatible with the results of the Asian Disease case too. The point so

far is only that the case does *not* debunk our ethical intuitions, or traditional ways of accounting for them, contrary to prestigious and influential opinion.

This completes the first part of our main argument. What remains is to show that our narrative holds up much better than the prospect theory narrative, when we try to account for a broader range of cases all together, including rescue scenario cases like (1) and (2) above. As we press on to those and other famous cases (§5), we will see that Prospect Theory introduces serious complications with its crucial idea of a baseline—complications which show that the traditional accounts are to be preferred to the prospect theoretic one.

Kahneman has adduced other cases in support of Prospect Theory's skeptical account of our intuitive ethical judgments; we will be considering these in the next section (§4). But we conclude this section by emphasizing the special importance of the Asian Disease Problem. The so-called framing effects in it are amazingly robust and tenacious. It's worth quoting Kahneman's recent summary of the implications he draws from the case (and that we have been working to refute in this section):

> The failure of invariance [i.e. of consistency] is both pervasive and robust. It is as common among sophisticated respondents as among naive ones, and it is not eliminated even when the same respondents answer both questions within a few minutes. Respondents confronted with their conflicting answers are typically puzzled. Even after rereading the problems, they still wish to be risk averse in the "lives saved" version; they wish to be risk seeking in the "lives lost" version; and they also wish to obey invariance and give consistent answers in the two versions. In their stubborn appeal, framing effects resemble perceptual illusions more than computational errors.[15]

The respondents want it all: they want to be able to stick to their choices even when they realize how they have been affected by the changes in wording, and yet they also want to think of their choices as consistent. They seem to have sensed that there must be a reasonable explanation of their selections, even if they could not articulate it. We have an explanation that vindicates that sense.

The social scientific literature presents us with many other cases of framing—and, more generally, of supposed evidence for the unreliability of intuitions—that deserve attention. But we submit that the Asian Disease Problem can no longer be cited as evidence for arbitrariness and inconsistency in our intuitive judgments, nor as evidence against the working assumption that they are influenced by reasonable ethical considerations, at least considerations of a fairly standard non-consequentialist type.

§4. There have been other critiques of Prospect Theory, which, like our own, object to it on conceptual or ethical grounds (as opposed to empirical).

Mark van Roojen (1999) has argued that Prospect Theory has not established itself as an alternative to Doing and Allowing, or like ethical principles. But his argument is different from ours. He argues that the theory has done nothing to show that Doing and Allowing, and related principles, cannot be *justified* on the basis of the judgments we make in test cases. His point is that Doing and Allowing

is supposed to explain the *truth* (if it be such) of our judgments in the cases. In rescue cases such as those we will discuss in §5, for example, the point of ethical principles and theory is *not* to explain why we *think* some can be left to die and the others saved. The prospect theoretic explanation of why we *think* it, even if true, is not in conflict with the proper purpose of ethical reflection on cases and the principles meant to show which judgments are correct, and why. Reflective equilibrium between our judgment in cases and the principles supporting those judgments is obtained not by attending to *why* people think the way they do, but rather by attending to the *content* of the judgments and principles themselves. Prospect theory can debunk ethical theory only if it can show that the reasoning of the respondents in the Asian Disease and other test cases is fallacious (van Roojen 1999, p. 846–847).

However, we have been working with the common assumption—and believe it is important to do so—that our thinking and judgment, besides not being fallacious, *is influenced* by ethical principles that can be arrived at upon reflection on these cases. Or, more cautiously, that at least Prospect Theory has not shown anything to the contrary. van Roojen seems unconcerned with this issue. But we think he ought to be, because, like many other ethicists, we think that ethical theory is not only in the business of arriving at the truth or correctness of certain principles and judgments, but also of exploring whether those principles exert a (doubtlessly imperfect) guiding influence on our intuitive judgments in cases like the Asian Disease case. Non-consequentialists in particular tend to be concerned with *being* moral—with acting *from*, as opposed to *just in accordance with*, the moral principles they explicitly and correctly avow. It matters to them that the action not be simply the result of a mechanism subject to arbitrary and irrational psychological forces.

We hope that this clears up any objection to the effect that we have mixed apples and oranges in our discussion, by wrongheadedly trying to compare an empirical/descriptive theory to normative ones. Even when normative theories draw a strict distinction between the empirical and the normative, it can be important to them that the norms show some sign of having a life in our actual thinking and judgment.[16] Furthermore, our intuitive judgment in ethical cases serves as *evidence* for the truth of certain ethical principles, and counter-evidence against others, because we take our intuitive judgments to have some authority in ethics. The authority of those judgments is seriously called into question if they are generally the result of a non-ethical (and possibly irrational) mechanism. If they are, it would be fair to wonder whether reflection was nothing but a post hoc rationalization of the judgments (as Jonathan Haidt and others argue).[17] Again, van Roojen seems unconcerned with the issue.

F.M. Kamm is perhaps the best-known contemporary philosopher who works under the assumption that our ethical intuitions have authority, and that we can make that authority explicit by discovering the principles that inform and influence them.[18] Moreover, she has been especially effective in showing, across a variety of ethical cases cited in support of Prospect Theory, that the cases have *not* been constructed with enough care to serve as evidence for the theory. We think it worthwhile to summarize some of her main results, since they can be cited in

support of our general line, and (more important) have been remarkably and unduly ignored by both ethicists and prospect theorists alike, including Kahneman.

Kamm was the first to point out, we believe, that conceptual analysis seems to indicate that Prospect Theory's loss-versus-no-gain distinction cannot be collapsed into the traditional ethical harm-versus-not-aid distinction (a version of the distinction marked by Doing and Allowing), noting at the same time that the one does not explain the other.[19] For one thing, loss imposed by an agent is different from loss due to natural causes or accident or whatnot. Also, loss and no gain is about what the victim undergoes, while harm and not-aiding is about both what the victim undergoes and what the agent does or doesn't do. Moreover, given the way Prospect Theory describes a loss, it is possible for someone to suffer a loss as a result of someone's not-aiding. Kamm notes, for example, that in the Asian Disease case, not-aiding the 600 results in a loss when they are framed as alive and well. She also suggests that the traditional distinction will not be subject to the same framing effects. In anticipation of empirical tests of her suggestion she cautions that even if conceptual analysis and theory can articulate moral distinctions that are resistant to framing, experimental subjects might nevertheless be duped into deploying the distinctions in inconsistent ways under certain frames (the Asian Disease frames, for example).

We are of course sympathetic to these pertinent and pointed observations, and, if nothing else, wish to recover them from general neglect. In our discussion, later, of Prospect Theory's baseline troubles, we defend and develop an important lesson from Kamm about these distinctions. However, we are in general conducting a more aggressive critique of the implications Prospect Theory wishes to draw from the Asian Disease Problem and related cases; our critique does not depend upon a predicted resistance to framing, nor, alternatively, on a plea that experimental subjects were duped into inconsistent application of a moral distinction. The non-consequentialist thinking we've described does *not* resist, but rather *follows* the effects of the frames in the case, while being at the same time coherent and reasonable. We therefore take ourselves to have refuted the claim the intuitive use of the moral distinctions in the case are worthless, and in need of replacement by the prospect theoretic ones. We have also made the point that such replacement would lead, at least in some instances, to absurdity, as does the suggestion that the neutral outcome for Group 1 (all lost) is ethically neutral.

Of greater interest to us then is how Kamm's wide-ranging discussion supplements our critique of the Asian Disease case. We illustrate with two non-rescue cases that she discusses, before we address in the next section, the famous Original Trolley case, and related rescue cases.

Consider the following:

Snow Shovel Case

Loss Case. A spring blizzard leads a store to raise the price on its snow shovels.

No Gain Case. A store does not reduce the price of its snow shovels when it gets them cheaper from its dealer.

Although the final price is the same in both cases, respondents tend to say that the first case is unfair, the second fair. Kahneman maintains that respondents set the pre-blizzard price as the baseline for the first case. This of course makes the price hike a loss. The sense of unfairness can then be explained as the predicted aversion to loss from the baseline. The store's not reducing its price in the second case, when it gets the shovels cheaper than usual, is simply a no-gain from the same baseline.

What stands out right away as problematic about the cases is that in the one case there is a blizzard and in the other there is not. Raising prices in the face of a blizzard seems to take advantage of people, and even suggests danger. Plus, two different sorts of loss are involved: those imposed by an agent, those not. If we equalize the cases by taking the blizzard out of the first one, then we have a case where a store simply raises its prices (no increased need). This may be irrational but it does not seem unfair, even though it is a loss from the baseline. If we equalize the cases the other way, by putting the blizzard into the *No Gain* case, is it as bad for the one store not to pass on its savings as it is for the other to impose a higher price? If not, then perhaps it's because of the difference between harming and not aiding.

In this way, Kamm shows how the keeping of all things equal vitiates the case for Prospect Theory.[20]

There is one other non-rescue ethical test case worth considering, since it is Kahneman's favorite case of framing:[21] namely, the famous tax case, put by the economist Thomas Schelling. Here, condensed by us, is how Kahneman presents it:

Schelling's Child Tax Case

> (I) A standard exemption is allowed for each child and the amount of the exemption is independent of the taxpayer's income.
>
> Should the child exemption be larger for the rich than for the poor?

Most people find the idea of favoring the rich by a larger exemption completely unacceptable. Schelling then points out that the tax law is arbitrary. It assumes a childless family as the default case and reduces the tax by the amount of the exemption for each child. The tax law could be rewritten with another default case, as below.

> (II) Families with fewer than the two children pay a tax surcharge.
>
> Should the childless poor pay as large a surcharge as the childless rich?

Most people reject this with as much vehemence as the first. But the difference between the tax due by a childless family and by a family with two children is described as a reduction of the tax in the first version and as an increase in the second. If in the first version you want the poor to receive no less a benefit than the rich for having children, then it seems that in order to be consistent, you would be OK with the poor paying the same penalty as the rich for being childless (contrary to intuitive judgment). Whatever the imagined surcharge turns out to be,

the childless rich and childless poor would be out the same amount of money due to it, just as in the exemption case. If you think the surcharge should be greater for the rich, then that is equivalent extensionally to their getting a bigger break than the poor for having two children. Assume a surcharge of $1000 for not having two children. This is extensionally equivalent to an exemption of $1000 for having two children, because whether under the exemption or surcharge, if childless, you are $1000 poorer than you would be if you had two children.

So, we seem to get an intuitively baffling inconsistency in our ethical judgment, just by reframing. Kahneman's view is that, as with the Asian Disease Problem, Schelling's Tax Case shows that our intuitions are worthless as a guide to principled reasons for our selections (*if* there are any principled reasons for them, which Kahneman doubts). Our intuitive judgment or imperative—Favor the poor!—is unreliable. As Kahneman says,

> It generates contradictory answers to the same problem, depending on how that problem is framed. And of course you already know the question that comes next. Now that you have seen that your reactions to the problem are influenced by the frame, what is your answer to the question: How should the tax code treat the children of the rich and the poor?
>
> Here again you will probably find yourself dumbfounded. You have a moral intuition about differences between the rich and the poor, but these intuitions depend on an arbitrary reference point, and they are not about the real problem. This problem—the question about actual states of the world—is how much tax individual families should pay … You have no compelling moral intuitions to guide you in solving that problem. Your moral feelings are attached to frames, to descriptions of reality rather than to reality itself. *The message about the nature of framing is stark: framing should not be viewed as an intervention that masks or distorts an underlying preference. At least in this instance—and also in the problems of the Asian disease … —there is no underlying preference that is masked or distorted by the frame. Our preferences are about framed problems, and our moral intuitions are about descriptions, not about substance*.[22] [Emphasis ours]

But as Kamm points out, the word "exemption" suggests a *benefit* that the government gives you with money it was entitled to take from you. "Surcharge" suggests the taking away of money you were entitled to, and thus a *harm*. Whether the surcharge *is* a harm of course depends on whether the purpose of the surcharge (e.g. increasing the birthrate) justifies cutting into what you were entitled to. It may be justified if imposed on the rich who can afford it, but not on the poor who can't.[23]As in our own narrative for the Asian Disease selections, the ethics of benefits is different from the ethics of harm (or avoiding harm), even if the outcomes are extensionally equivalent in terms of expected utility (dollars likely held or not, or people likely alive or not, at the end of the day).

Like the Asian Disease case, the celebrated Snow Shovel and Schelling Tax cases are adduced by Kahneman as evidence for his theory, in his recent best selling book (Kahneman 2011). We thought it worthwhile to restate in detail Kamm's criticisms of the two celebrated non-Asian Disease cases, since, curiously, Kahneman's book

makes no mention of Kamm's criticisms of these cases, (nor of Kamm at all), nor any adjustments to his arguments in light of her criticisms, which were delivered years ago in her role as commentator to Kahneman's 1994 Tanner Lectures.

We maintain that Kamm's analysis helps establish that Prospect Theory is not yet ready to infiltrate the ethical domain, if it ever will.

§5. It's time now to consider some of the standard rescue cases discussed in the philosophical literature, including Foot's especially productive and important runaway tram, or, as we will call it, the Original Trolley case (somehow the literature changed the tram into a trolley).

Horowitz's paper (1998) is of particular relevance here. She concentrates on two cases, Rescue 1 and Rescue 2. These are minor variations of cases (1) and (2) above, and often discussed in connection with the Original Trolley.

Famous rescue cases

> Rescue 1: We can either save five people in danger of drowning in one place or a single person in danger of drowning somewhere else. We cannot save all six.
>
> Rescue 2: We can save the five only by driving over and thereby killing someone who (for an unspecified reason) is trapped on the road. If we do not undertake the rescue, the trapped person can later be freed.

In discussing Horowitz's treatment of these cases, Kamm (1998) and Sinnott-Armstrong (2008a) have addressed the following two very similar cases.

> Rescue 1': We are about to start rushing to the hospital in a car with five dying people in the back. Beside the road, someone, whom only we can help, is dying, but if we stop to help her, we will not save the five.
>
> Rescue 2': We are about to start rushing to the hospital in a car with five dying people in the back. In the middle of the road is a healthy person who cannot be moved. If we go on to the hospital to save the five, we will, as a foreseen side effect, run over and kill her.

We lay out both sets of cases in order to keep track of the dialectic across authors, who have unfortunately confused things a little (though not essentially) by addressing slightly different cases.

Most people think it clearly right to save the five in Rescue 1 and 1', but not in Rescue 2 and 2'. This seems to reflect sensitivity to the distinction in Doing and Allowing, according to which the presumption against actively killing is stronger than the one against allowing to die.

On Horowitz's prospect theoretic account, we select as we do in Rescue 1 because we conceive of everyone as on the brink of death (about to drown). Then, only gains/no losses are involved, and we are free to maximize gains without any worry about risk of loss. In Rescue 2, the person on the side of the road is alive and well. There is a loss therefore if we drive on over him, and simply no gain if we don't.

Looked at this way, loss aversion dictates that we not run him over. This is not offered as a justification, but only as an explanation, in accordance with Prospect Theory (Horowitz, p. 378).

In Rescue 1', as in Rescue 1, everyone is about to die, so the baseline is all lost (as good as dead) again. So we are again free to maximize gains without risk of loss. In Rescue 2', as in Rescue 2, running over the one is a loss, not doing so is a no-gain. Again, loss aversion dictates that we do not proceed roughshod over the one.

Consider now this critical response from Kamm, imaginatively true to form. Change the frame in Rescue 1' by supposing that the one person is alive and well, but about to be devoured by a wild animal if we do not stay with him (presumably a loss, as in the Asian Disease "will die" frame). In order to prevent this loss, it does not seem permissible to let the five die even though doing so involves no loss/no gain. Prospect Theory does not account, therefore, for this intuitive response. (Rescue 1 could be changed in similar ways to the same effect.) Suppose, on the other hand, the five in Rescue 2' are in excellent shape, but will die of a disease about to hit town unless we go right away to get a drug to prevent their death (i.e. prevent a loss of five). In our effort to get the drug, it still doesn't seem permissible to run over one who is dying, even though doing so involves no loss/no gain.

Sinnott-Armstrong (2008a) replies, in defense of Prospect Theory, with this:

> [In Kamm's frame of Rescue 1'] a failure to save the five is supposed to involve losses to the five, because they are alive and well at present, so the baseline is healthy life. There are, however, other ways to draw the baseline. The disease is headed for town, so the five people are doomed to die if they do not get the drug (just as a person is doomed when an arrow is headed for his heart, even if the arrow has not struck yet). That feature of the situation might lead many people to draw the baseline at the five people being dead. Then not saving them would involve no-gains rather than losses, contrary to Kamm's claim. Thus prospect theory can explain why people who draw such a baseline believe that we should not cause harm to save the five in this case. (57)

(Something similar might be said about Kamm's other variation.)

Sinnott-Armstrong anticipates a possible objection, namely that in the Asian Disease case (Group 2, "will die" frame) "the baseline was not drawn in terms of who is doomed " (as he puts it), even though the disease is, as above, headed for town and about to kill 600 people. But he adds this:

> However, prospect theory need not claim that the baseline is always drawn in the same way. People's varying intuitions can be explained by variations in where they draw the baseline, even if they have no consistent reason for drawing it where they do. Thus, Horowitz's explanation does seem to work just fine in such cases. (57)

Our response to this defense of the theory is to note that so much flexibility in the attribution of baselines (with no independent evidence offered for the proposed variations) threatens the theory with vacuity, inasmuch as it suggests that we can

attribute to respondents, with no other rhyme or reason, whatever baseline works for the theory to turn out right.

We can make more trouble of this sort for the baselines, through consideration of the Original Trolley. The trolley is heading right for the five, so it's like the deadly arrow. The person on the other track is stuck (like them) but otherwise fine. We can let the trolley run over the five, or else redirect it over the one. Then, presumably, the baseline is five as good as dead and one alive and well. This is so especially if baselines are drawn in accordance with what would happen if there were no intervention.[24]As in Rescue 2, loss aversion dictates that we *not* redirect the trolley. But most everyone elects to redirect it over the one (a loss), rather than let it travel over the five (no gain/no loss). So, using this baseline, Prospect Theory does not predict the great majority of intuitive responses to the Original Trolley. This is a damaging result, since the record of intuitive responses to the Original Trolley is about the most robust and stable of any ever studied in this area.[25]

In the spirit of Sinnott-Armstrong's remarks we could suppose the baseline to be different in the Original Trolley, from Rescue 2 type scenarios. We can get the right prediction for Prospect Theory if in Original Trolley the baseline is "all as good as dead" or "the one as good as dead and the five alive and well." But we see no reason to make these suppositions except that they save the theory. In fact the suppositions seem plainly false. The only perhaps not entirely implausible way to save Prospect Theory here would be to change the baseline to "all alive and well," in which case redirecting the trolley simply minimizes losses. We have not seen any independent evidence offered for thinking that the baselines could change in this way from the exceedingly similar arrow to the heart case, or Rescue 2 type cases (where 5 are about to die), and doubt that any offered would bear scrutiny. The trolley is a juggernaut of impending doom, like the arrow to the heart, or the threat to the five in any of the Rescue 2 type scenarios. Unless we are in the grip of the theory, it is highly implausible that these very similar impending lethal threats would induce respondents to draw incompatible baselines, across very similar scenarios. The only reason we can see for the suggestion that they do is that they save the theory.

It is implausible to suggest that respondents would *ever* do this, and perhaps worse to suggest that they at times do and at other times don't. Recall that in Rescue 2 type scenarios, we must run roughshod over a stuck, but otherwise fine, person on the road in order to save five others from some impending doom. The most common intuition in those cases is that we are not permitted to run him over. But if the Rescue 2 baseline is "all alive and well," as in the just mentioned (*implausible*) attempt to save Prospect Theory in the Trolley case, then the theory predicts that we will minimize losses by running roughshod over the one—the wrong prediction. The only way to make the theory work here would be to return to the *plausible* baseline of "the one alive and well and the five as good as dead." Then loss aversion would dictate not running roughshod over the one. This accounts for the most common intuition in the case, but of course causes trouble again in the Original Trolley. The only way to get the right outcome across the two types of cases—Original Trolley and Rescue 2 type cases—would be to

suggest that those about to die in the one case are alive and well, and those about to die in the other, as good as dead. This of course threatens the theory with vacuity.

The only truly plausible baseline across these cases is "the one alive and well, the five as good as dead." It might be suggested that Prospect Theory should just stick to this baseline in the Original Trolley and account for the intuition that we ought to divert the trolley, by maintaining that the gains are great enough (5 saved from death) to outweigh the generally greater aversion to loss (since it's just 1 lost). But then the theory will still not be able to account for the most common responses in Rescue 2 type cases.[26]

In a similar vein, we might ask why, on the theory, respondents are induced to change the baseline in the Asian Disease case from where the prompt seems to set it. The prompt says that the disease is *expected to kill* 600, which would seem to put the baseline at 600 as good as dead, on analogy with arrow to the heart, or trolley coming at you, or (equivalently) to disease about to hit town. But then somehow the "will die" language disposes Group 2 to switch the baseline to 600 alive and well. Why?

It cannot be suggested that the prompt sets no baseline, and thus no baseline to be changed. It obviously is playing *some* role, presumably one similar to the role played by the prompts in the Richer Decision Problem, where being told you are $500 richer, or $300 poorer, changes your baseline from whatever it was before. Moreover, "the will be saved language" makes sense only as "will be saved from being killed," i.e. from the all as good as dead baseline set by the prompt (like the saved language in the drowning Rescue 1 scenario).

On the other hand, that people *will be killed* seems not that different from saying that people *will die*. So, if you think "will die" sets the baseline at "alive and well," then perhaps that is a reason to set the prompt baseline at alive and well. But then somehow, the "will be saved" language disposes Group 1 to switch it to as good as dead. Why?

Or else, why does "expected to kill" and "will die" dispose us to set the baseline at alive and well; while "dying," "about to die," or "arrow to the heart" dispose us to set it at as good as dead?

Any way one looks at it, something's funny about how baselines are supposedly drawn on the theory. We need a reason to attribute such perplexing shifts in baseline to the experimental subjects, and the reason cannot of course be simply that the shifts are what make the prospect theoretical predictions come out right. So, there seem to be serious difficulties with baselines, or reference points—a crucial element of the theory. At the very least, there is some unclarity about how to attribute baselines, as Horowitz candidly admits. Although, she also admits (like Kahneman) that the baseline is often the status quo.[27] But the status quo baseline (what happens without intervention) will not help with the general difficulties we have been discussing (it would, for example, still give the wrong prediction in one or another case: Original Trolley or Rescue 2 type cases).

Now it might be countered that these difficulties also affect our own narrative of the results of the Asian Disease Problem, and that therefore we also must

concede arbitrariness or irrationality in our account of how baselines are drawn by respondents confronted with that decision problem.

However, we do not think this is so, because what's crucial to our narrative is only that the "will die" language disposes respondents to be guided by the ethics of harm and that the "will be saved" language instead disposes them to be guided by the ethics of benefits, in considering which intervention to execute. We therefore retract, as unnecessary to our narrative, the earlier assumption that framing causes the baselines to be set, alternatively, at "alive and well" or "as good as dead." In this way, we avoid having to explain why baselines shift from the prompt, in the perplexing ways that they appear to on Prospect Theory's assumptions (not to mention the perplexing shifts needed to account for the Original Trolley and related rescue cases). In the end there is really no issue about baselines for us in the Asian Disease case or the others; harm or benefit need not be a loss or gain from a baseline. Harm in particular can come from the means as much as from the ends, in non-consequentialist views. Earlier, in the context of the "will die" frame, we spoke in terms of the side-effects killing the "alive and well"; but, again, that phrase can be dropped. All that matters is that innocent people will die for sure, perhaps due to a harm introduced by the intervention itself.

To sum up, Prospect Theory must refer, as it always does, to a reference point or baseline in order to explain the effects of the framing (in contrast to our own account). Worse, it seems that Prospect Theory must shift around reference points in ways that threaten the theory with vacuity.

We look forward to empirical testing of our own account. However, and at long last, we wish to caution against the evident tendency to overrate and over-generalize the statistically significant results of such investigations,[28]—especially when they float free of a promising theory (as most do)—and to lose sight of the larger lessons. What we have learned is how hard it is for the best researchers (including philosophers) to tell whether cases have been written properly—in particular, whether they have, among other things, kept things ethically equal across cases. It can take a lot of thought and time before we discern the genuine implications of even the best empirical studies (the cleverest and best known and confirmed), as opposed to what might seem at first (and for a long time after!) to follow from them. The tendency to err here, without recognizing it, is evidently strong and may be due to the fact that remarkable results are attention getting, while confirmations of traditional and widely accepted views are not. In any case, our argument should be seen as part of a recent counter-trend to the new wave of empirical (or experimental) philosophy: a counter to the proliferation of empirical studies and a call to focus more careful conceptual attention on the most remarkable and widely confirmed studies (especially when the empirical results seem to have a good theory behind them) supposedly calling into question the rationality or reasonableness of our judgments.

§6. We conclude with a summary of our paper.

The Asian Disease Problem does nothing to support Kahneman's influential claim that our intuitive judgments are subject to the arbitrary effects of framing

and thereby rendered hopelessly inconsistent. We show this without questioning the empirical results of the case, but rather by supplying a ready-to-hand ethical narrative that reflects non-consequentialist sensibilities, along the lines of the principle of Doing and Allowing. What everyone has missed in discussions of the Asian Disease Problem is that the effects of the changes in wording cannot be assumed to be arbitrary effects, with no moral significance. Once this is seen there is no need to question the empirical results obtained from respondents (which are in any case extremely well-confirmed), nor to suppose that the respondents are being duped by the frames into making non-rational or irrational judgments, without moral significance. On the contrary, the judgments are easily interpreted as reasonable and ethically principled.

F.M. Kamm has given similar arguments against the Snow Shovel Case and Schelling's Tax Case, cases which, like the Asian Disease Problem, are among the most important cases cited in support of the claim that our intuitions are subject to arbitrary effects of framing and thus worthless as data for principled ethical accounts. So far as we can see, then, the prospect theoretic case has yet to be made for the debunking of our moral intuitions.

Moreover, the prospect theoretic account of our intuitions in the Asian Disease case suffers from complications introduced by its theory of how we set baselines or neutral outcomes. The complications occur quite generally across standard life and death rescue cases—including Foot's Original Trolley Problem—that prospect theorists have thought they could explain better than traditional ethical principles, particularly non-consequentialist ones along the lines of Doing and Allowing. We demonstrate that in order for the theory to correctly predict the most common intuitive judgments across these life and death cases, including the Asian Disease case, the baselines must be assumed to shift around with no rhyme or reason, other than that the shifts make the predictions come out right. This of course tends to make the theory vacuous, unless independent empirical evidence can support such baffling shifts in baselines. We know of no such evidence, and note that Kahneman and Horowitz have been candid about difficulties with the account of how baselines are set. They have proposed the general rule of thumb that baselines are set by considering what outcome would occur if there were no intervention. But this does not resolve the complications we have raised.

Our own account of the Asian Disease results (and rescue cases, etc.) does not require the positing of baselines in the manner proposed by Prospect Theory. This is because unlike the gains and losses of Prospect Theory, the harms and benefits of non-consequentialism are not necessarily linked to outcomes from a baseline. In particular, non-consequentialism recognizes certain harms as intrinsic to interventions themselves, independently of outcomes (whether from baselines or not). This is the heart of non-consequentialism, as represented in the distinction between Doing and Allowing, and in non-consequentialist ideals of the intrinsic value of persons and non-violability of their rights. Such ideals fare better than Prospect Theory as explanations of our judgments when confronted with the spectrum of cases: Asian Disease, Rescue cases 1 and 2, and the Snow Shovel case, for example. They are also consistent with our intuitions in Schelling's famous tax case.

Warren Quinn's well known attempt to solve the problem of the Original Trolley, through a unique and subtle reading of Doing and Allowing, was the target of Horowitz's pioneering, prospect theoretic attack on the distinction. Whatever one thinks of Quinn's solution, Prospect Theory presents no reason to oppose it, much less a viable alternative to it.[29]

Notes

[1] Horowitz 1998.

[2] See §5, below.

[3] Which says, roughly, that it is sometimes permissible to allow a harm to occur that it would not be permissible to actively bring about.

[4] Good places to begin investigating the wave are Knobe and Nichols 2008, and Sinnott-Armstrong 2008b. See also Appiah 2009 and Doris 2010.

[5] First presented in Kahneman and Tversky 1981, and widely cited since then.

[6] In addition to the pioneering article by Howoritz, see, for example, Driver 2005, note 39, Sinnott-Armstrong 2008, pp. 54-57, Appiah 2009, pp. 82-88, and Merritt, Doris and Harman 2010, p. 359.

[7] See §4, below.

[8] See Kahneman and Tversky 1981, p. 453.

[9] Kahneman 2011, p. 368.

[10] The case is discussed in Horowitz 1998, p. 369.

[11] See Kahneman 2011, pp. 369 and 437, and also pp. 17&18, below.

[12] See Kahneman 2011, pp. 370 & 437 (quoted at length and discussed below). Cf. also the literature cited in note 6, above.

[13] Kahneman also candidly shows, by the way, that the idea of the reference point being neutral can lead to other absurdities, even just in the non-ethical domain of money. See Kahneman 2011, p. 287.

[14] A law of what Kahneman calls System 1 of the mind: i.e. the "lazy," or more or less unreflective, unconscious and automatic system that, on his theory, generates many of our intuitions. See Kahneman 2011, chapter 1.

[15] Kahneman 2011, p. 437.

[16] Kant is a good example of this. No one is stricter or more adamant about the distinction between the moral and the empirical. Yet it is crucial to Kant's theory that the moral law necessarily has an influence on the human will, even a *felt* influence. See his *Critique of Practical Reason*, Chapter III: "On the Incentives of Pure Practical Reason."

[17] See Jacobson 2012 for a critical review of Haidt and others writing in this vein.

[18] What she says in a recent interview is typical of her attitude: see Voorhoeve 2011, chapter 1. See also the introduction to Kamm 2007, pp. 4–5, and p. 8, note 4, where she describes her method and suggests that intuitive ethical judgments might originate in deep ethical internal structures, along the lines of Chomsky's theory of grammatical judgments.

[19] Kamm 1998, pp. 470f.

[20] Our presentation of the Snow Shovel case follows Kamm's in Kamm 1998, which is based on her invited commentary to Kahneman's Tanner Lectures of 1994. Kamm reproduces virtually the same commentary in Kamm 2007, chapter 14. Kahneman adduces the *Loss* Snow Shovel case, with minor variations, in Kahneman 2011, p. 306. He does not mention or engage Kamm's earlier criticisms, except perhaps tacitly by omitting the *No Gain* case from the account (which, in fairness, does not appear in the article where the *Loss* case first appears, see Kahneman et als 1986, p. 729. However, the discussion in that paper is concerned with economic theory rather than moral theory). Still, even without the *No Gain* case, it is problematic to conclude that the perceived unfairness in the *Loss* case confirms Prospect Theory's account rather than one which appeals to the idea that endangered people are being taken advantage of.

Kahneman (2011, p. 306) presents one more case (the Photocopy Shop case), and cites still more cases (a bit later on), as evidence of the fact that people tend to normativize the status quo, but again without considering the suggested evidence to the contrary we've just seen in Kamm's critique of the

Snow Shovel cases (e.g. in her suggestion that it's not unfair to raise prices when there is no blizzard and no increased need).

[21] See Kahneman 2011, p. 369.

[22] Kahneman 2011, p. 370.

[23] Kamm 1998, pp. 483–484.

[24] See Kamm 1998, p. 45, where Kamm cites Jonathan Baron's suggestion that this is where people tend to draw the baseline.

[25] Otuska 2008, note 2, cites major studies.

[26] In fairness to Horowitz, she never explicitly tries to account for the common intuition that it is wrong to run roughshod over the one, as in the Rescue 2 type cases, nor does she discuss the Original Trolley. Her target is Warren Quinn's intuition and his claim that it is explained by Doing and Allowing. Quinn's intuition is that it is clear that we can save the five (at the expense of the one) in Rescue 1 type cases, but not so clear that we can run roughshod over the one to save the five in Rescue 2 type cases (Quinn p. 290). Of course, her argument is that Prospect Theory explains the intuitive difference better than Doing and Allowing, like so: we take killing to be loss from a baseline of alive and well, and letting die to be no gain/no loss. Due to loss aversion, the drop in value from the baseline in killing the alive and well is greater than the increase in value from the baseline when saving the all-but-dead. This difference in value is "perceived as a difference in the force of one's reasons in the two cases" (Horowitz 378). So Horowitz is focused on the intuition that saving the five in Rescue 1 type cases is morally unproblematic, while doing so in Rescue 2 type cases is not. This puts the intuitive data more mildly than we have been in discussing Sinnott-Armstrong and Kamm. But our position holds all the same, given the common intuitive judgments in the Original Trolley, where most people confidently think one can (or ought) to divert the trolley (but are not so clear about roughshod). How can this be if the one is alive and well and the five as good as dead? Loss aversion would seem to make us treat this case more like Rescue 2, than like Rescue 1 cases. But the opposite seems to be true. People must be drawing a different baseline. But, as above, we can't see any reason for thinking so, apart from commitment to the theory.

[27] Horowitz 1998, p. 371. Cf. Kahneman 2011, p. 287.

[28] Considering the once iconic, but now beleaguered, p-value. See, http://www.nature.com/news/scientific-method-statistical-errors-1.14700, https://edge.org/response-detail/25414, and https://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/.

[29] This paper has its origins in Sandra Dreisbach's doctoral thesis (Dreisbach 2012). We would like to thank committee members Jorge Hankamer, Paul Roth and Ellen Suckiel for their help in making it a successful thesis, with so much proven potential. We would also like to express our gratitude to David Hoy who assisted the early topic development and research, the graduate students in the UC Santa Cruz Philosophy department for their support, and a special thanks to Kaija Mortensen and Matt Dreisbach for their thoughtful discussion and constructive feedback through the development of both the thesis and the paper. Daniel Guevara began teaching and reflecting on the topics in this paper several years ago, and would like to thank the students in his Advanced Ethics course (2012–13), who helped him develop the first stages of his critical view of Prospect Theory. The paper was delivered at UC Santa Cruz (Philosophy Department Colloquium, 2012, and Audun Dahl's Graduate Seminar, Psychology, 2016), Pomona College (Philosophy Department Colloquium, 2014), and UC Riverside (Philosophy Department Colloquium, 2014). We are grateful to those institutions and to those who participated in the very helpful discussion of the paper—special thanks to Jonathan Ellis, Andrews Reath, Kyle Robertson, Eric Schwitzgebel, and Julie Tannenbaum, for also providing written comments on various drafts. Audun Dahl and Dana Nelkin read and commented on a late draft; we are thankful for their very supportive remarks, which helped secure our confidence in the paper. We also gratefully acknowledge the careful and stimulating comments of the anonymous referees for *Noûs*.

References

Appiah, Kwame Anthony. 2009. *Experiments in Ethics*, Cambridge: Harvard University Press.

Doris, John M. 2010. *The Moral Psychology Handbook*, New York: Oxford University Press.

Dreisbach, Sandra. *The Impact of Psychological Research on Moral Decision Making: Prospect Theory and the Doctrine of Doing and Allowing*. Diss. University of California Santa Cruz, 2012.
Driver, Julia. 2005. "Normative Ethics." In *The Oxford Handbook of Contemporary Philosophy*, Frank Jackson and Michael Smith (eds.). Oxford: Oxford University Press, pp. 31–62.
Foot, Philippa. 2002. "The Problem of Abortion and the Doctrine of Double Effect," in Philippa Foot, *Virtues and Vices*, Oxford: Oxford University Press, pp. 19–32. Originally in the *Oxford Review*, 1967, pp. 5–15.
Horowitz, Tamara. 1998. "Philosophical Intuitions and Psychological Theory," *Ethics*, pp. 367–385.
Jacobson, Daniel. 2012. "Moral Dumbfounding and Moral Stupefaction," *Oxford Studies in Normative Ethics: Volume 2*.
Kahneman, Daniel. 2011. *Thinking Fast and Slow*, New York: Farrar, Straus and Giroux.
Kahneman, Daniel, Knetsch Jack L., Thaler, Richard. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *The American Economic Review*, pp. 728–741.
Kahneman, Daniel and Tversky, Amos. 1981. "The Framing of Decisions and the Psychology of Choice," *Science*, pp. 453–458.
Kamm, F. M. 2007. *Intricate Ethics*, New York: Oxford University Press.
Kamm, F. M. 1998. "Moral Intuitions, Cognitive Psychology, and the Harming-Versus-Not-Aiding Distinction," *Ethics*, pp. 463–488.
Knobe, Joshua and Nichols, Shaun, eds., 2008. *Experimental Philosophy*, New York: Oxford University Press.
Merritt, Maria W., Doris, John M., and Harman, Gilbert 2010. "Character." In Doris 2010, pp. 355–401.
Otsuka, Michael. 2008. "Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases," *Utilitas*, pp. 92–110.
Quinn, Warren 1989. "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing," *The Philosophical Review*, pp. 287–312.
Sinnott-Armstrong, Walter. 2008a. "Framing Intuitions." In Sinnott-Armstrong 2008b, pp. 47–76.
Sinnott-Armstrong, Walter, ed., 2008b. *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, Cambridge: The MIT Press.
van Roojen, Mark. 1999. "Reflective Moral Equilibrium and Psychological Theory," *Ethics*, pp. 846–857.
Voorhoeve, Alex (ed). 2011. *Conversations on Ethics*, Oxford: Oxford University Press.