

## CHAPTER 10

### THE NONIDENTITY PROBLEM AND THE TWO ENVELOPE PROBLEM:

#### WHEN IS ONE ACT BETTER FOR A PERSON THAN ANOTHER?

M. A. ROBERTS

[Appears in *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, Roberts and Wasserman, eds. (Springer, 2009), pp. 201-228]

**Abstract.** The nonidentity problem and the two envelope problem have in common an ongoing resistance to intuitive analysis. They also share certain structural features. I argue that the two problems proceed under the same error, imagining subjects to draw haphazardly from a potpourri of actual and expected values to generate results about betterness and harm rather than, as we naturally do and always should, drawing in a more discriminating way from a more orderly array. When we play by the same set of rules in calculating the values that we then compare, we, in particular, become able to discern (1) *harm* in just the cases in respect of which one important type of nonidentity problem has long been thought to show “no harm done” and (2) *no harm done* in the two envelope problem, which purports to show “harm done” when the subject refuses endlessly to switch from one envelope to another and back again.

### 10.1 PARALLEL PROBLEMS

The nonidentity problem and the two-envelope problem may seem like games. But they are hard games, perennially resisting intuitive resolution. Moreover, the nonidentity problem is widely considered to have destroyed any hope that moral theory can be grounded in the highly intuitive “person-affecting,” or “person-based,” approach to ethics (“PBA”). According to PBA, the moral status of a given act is determined by facts that are *person-based* in nature. Included in PBA is, for example, the core idea that “what is bad must be bad for someone.”<sup>1</sup> That core idea—the “person-based intuition” (“PBI”)—expresses nothing more than a simple necessary condition on wrongdoing: the act that is “bad for” no existing

and no future person cannot be morally wrong. PBA, in contrast, includes both PBI and additional necessary and certain sufficient conditions on wrongdoing as well, with the principles that constitute PBA united in their effort to connect wrongdoing with how persons—as individuals, not in the aggregate, and including at least some non-human animals alongside many humans—are *affected* for better or worse.

On the face of things, PBA is a compelling way of thinking about how moral law is structured. It seems plausible that an act’s moral status can be tied directly to what that act does—or, at least, can be expected to do—for people. Moreover, the proposition that moral theory is not person-based—that it must be “impersonal” at least in part—creates its own set of challenges. Impersonal, or mixed, theories cannot easily make sense of straightforward questions of moral obligation.<sup>2</sup> Nor can they serve as the basis of a Dworkinian “political morality” we would like to be able to appeal to in addressing hard questions in respect of which the text of the law is indeterminate.<sup>3</sup>

If we do find PBA—including PBI—at least credible, the nonidentity problem will be of grave concern to us. That problem purports to show that at least some “bad” acts are “bad for” no one at all. In this paper, I will focus on one especially powerful type of nonidentity problem—what I will call the “can’t-expect-better” problem.<sup>4</sup> That problem type includes Kavka’s slave child and pleasure pill cases, Parfit’s depletion and risky policy cases and Sher’s and Shiffrin’s transgenerational compensation puzzles.<sup>5</sup> Global warming is perhaps the world’s biggest can’t-expect-better problem, compelling Broome, for one, to jettison PBA in favor of an impersonal approach.<sup>6</sup> I want to concede that, *if* these problem cases have been understood correctly—if, in particular, the wrong acts I concede they involve really do harm *no one*—then these problem cases decisively refute PBI and hence PBA. My argument

here, however, is that these problem cases have not been understood correctly: we *think* they establish the “no harm done” result they purport to establish *only* because we have committed a certain fallacy in connection with our reasoning about just when what we do today harms, or makes things worse for, persons who will not exist until tomorrow.<sup>7</sup> If I am correct, then this particular type of nonidentity challenge against PBI fails.

A very different type of nonidentity problem—the “can’t-*do*-better” problem—includes, among others, Parfit’s two medical programs case and cases of “wrongful disability.”<sup>8</sup> The argument made in connection with the can’t-*do*-better problem appeals to the fact that, in some cases, it is simply not biologically, or physically, possible for the particular person who has been brought into existence by the act under scrutiny to exist and *not* suffer a particular impairment. From there it is inferred that the procreative act itself does not make things worse for, or harm, that person. In the rare case, then, in which that same act harms *no one else*—*no other existing or future person*—PBI dictates that the act itself cannot be wrong. Here, the result that the procreative act does not harm the impaired child, so long as that child’s life is not less than worth living, is one that I believe we cannot avoid. I argue, however, that the further result that the act is not wrong is one that we can accept. If I am correct, then this second type of nonidentity challenge against PBI fails as well.

This is not to say that the can’t-*do*-better problem is not a difficult one. But my own view is that the can’t-*expect*-better problem is still less tractable. In the can’t-*expect*-better problem, the position that the act under scrutiny is permissible is simply not plausible. Moreover, the phenomenon that Kavka describes as the “precariousness” of existence, which the can’t-*expect*-better problem appeals to for purposes of demonstrating that the act under scrutiny—the “bad” act—is “bad for” no one, is riveting in its own right and seems

unassailably linked to the question of harm.<sup>9</sup> Each of us, Kavka notes, has made it into existence *against all odds*. Eliminating any act—including the clearly “bad” act—in the causal sequence of highly particularized acts and events that ends in the conception of any one person, and substituting in the place of that act any clearly permissible act, very probably will result not in a different or better life for that one person but rather in *no life at all* for that one person. As Parfit puts it, how many of us would have existed had “motor cars . . . never been invented”?<sup>10</sup> But the precariousness insight applies not just to earth-shaking events like the invention of motor cars or Hitler’s baser acts not having been countermanded earlier on, but also to things like whether a couple dines at one restaurant rather than another—or pauses, in one of Kavka’s cases, to take a teratogenic “pleasure pill”—just before conceiving their first child.<sup>11</sup> Any little variation in the causal sequence can easily affect, at least in some slight way, the timing and manner of conception—and the slightest variation in the timing and manner of conception would have all but assured that the particular person who exists and suffers would have been taken off-track for existence altogether. The permissible act might have meant the conception of another, “nonidentical,” better off person in place of the one. But it would, it seems, not have made things any *better for* the particular child who in fact exists and suffers as a result of the clearly “bad” act. If anything, the permissible act, by, very probably, leaving that child out of existence altogether, would, very probably, just have made things *worse* for that child.

According, then, to the can’t-expect-better problem, once we recognize the precariousness of existence, we are forced to accept that the clearly “bad” act is not “bad for” what might seem to be its clearest victim. Surely, after all, it is better to have an existence that is, if flawed, nonetheless worth having than it is never to have existed at all. One act in

fact produces the former outcome for a person, while any other act, very, very probably, would have produced the latter. How, then, can that one act be “bad for” that person? If the broken arm inflicted as a matter of necessity in the context of a rescue is *not* a harm, then how can the suffering that accompanies as a matter of *near*-necessity all that is precious about life from the perspective of the one who lives constitute a *harm*?

I will argue, however, that the gap between necessity and near-necessity is much wider than many theorists have taken it to be. My argument will not involve any particularly creative account of when it is that an act is “bad for”—or *harms*—a person. I, instead, take it for granted that the comparative approach exploited by the can’t-expect-better problem is, in itself, unproblematic. According to that account, an act *harms* a person if that act makes things “worse for” that person than they needed to have been—if, that is, an alternative act would have made things “better for” that person than they in fact are. An act harms a person, in other words, if that act creates less wellbeing for that person when the agent—or group of agents—had the alternative of creating more. Thus, I harm you (under ordinary circumstances<sup>12</sup>) when I shoot you in the arm, not because (1) you have ended up shot in the arm and you suffer<sup>13</sup>, and not because (2) you “would have been” better off had I not shot you in the arm<sup>14</sup>, but rather because (3) I had the alternative of not shooting you at all and that alternative would have made things better for you than you are, my having shot you in the arm.

More specifically, my argument will be that the can’t-expect-better problem fails, not when it insists on a comparative approach to harm, but when it beguiles us into drawing haphazardly from a potpourri of *actual* and *expected* wellbeing levels (“values”) and, on the basis of a comparison between some such pair of values, deciding that an act does not make a

person any worse off than that person would have been under any alternative act. When we instead take care to select our value pairs in a more discriminating way from a more orderly array—as we naturally do and always should—we come to quite different results on both betterness and harm. We come to results that seem both intuitively plausible (surely, e.g., the child *is* harmed when its parents take the teratogenic pleasure pill prior to conceiving that child, and, just as surely, taking the pill *is* wrong) and not at all at odds with either PBI or PBA.

This is not to suggest that it is a mistake to bring expected value to bear in determining betterness. Doing so allows us to construct a moral theory that determines on a *prospective* basis what our moral obligations are. And we will consider that capacity critical if we think that morality has an action-guiding function. But betterness can be tricky, in two ways that are of particular import for purposes of evaluating the can't-expect-better problem. First, betterness between acts cannot be reliably determined by a comparison between actual and expected values. And, second, even if we do discipline ourselves to compare (just) expected value against expected value (or actual value against actual value), any calculation of expected value will come with its own hazards. In particular, once we know the future has unfolded in a certain way, it can be very hard, as a practical matter, to calculate *as though* we have no knowledge of that fact whatsoever. Yet that is exactly what we must do, if our expected value against expected value comparison is to determine in any reliable way whether one act is worse for a person than another.

The *two-envelope problem* exploits our epistemic vulnerabilities in the reverse way. We are urged to imagine that the future has *not* unfolded in a particular way when we are quite aware that it has. We are urged to think that we do *not* know a certain thing that we in

fact know quite well. Again, as a practical matter, it is very hard to calculate as though we *do* know a certain thing when the very design of the case strenuously urges us to think we *don't*.

Unlike the nonidentity problem, the two-envelope problem has made barely a dent in the thinking of philosophers aiming to understand the structure of morality. But perhaps more note should have been taken of that problem, for it so nicely demonstrates just how much can go wrong in how we think about when it is that one act is better for a person than another. Illicit assumptions, whether they cover things we know but are not “supposed” to know, or things that we really are *supposed* to know but think we are not, quietly create chaos in our expected value calculations. But without those essentially confabulated calculations, we never obtain the problematic results we would so love to avoid: that clearly “bad” future-directed acts are “bad for” no one at all, and that (endlessly) switching from one envelope to the other can somehow constitute a worthwhile endeavor.

## 10.2 A PERSON-BASED APPROACH TO PROCREATIVE CHOICE

### 10.2.1 *The choice not to conceive a child*

The person-based intuition (“PBI”) provides a sensible account of many issues relating to procreative choice. According to PBI, an act (including any omission) is wrong at a world *only if* it creates less wellbeing for a person who does or will exist at that world when the agent had the alternative of creating still more wellbeing for that same person at some other world.<sup>15</sup> I will reserve the term “harm” for the case in which that necessary condition is satisfied. Correspondingly, an act that (arguably) harms *only* those persons who count as *merely possible* relative to a world, by way of failing to bring those persons into existence to begin with, must be deemed permissible. According to PBI, then, losses incurred by the

merely possible in virtue of their never having existed at all are without moral significance: they cannot make an otherwise permissible act wrong.

Important practical implications ensue. Consider, for example, the choice whether to conceive a child. That that child—that any child then conceived—would have a life worth living—a positive lifetime wellbeing level—if he or she were brought into existence does not, according to PBI, put *any* moral pressure on us to bring that new person into existence.<sup>16</sup> The choice *not* to procreate becomes—in many instances—a clearly morally permissible alternative.

In contrast, impersonal, aggregative forms of consequentialism suggest, in surprising scenarios, that the choice to procreate is obligatory. Such views focus on whether the choice to bring the new person into existence increases total or average aggregate wellbeing or, under pluralism, increases aggregate wellbeing enough to counterbalance any values or ideals that weigh against that choice.<sup>17</sup> The implication (often) will be that we are wrong not to bring the new person into existence—even if the choice not to bring the new person into existence makes things better for some persons who do or will exist (e.g., the woman who bears the child) and worse for none. That seems implausible. That moral law imposes on us such stringent procreation obligations—to produce a first child, a fifth child, a tenth child—seems highly implausible, however happy the child we might have had would have been.

### 10.2.2 *Abortion*

Not conceiving a child is one thing. Aborting a fetus *may* be quite another—depending on the timing of the abortion. Until that point in the pregnancy at which a *person* has come into being, PBI will deal with abortion just as it does non-conception. The

implications of the woman's abortion choice for the *non-person fetus* will not, according to PBI, create any grounds whatsoever for a moral objection to that choice. The abortion choice will be treated differently, however, at that point in the pregnancy at which a new person has commenced existence. In that case, PBI leaves the door open for a finding that the choice is wrong. But, as a simple necessary condition on wrongdoing, PBI cannot do any more than leave that door open, even in the case of abortions performed in the last minutes of the full-term pregnancy when (it seems) we clearly do have a person. Yet PBI can be understood to be part of a broader view—a *person-based approach* (“PBA”) that includes certain sufficient conditions as well as additional necessary conditions for wrongdoing.<sup>18</sup> One such sufficient condition would be the following (Paretian) principle: an act is wrong at a world when there exists an alternative to that act that creates additional wellbeing for at least some persons who do or will exist at that world without creating less wellbeing for any person who does or will exist at that world and without bringing any additional persons into existence. According to that principle and on the assumption that the late pregnancy involves a person whose life will be worth living, if there is no cost to the woman or to anyone else who does or will exist in allowing the pregnancy to continue, and if it's not the case that (somehow) allowing the pregnancy to continue would mean that one or more additional persons would be brought into existence, then the abortion is wrong.

Of course, there may well be some diminution in the woman's wellbeing involved in forgoing even the very late-term abortion and having the child. The very late abortion could, in other words, involve a *tradeoff*—a situation in which the agent can increase wellbeing for one person only by decreasing wellbeing for someone else. While I will not try to articulate a set of person-based tradeoff principles here, there is certainly no reason to think that PBA

will not include them.<sup>19</sup> PBA determines wrongdoing by reference to morally significant, person-based facts—and tradeoff scenarios are replete with just those sorts of facts. Thus, where the agent’s only alternatives are between (1) reducing the fetal-person’s lifetime wellbeing to the very low level it will have as a function of having had only a very abbreviated time in existence and (2) reducing the woman’s lifetime wellbeing in some more modest way, PBA can be expected to say that the abortion is wrong.

PBI and PBA thus serve to center the abortion debate on issues we intuitively take to have moral relevance to that debate, including, for example, the neurological status of the developing fetus, the emergence of consciousness and, most generally, when it is during the pregnancy that the developing organism counts as a *person*. In contrast, the aggregative approach sets aside, or at least minimizes, issues that intuitively seem of grave moral significance in favor issues that seem peripheral at best. Thus, under both totalism and averagism, the non-person fetus has just about the same moral status as the fetal-person does. Similarly, under pluralism, the distinction between aborting the non-person fetus and aborting the fetal-person becomes far less relevant than it seems that it ought to be.

### 10.2.3 *The moral significance of merely possible persons*

The appeal of PBA is increased when we realize that it can, and I think should, be understood not to place on a moral pedestal either (1) *actual* persons—persons who do or will exist at the *actual* world—or even (2) persons who *would exist* were the act under scrutiny in fact performed. The fact that less wellbeing has been created for a person when the agent had the alternative of creating more can bear on the moral status of a given act, according to PBA, even if that person is neither actual nor someone who would exist were

that act performed. Thus, PBA does not claim that the losses incurred by the merely possible *never* have moral significance or that the merely possible *never* matter morally. Those losses do matter morally, so long as they are incurred at worlds where the persons who incur them do or will exist, for purposes of determining the moral status not just of acts performed at those worlds but of acts performed at still other worlds as well. Thus, PBA recognizes that the plights of the merely possible may imply that a wrong has been done—not at the world at which those persons are merely possible but rather at a second world where those same persons do or will exist and suffer as compared to how those same persons fare at still a third world. Under PBA, the “genocidal adventures of nonactual dictators” remain a no-no, even where the victims of genocide are all “nonactual” as well.<sup>20</sup> Moreover, there is no reason to think that the fact that the merely possible and their losses matter *there* does not bear on the moral status of an act performed *here*. An act may thus be deemed perfectly permissible at a world in the case where the only way to create more wellbeing for the relatively well-off persons who do or will exist at that world without creating less wellbeing for any of those same persons is to bring into existence still additional persons who are then treated very badly relative to some third world—who are, for example, made to serve as slaves or organ donors for the rest of us.<sup>21</sup>

#### 10.2.4 *The moral status of future persons*

It is also a plus for PBI that PBI is consistent with—and PBA *requires*—our denial of a view Narveson seemed to suggest decades ago and that other theorists have at least flirted with since: that, if the “children produced have a good chance at a good life, we think people should have them if they want them.”<sup>22</sup> But we do not think that a procreative choice is

*always* permissible so long as the new persons are “happy” or that it is *always* enough to give our own offspring a “good chance at a good life.” Completely independent of the nonidentity problem, we think that making “happy people” often isn’t morally neutral.<sup>23</sup>

PBA, which I have set forth here as an inherently *maximizing* approach (that it does not aggregate does not mean it does not, on a person-to-person basis, maximize), reflects that stringent standard. Suppose the difference between a child being born with spina bifida or not is (just) the difference between the woman’s taking a vitamin just after conception or not. Clearly, the woman’s refusal to take the vitamin harms the child then born with spina bifida. That is so even where that child’s life will be unambiguously worth living if the woman does not take the vitamin.<sup>24</sup> But once we establish that the act harms an existing or future person, we immediately avoid, under PBI, the implication that no wrong has been done. If the case is also one in which the woman’s taking the vitamin does not harm any other existing or future person, including the woman herself, and does not bring still other persons into existence who are then maltreated, PBA can be expected to do still more—that is, to imply that what the woman has done is wrong.

#### 10.2.5 *Genetic interventions*

A last point in favor of PBA is that it adjusts its results nicely in the face of technological change. Suppose that a genetic technology has been developed that corrects the Huntington’s gene in the newly formed embryo. And suppose that that technology has been made available to the couple but they refuse to allow it to be used to correct “their” embryo’s mutated gene. However ample their resources and unambiguously worth living the child’s life will be even without the correction, the couple cannot rely on PBI to make the case that

what they have done is permissible. For their choice clearly harms their child: they have created less wellbeing for that child when they could have created more. Moreover, depending on the tradeoffs—if any—that are involved, PBA will imply that the couple’s refusal to allow the correction to be made is wrong.

#### 10.2.6 *Fit with law*

The particular branches of the law that most directly address issues relating to future persons—constitutional privacy law, tort law and family law—themselves seem intransigently person-based in nature. Tort law ties wrongdoing to the harming of some person or another. And constitutional privacy law allows the state to regulate what it considers to be wrongdoing only when the state can tie that wrongdoing to the harming of persons—existing or future.<sup>25</sup> Family law focuses on risks to offspring. Where would-be parents can demonstrate their conduct is devoid of any risk to any existing or future person, the provisions of family law that would normally permit agents of the state to intervene in the parent-child relationship or remove children from the custody and control of their parents have no application.

What this means is that a person-based approach has the capacity to provide guidance in dealing with hard cases involving future persons in respect of which the law itself is indeterminate. In contrast, the legal system and totalism (or averaging or pluralism) will be like two great ships passing in the night (and one of them will be the Titanic). Moral theory will generate findings of harmless wrongs that courts will find legally irrelevant—or will consider acts to be permissible that courts will view with alarm. More generally, an impersonal, aggregative form of consequentialism will not have a sufficiently good “fit” with

the law to be useful in helping us to sort through the complex array of future person cases we now face—from abortion to the “custody” of frozen embryos to supernumerary pregnancy to global warming.

We could, of course, always restructure the law—a mere human product—along impersonal, aggregative lines. But such a restructuring is one that we should at least hesitate to undertake. PBA is well-entrenched within the law, and the law itself is a product of thousands of years of intense human effort. Moreover, the aggregative approach itself is hardly without blemish. None of this means PBA is correct, of course. But it does mean that it is worth a close look.

### 10.3 THE NONIDENTITY PROBLEM

#### 10.3.1 *Two types of nonidentity problems*

The nonidentity problem is really just a collection of different problems displaying distinct logical features. Those problems, accordingly, can be typed. When we fail to sort nonidentity problems in accordance with their types, we may well think “the nonidentity problem” shows that some “bad” acts are “bad for” no one. In contrast, when we analyze the problems in accordance with their type, we come to quite different results. We can then see that (1) the problems that really do demonstrate “no harm done” to any person—the “can’t-*do-better*” problems—are exactly those in respect of which it never becomes quite clear that a wrong has been done, and (2) the problems that involve acts that are clearly wrong—the “can’t-*expect-better*” problems—are exactly those in respect of which it never becomes clear that that same person has not been harmed.

### 10.3.2 *The can't-do-better problem*

The can't-*do*-better problem arises when the agent's procreative choice harms no existing or future person—that is, when it creates at least as much wellbeing for each existing and future person as any alternative choice the agent might have made instead.<sup>26</sup> Suppose that a couple's choice to bring a Huntington's child into existence meets that condition. Suppose, also, that the couple had the alternative of bringing into existence a nonidentical, genetically healthier child in place of the Huntington's child. The can't-*do*-better problem successfully challenges PBI only if we agree that, on those facts, the couple's choice is wrong. But is it really so clear to us that it is?

In answering this question, it is critical to note just how rare the bona fide “no harm done” case really is. Even if the *procreative* effect of the couple's procreative choice does not constitute a harm to the impaired child, the *distributive* effects of that very same choice may well constitute harms—to the impaired child's older or younger siblings or to the impaired child or to both. Distributive harms will arise when the couple scrimps on the resources—time, energy and money—they would otherwise expend on the impaired child in order to insure that that child's older and younger siblings are protected from any ill effects of their choice. Alternatively, the couple may scrimp on expenditures for those other children in order to insure that the impaired child has the care that he or she requires. Either way, they have put themselves in a moral bind by making the procreative choice in the way that they have. They have insured that harm will befall some or all of their offspring.<sup>27</sup>

But if there is harm—if the act is “bad for” someone or another—then the case is not an instance of the can’t-do-better problem. The can’t-do-better problem arises *only if* the couple’s choice harms *no one*—*only if*, that is, it is *maximizing* for each existing and future person. It seems to me that, in that rare case, any sentiment that we might have that the couple’s choice is wrong will itself begin to fade. If so, then our conviction that the case proves PBI false must begin to fade as well.

### 10.3.3 *The Can’t-Expect-Better Problem*

In contrast to the can’t-*do*-better problem, the can’t-*expect*-better problem includes many cases in which the choices under scrutiny seem incontrovertibly wrong. In any such case, rescuing PBI requires a showing that the logic that takes us to the “no harm done” result is itself mistaken.

I will focus on just one instance of the can’t-*expect*-better problem here—Kavka’s slave child case.<sup>28</sup> There, a couple enters into a binding, enforceable contract with a wealthy man according to which the couple will conceive and bear a child who will be transferred at birth to the wealthy man as a slave. In exchange, they will receive \$50,000, which they will use not to save the world but to buy a yacht. The couple then produces a child as a slave.

Let’s call that child “p.” Despite p’s status as slave, p’s life is worth living.

Has the couple’s act of bringing a child into existence in this particular way *harmed* p—created, that is, less wellbeing for p when the couple had the alternative of creating more?<sup>29</sup> The couple, of course, had a number of alternatives—including not entering into the contract and still taking steps to produce a child, and not entering the contract and *not* taking

steps to produce a child. The latter of those two alternatives obviously would not have generated any additional wellbeing for p. But what about the former?

Kavka, citing the “precariousness” of existence, argues that it, too, has little to offer p.<sup>30</sup> After all, had the couple not acted just as they did, what would the chances have been that the very same gametes that happened to combine to produce p would still have combined? What would the chances have been that the very same sperm (out of hundreds of millions!) would still have inseminated the very same egg? Practically none at all. Surely, then, from p’s own point of view, it is better for the couple to enter into the contract. Surely their choice, if anything, benefits p rather than harms p.

#### 10.3.4 *A closer look at the can’t-expect-better problem*

We need to look more closely at this argument. Let “A” be the couple’s act of entering into the contract and taking steps to produce a child (their entering into the contract and, let’s suppose, having sex). That is what the couple has in fact done. Let “B” be their act of *not* entering into the contract yet still taking steps to produce a child. B clearly existed as an alternative for the couple at the critical moment just prior to their performance of A. That is so, even if it is also true that, had the couple *not* performed A, they *would not* have chosen B—even if, that is, they would have instead refrained from producing a child altogether (“C”).<sup>31</sup>

We should agree that the *probability* that p will come into existence, given B, is very, very low.<sup>32</sup> That is simply to recognize the phenomenon of the precariousness of existence. Moreover, by hypothesis, A confers on p an existence worth having. One act, B, generates

for p an unbelievably tiny chance of ever coming into existence at all, while another act, A, generates for p an existence worth having.

These points seem clearly correct. Nonetheless, if we try to determine on the basis of just these points whether A really is at least as good for p as B is, or whether A harms p, then we allow ourselves to be rushed down the garden path. This is so, for two reasons. (1) The claim that p's chances of existence are very small, given B, at least bears rewriting. For surely just how small p's chances are of coming into existence, given B, depends on just how B itself is realized. (2) Moreover, betterness presumably is not established by comparing the *probability* of p's achieving a certain wellbeing level given one choice against the actual wellbeing level in fact generated for p by another. After all, probability is just a number between 0 and 1. Yes, the *actual wellbeing level* that A generates for p may well be substantially greater than any such number. But it would be premature to conclude, on that basis alone, that A is better than B for p. We can take these two points in turn.

(1) *Choice of B as a way into existence for p.* Let's call the highly particularized way in which the couple in fact realizes the generic A—how they realize A at, e.g., the *actual* world—"A\*." We can note, then, that there are a lot of ways in which the couple could have realized A that do not bring p into existence. It is their concrete performance of A in all its identity-influencing detail—their performance of A\* in place of A\*\*, A\*\*\*—that brings p into existence. But those details have *nothing* to do with the fact that A\* is an act of entering into the slave child contract and *everything* to do with the various spatial-temporal-mechanical characteristics of A\* that put the right sperm in the right place at the right time.

That, in turn, means that the alternative act B *can* equally well put the right sperm in the right place at the right time. It will all depend on *just how* the couple goes about

performing B. Clearly, among the many, many alternative ways of realizing the generic B, there exist some that *perfectly mimic A\* in all respects relevant to p's coming into existence*—that include, that is, the very same spatial-temporal-mechanical identity-influencing details we find in A\*. Call some such concrete, particular way of realizing B “B\*.”

We should note, as well, that B\* clearly exists as an alternative for the couple at that critical time just prior to performance. Nothing in natural law or the acts of other agents prevents the couple from performing B\* in place of A\*. They *can* perform B\* just as easily as they *can* perform A\*.

Moreover, as between A\* and B\*, there is no basis for thinking that the couple's performance of B\* will end in p's never coming into existence at all. If A\* fits into how the future will otherwise unfold in a way that brings p into existence as a slave, then B\* will fit into how the future will otherwise unfold in a way that will bring p into existence as a non-slave. If A\* makes p's coming into existence likely, then so does B\*. If A\* makes p's coming into existence a tiny bit less than highly improbable, so does B\*.

Can't we then, after all, say that the couple has created less wellbeing for p when they could have created more—that A\* is not, after all, at least as good for p as B\* is—and hence that what the couple has done in performing A\* harms p?

We *might well* say that—but only if we are *actual value* consequentialists.<sup>33</sup> But we might not be actual value consequentialists. We might think instead that moral law is based not on *actual* value but rather on *expected* value. For we might (not implausibly) believe that the moral assessments that we make, at least in theory, can have an *action-guiding function*. An assessment of wrongdoing based on actual value is one that can be reached only in

hindsight (if at all). That approach leaves the agent open to charges of wrongdoing even when the agent is conceded to have made the best possible choice given all the information within the agent's grasp at that critical moment just prior to choice. The great thing about expected value is that it is something we can calculate, at least in roughly, before we act. It takes into account that we cannot know precisely how the future will unfold and, indeed, that how the future will unfold may not be a determinate matter of fact.

If we think that expected value is critical to the issue of permissibility, then we will find in the slave child case a far more formidable challenge to PBI. We agree that A\* and B\* are on par in terms of bringing p into existence—that if A\* does so (and it does), B\* would have done so as well. Yet even then the can't-expect-better problem may seem able to argue its way to the results that A\* is at least as good for p as B\* is and that A\* does not harm p.

(2) *Relevance of probability.* If probability is important in connection with the slave child case, it is important because it tells us something about how much *expected value* B generates for p. And hence we call the particular type of nonidentity problem that relies on probability the “can't-expect-better” problem.

Let's just pause here to note an important constraint that our analysis will be subject to if we do indeed opt to jettison an actual value theory in favor of an expected value theory. For a moral theory to have the action-guiding function that the concept of expected value is meant to make possible, expected value must itself be something that we can calculate, at least roughly, *just prior to choice*. That fact, in turn, constrains—and, indeed, helps us to identify—just which probabilities can properly be understood to bear on that calculation. In particular, we must accept the following rule:

*Calculation-prior-to-choice constraint:* The probabilities relevant to the calculation of expected value are just those determined on the basis of the information within the agent's grasp at the critical moment just prior to performance.

This constraint sometimes does not make itself felt in any obvious way. It does not, for example, lead to any surprising perturbations in how we calculate the expected value for the generic B. Let "t0" be the moment just prior to performance. Suppose that as the future unfolds and given the couple's choice of A\*, p's level of wellbeing is +100. And suppose that p's level of wellbeing at a world where B is chosen over A\* and p (against all odds) exists as a nonslave is (estimating wildly) +200. We have already conceded that the probability of p's coming into existence, given B, is very low. And that assessment does not change when we explicitly restrict the information on which the assessment is based to just that information within the couple's grasp at t0. The probability of p's coming into existence, given B, at t0 remains very low. Suppose that that probability (estimating still more wildly now) is .0000000001.

Adapting, then, the standard expected value formula for purposes of PBA, we take the summation of the actual values for p of each possible outcome of that act B multiplied by the probability that that outcome will obtain, given the performance of B. On the assumption that the value of never having existed at all for p is itself zero, we then can write:

$$EV(B, p, t_0) = .0000000001(200) + .9999999999(0)$$

And we conclude—as anticipated—that the generic B generates precious little expected value for p.

We are now in a position to compare—not a mere probability, a number between 0 and 1—but a *value*—the *expected* value generated by B for p. But against what? What the can't-expect-better problem seems to ask us to do is compare that very low *expected* value

generated by B for p against the very high *actual* value generated by A\* for p—that is, +100—and then to obtain that A\* is, after all, at least as good for p as B is.

Now, I argue, in part 10.3.5 just below, that we cannot validly reach that particular betterness result. But first let's complete the argument, to see just how the can't-expect-better problem is *supposed* to achieve its ultimate result that A\* does not harm and is not “bad for” p.

It would be premature to infer, from the bare fact that A\* is at least as good for p as B is, that A\* does not harm p. We must first determine more generally that A\* is at least as good for p as *any* other alternative is for p. And in that connection, we must obviously consider B\*, which includes all the identity-influencing spatio-temporal-mechanical details that we find in A\*. It has been conceded that p would have come into existence, given the substitution of B\* in for A\*. It might thus seem that, while  $EV(B, p, t_0)$  is very low,  $EV(B^*, p, t_0)$  should be substantially higher.

But that would be a mistake. Here, the calculation-prior-to-choice constraint does make itself clearly felt. Under that constraint, the optimistic probability of p's coming into existence that we may think attaches to B\*—a number that, in some objective sense may well be correct and that we have no reason to abandon<sup>34</sup>—must nonetheless be set aside. If expected value is to have the action-guiding function that is its very purpose, then the probabilities we rely on in calculating expected value are limited to those that are determined *on the basis of just that information that is within the agent's grasp at t0*. But think about how little is within the couple's grasp at t0—just how little is settled for them at t0! Perhaps it is settled for them at t0 that they will choose A rather than B or C and will together produce some child or another—call it “p”—as a slave.<sup>35</sup> It is not at all settled for them that the

probability of p's coming into existence, given B\*, is anything more than "very low."

Perhaps it is true that (in some objective sense) p's coming into existence, given B\*, is greater than "very low." The problem is that that is nothing the couple can grasp at t0. The couple has no basis for correlating their performance of B\* with the coming into existence of p. For all they know at t0, B\* is just one of the many, many ways of realizing B—ways that include B\*\*, B\*\*\* etc.—that would have taken p off track for existence altogether.

Calculated, then, on the basis of *just that information that is within the couple's grasp at t0*, the probability of p's coming into existence, given B\*, remains very, very low.

Things would be quite different if the couple could grasp at t0 that the particular way they will realize A will be by performing A\*. For then they would have a basis for correlating their performance of B\* with the coming into existence of p. That is: having made the correlation between A\* and p, they could then make the correlation between B\*, which exactly mimics A\* in all of its identity-influencing spatial-temporal-mechanical details, and p. The problem is that, at t0, it remains highly unsettled for them, and may not then even be a determinate matter of fact, that their choice of A will be realized by A\*. For all they then know, they will realize A not by the performance of A\* but rather by the performance of A\*\* or A\*\*\* or etc.<sup>36</sup>

We must now concede that the probability that p will come into existence, given B\* and calculated on the basis of just *the information within the couple's grasp at t0*, remains very low. But that means that the *expected* value B\* generates for p is very low as well. Yet the *actual* value A\* generates for p has not changed. It is still quite high (+100). Yet B\* constitutes p's last, best hope of coming into existence as a non-slave. And thus the argument concludes: no alternative is better for p than A\* is—and A\* thus does not harm p.

### 10.3.5 *The mistake in the “can’t-expect-better-than-get” problem*

The difficulty with the argument we have just described is that we cannot reliably determine betterness, and ultimately harm, on the basis of a comparison between two radically distinct sorts of value—the very low *expected* value generated by B (or B\*) for p against the relatively high level of *actual* value that has in fact been generated for p by A\*. We can certainly compare those two numbers and obtain the result that  $AV(A^*, p) > EV(B^*, p, t_0)$ . But it is a mistake to think that comparison has anything to do with betterness, or harm.

And we can count the ways why that is so. For one thing, a rule that determines betterness by comparing actual against expected values is inconsistent. After all,  $AV(B^*, p)$  may be quite high. Suppose that it is. Yet  $EV(A^*, p, t_0)$  is very low (we come back to this point in what follows; for now, it is enough to note that, for the same reason  $EV(B^*, p, t_0)$  is very low, so is  $EV(A^*, p, t_0)$ ). If we think betterness can be established by a comparison between actual and expected values, we must now say that B\* is better for p than A is. But this way of thinking has already committed us to the view that A is at least as good for p as B\* is. Given that betterness is anti-symmetric, we now face an inconsistency.

For another, it intuitively seems that the only way we can apply an actual-against-expected comparison to get the result that A\* is at least as good for p as B\* is will be to equivocate on “value.” “Value” has to mean, first, *actual* value and, second, *expected* value if the comparison is to connect in the right way to the facts of the case. But betterness for a person has to do with producing *more of a certain stuff* for that person. It is true that the *number* representing  $AV(A^*, p)$  is greater than the *number* representing  $EV(B^*, p, t_0)$ . But it

does not follow that  $A^*$  is better for  $p$  than  $B^*$  is, any more than a comparison between five sifted and four unsifted cups of flour demonstrates betterness in respect of—i.e. more—flour.

Once we have the two numbers—the actual value of  $A$ ; the expected value of  $B^*$ —in which we have a great deal of confidence—and we do—it then becomes very hard *not* to compare those two numbers and think we are validly coming to an accurate result on betterness, and ultimately harm. But we aren't.

We can call this variation on the can't-expect-better problem the “can't-expect-better-than-*get*” problem. It is a fallacy—a bit of reasoning that seems compelling on its face but in fact is mistaken and should be rejected.

#### 10.3.6 *A variation on the can't-expect-better problem*

A critic might object that I have misconstrued the slave child case. Perhaps the “no harm done” result is supposed to be derived not from the problematic actual-against-expected comparison but rather from a seemingly more legitimate expected-against-expected comparison. We can call this variation on the argument the “can't-expect-better-than-*expect*” problem.

We have said that the probability, calculated on the basis of just the information within the couple's grasp at  $t_0$ , that  $p$  will come into existence, given  $B^*$ , is very low. It may seem that  $A$ —and surely  $A^*$ !—makes  $p$ 's coming into existence significantly more likely than *that*. It may seem, then, that we can conclude that  $EV(A, p, t_0) > EV(B^*, p, t_0)$ —or at least that  $EV(A^*, p, t_0) > EV(B^*, p, t_0)$ . Either way, we can conclude, now under the more legitimate expected-against-expected comparison, that  $A$ , or at least  $A^*$ , is better for  $p$  than  $B^*$  is and thus does not harm  $p$ .

### 10.3.7 *The mistake in the “can’t-expect-better-than-expect” problem*

This argument, however, fails as well. There are three questions that we should closely consider.

(1) *How much expected value does A generate for p?* The difficulty with this argument is that the probability of p’s coming into existence, given A and calculated on the basis of just that information that is within the couple’s grasp at t<sub>0</sub>, is itself very low.

Suppose, as before, that it is settled for the couple at t<sub>0</sub> that they will choose A rather than B or C and that, as a result of that choice, they will together produce some child or another p whose existence as a slave will be flawed yet still worth having. That information is within their grasp at t<sub>0</sub>. And it is a lot of information: on the basis of that information, the couple can, for example, reasonably project at t<sub>0</sub> that the *actual* value A will generate for that child p will in fact be relatively high—around +100.

But the *actual* value A will generate for p is no longer at play. Pertinent now is the question of how much *expected* value A will generate for p. To reach the result that A generates so much *expected* value for p that A is at least as good for p as B\* is, the couple would also need to grasp that the probability of p’s coming into existence, given A, is itself significant (is greater, that is, than the very low probability the couple is in a position to assign to p’s coming into existence, given B\*). And, according to the calculation-prior-to-choice constraint, that would be something they would need to grasp prior to choice—prior, that is, to t<sub>0</sub>. But they can’t. At least, on the highly plausible assumption that it remains unsettled for them, at t<sub>0</sub>, that they will happen to realize A by performing A\* rather than

A\*\*, A\*\*\*, etc., all they can grasp at  $t_0$  is that there are a lot of ways of performing A that will mean that p will never exist at all.

In short: all the couple can grasp at  $t_0$  is this: some child or another will be brought into existence by their choice of A, *and* whichever child that happens to be—call it “p”—will have been brought into existence *against all odds*. But since the probability of p’s coming into existence given A, calculated on the basis of just that information within the couple’s grasp at  $t_0$ , is very low, so is  $EV(A, p, t_0)$ .

(2) *How much expected value does A\* generate for p?* Do things change when we turn to A\*? Is  $EV(A^*, p, t_0) >$  than  $EV(A, p, t_0)$ ? More to the point, is  $EV(A^*, p, t_0) >$  than  $EV(B^*, p, t_0)$ ? It is always hard, when we know just how some question of fact has ultimately been settled, to keep in mind that we must think about things as though it has not. In calculating the expected value that A\* generates for p, however, that is exactly what we must do.

We retain the (highly plausible) supposition that it remains unsettled for the couple, at  $t_0$ , that they will happen to realize A by performing A\* rather than A\*\*, A\*\*\*, etc. We then find that, calculated on the basis of just that information within the couple’s grasp at  $t_0$ , the probability of p’s coming into existence, given A\*, remains very low. That is so, for exactly the same reasons that we said before that the probability of p’s coming into existence, given B\*, is very low. We concede (as we did for B\*) that the performance of A\* in place of A\*\*, A\*\*\*, etc. increases (in some objective sense) the probability of p’s coming into existence; we concede that it is true at  $t_0$  that p will, or probably will, come into existence, given A\*. Under the calculation-prior-to-choice constraint, however, that probability must be set aside as irrelevant to the calculation of expected value. The relevant probability is, instead, what

can be calculated on the basis of *just the information within the couple's grasp at t0*. But at *that* time the couple has no basis for correlating their performance of A\* with the coming into existence of p. For all they know at t0, A\* is just one of the many, many ways of realizing A—ways that include A\*\*, A\*\*\* etc.—that will take p off track for existence altogether. Calculated, then, on the basis of *just the information within the couple's grasp at t0*, the probability of p's coming into existence, given A\*, remains very low.

(3) *Abandoning the plausible supposition*. In addressing questions (1) and (2), we have supposed that the couple does not happen to know in advance that they will realize A by performing A\* rather than A\*\*, A\*\*\*, etc. Would that bit of foreknowledge have changed the analysis?

No. We still cannot reach the result that A\* is better for p than B, or B\*, is—or the critical “no harm done” result. Since the couple now is understood to grasp that they will realize A by performing A\*, they can also grasp just how to go about performing B in a way that will perfectly mimic A\* in all its critical identity-influencing, spatial-temporal-mechanical detail. Suppose, e.g., that part of A\* is the fact that the couple will expend exactly six seconds actually signing the slave child contract. All of that is within their grasp since (by supposition) it is within their grasp that they will realize A by way of performing A\*. They, moreover, understand that that particular sequencing will bring their child p “a step closer” to coming into existence. They are thus now in a position to identify an act B\* that has exactly those identity influencing features—an act B\* that involves, e.g., *feigning* to sign the slave child contract for exactly six seconds. The upshot is that—still calculating on the basis of just the information that is within the couple's grasp at t0—whatever the

probability of p's coming into existence, given A\*, will also be the probability of p's coming into existence, given B\*.

But since the actual value p will enjoy, given that B\* is performed and p comes into existence, will be far greater than it is under A\*, we now can obtain that  $EV(B^*, p, t_0) > EV(A^*, p, t_0)$ . We thus again block the inference to the result that A\* is at least as good for p as B\* is and that A\* therefore does not harm p.<sup>37</sup>

### 10.3.8 *Caveat*

My claim here is not that failing to maximize expected value for p in itself *harms* p. Rather, it creates a *risk*. It increases the chance of opening the door to a causal chain that will end badly for p. Where that risk eventuates—as it does in the case at hand; p has not somehow lucked out and gotten to exist as a nonslave; p's actual wellbeing level at the end of the day remains avoidably diminished by p's status as slave—we can say that what the couple has done *harms* p.

## 10.4 THE TWO-ENVELOPE PROBLEM

### 10.4.1 *The argument for switching*

In the two-envelope problem, two amounts of money are covertly placed in distinct envelopes. The envelopes are displayed to the subject, who is reliably told that one amount is twice the other. The subject is also told that he (or she; let's suppose he) may choose one of the two envelopes to keep. He arbitrarily chooses an envelope. He is then offered the option of switching. Where we let "S" stand for the act of switching and let "m" designate the value of the contents of the selected envelope (the "in-hand" envelope), and where we understand

that the probability is .5 that the “out-of-hand” envelope contains the greater amount, we calculate as follows:

$$EV(S) = .5(2m) + .5(.5m) = 1.25m$$

Since the value of not switching, that is, holding (“H”), is  $m$ , and since  $1.25m > m$ , we conclude that S is better for the subject than H is. But that is an odd result. Surely the subject’s initial arbitrary selection of the one envelope does not really involve a mysterious creation-at-a-distance of new value in the out-of-hand envelope.

Things get worse. Let “n” designate the contents of the out-of-hand envelope. We now obtain that  $EV(H) = 1.25n$ . Since  $1.25n > n$ , we also obtain that H is better for the subject than S is. But since betterness is anti-symmetric, we face a contradiction—just as we did in the can’t-expect-better-than-*get* variation on the slave child case.

Why not deny here, as we did there, that betterness can be reliably determined on the basis of an actual-against-expected comparison? We then remain free to take either the view that betterness is to be determined by an actual-against-actual comparison or the view that betterness is to be determined by an expected-against-expected comparison. Whichever view we take, however, we (appropriately) find ourselves unable to reach any betterness result at all.<sup>38</sup> We thus avoid both the odd result and the contradiction.

But not for long. Just as we did in the context of the slave child case, we can rewrite the two-envelope problem as an expected-against-expected problem. And we know independently that if we jettison actual-against-expected comparisons we will need to be willing to take the position that such a rewriting is sometimes plausible. Consider the idea that “a bird in hand is worth two in the bush.” If we think actual-against-expected comparisons do not reliably determine betterness, we can simply rewrite this perfectly cogent

aphorism as an expected-against-expected comparison. It's just that the probability of ending up with one bird is *extremely high*—close to 1—where the agent opts to hold onto the bird in hand. The EV of that choice thus approaches its AV.

A parallel approach to the two-envelope problem generates the following: EV(H) is surely  $m$  (or at least very close to  $m$ ).<sup>39</sup> EV(S) remains  $1.25m$ . We again infer that S is better than H—and by an analogous string of inferences that H is better than S.

#### 10.4.2 *The mistake in the argument*

What triggers the two-envelope problem is the inference from the standard formula to the result that  $EV(S) = 1.25m$ . But we can infer on the basis of that same formula that  $EV(S) = n$ .<sup>40</sup> Those two facts are on their own sufficient to establish a problem—at least on the assumption that, calculated for a given time, agent and world,  $EV(S)$  constitutes a unique value. For it then follows that  $1.25m = n$  and hence that  $n > m$ —all of which can be established prior to the point at which the subject chooses whether to switch. But it can't. The subject has *no idea* that the actual value of the out-of-hand envelope is greater than the actual value of the in-hand envelope. Moreover, it may not even be true that  $n > m$ . In fact, the chances are 50-50 that it *isn't* true. One of the two candidates for  $EV(S)$  must go.

But which? We originally let “ $m$ ” rigidly designate the value of the contents of the in-hand envelope and “ $n$ ” the value of the contents of the out-of-hand envelope. In doing so, we (implicitly) supposed the referents of “ $m$ ” and “ $n$ ” to be fixed—that is, that the contents of the two envelopes cannot themselves change from moment to moment or indeed at any time over the course of the game. Various theorists have relied on that fact to challenge the inference from the standard formula to the result that  $EV(S) = 1.25m$ .<sup>41</sup> On that basis, we can

rule out  $1.25m$  as a candidate for  $EV(S)$ . We are then left to compare  $n$  and  $m$  since those are, after all, the two expected values that we can legitimately calculate. But that comparison obviously provides no grounds to think that it is better for the subject to switch since, by supposition, we do not know what the referents of “ $n$ ” and “ $m$ ” are and hence can make no assessment regarding how they relate.<sup>42</sup>

It is worth explicitly noting why the supposition that the referents of “ $m$ ” and “ $n$ ” are fixed blocks the inference to the result that  $EV(S) = 1.25m$ . The point can be put somewhat roughly as follows: that the referents are fixed, together with the fact that subject has made his initial choice between the two envelopes, means that it is not an unsettled matter of fact what outcome will obtain in the case where the subject chooses to switch. One of the two outcomes that the problem urges us to consider possible outcomes of switching has in fact been taken off the table. The subject does not know which outcome has been taken off the table, but does know that one, or the other, has. That means that there is no future left to unfold in one possible way as opposed to another. The future has already unfolded.

We can put the point more precisely.<sup>43</sup> The original articulation of the problem takes it for granted that the facts of the case support the following application of the standard formula for expected value:

$$EV(S) = .5(2m) + .5(.5m) = 1.25m$$

Clearly, however, by its own terms this application generates the result that  $EV(S) = 1.25m$  only if *both* the probability is  $.5$  that switching will yield  $2m$  *and* the probability is  $.5$  that switching will yield  $.5m$ . But those are propositions the subject does not know. Rather, what the subject actually knows is the following conjunction: *both* the probability is  $0$  that switching will yield  $2m$  or the probability is  $1$  that switching will yield  $2m$ ; *and* the

probability is 0 that switching will yield  $.5m$  or the probability is 1 that switching will yield  $.5m$ . The subject knows, in other words, that *whichever* envelope he has initially selected—whether the lower or the higher-valued envelope—the preceding conjunction will be true. But that knowledge is itself a function of the subject's understanding that the value of the contents of the out-of-hand envelope is a settled matter of fact—that, in other words,  $n$  is fixed and will not vary over the course of the game. But that the subject knows that much—knows, that is, that the probability is 0 or 1 that switching will yield  $2m$  and 0 or 1 that switching will yield  $.5m$ —*precludes* the subject's *also* knowing that the probability is  $.5$  that switching will yield  $2m$  and the probability is  $.5$  that switching will yield  $.5m$ . And without that knowledge, the conditions on the particular application of the standard formula set forth above will remain unsatisfied.<sup>44</sup>

The upshot is that the two envelope game misuses the standard formula. It has taken it for granted that the conditions in the application set forth above have been met—in particular, that the probability is  $.5$  that switching will yield  $2m$  and  $.5$  that switching will yield  $.5m$ —when they aren't. It is not that the standard formula *cannot* be correctly used in the context of the two envelope problem. We can, for example, use the formula to infer that the expected value for switching approaches  $n$ . But we cannot infer anything from the standard formula when its conditions aren't met. We thus are never entitled to reach the result that  $EV(S) = 1.25m$ .

Why did we think, even for a moment, that we know that the probability is  $.5$  that switching will yield  $2m$  and the probability is  $.5$  that switching will yield  $.5m$ ? Because, *coming into the game*, the probability *is*  $.5$  that the envelope the subject will select is the higher-valued envelope and  $.5$  that the envelope the subject will select is the lower-valued

envelope. So it's a game worth playing—even if the price of admission happens itself to be whatever value the lower-valued envelope happens to contain. The problem then tries to beguile us into thinking that we know that those same statistics apply even after the initial selection of an envelope takes place. But they don't. That initial selection changes everything, when it is paired with the fact that the value of the contents of the out-of-hand envelope does not change.

Critical to this way of eliminating the basis for the claim that  $EV(S) = 1.25m$  is the very natural supposition that “m” and “n” are fixed for the course of the game—that is, that the values of the contents of the two envelopes cannot change as the game progresses. A look at an alternative, less natural (one might even say preposterous) supposition clarifies, I think, just how inapt the standard formula is—inapt in the sense that its conditions are not satisfied—when we make the more natural supposition instead.<sup>45</sup> What (we said) blocks the application of the standard formula in the original problem was the supposition that the contents of the two envelopes are fixed throughout the course of the game—that “m” and “n” rigidly designate particular quantities. If we reject that supposition, and put in its place the less natural supposition that the contents of the out-of-hand envelope *can* change, even from moment to moment, as the game progresses, then our analysis changes dramatically.

Suppose, then, that the value of the contents of the out of hand envelope *remains unsettled* even *after* the subject has made the initial selection between envelopes. Suppose, in particular, that the contents of the out-of-hand envelope shift in a way that makes the following claim—false in the context of the original problem—true: that, independently of whether the envelope initially selected (the “in-hand” envelope) contained the greater- or lesser-valued contents as of that moment immediately prior to selection, the probability that

the value of the contents of the out-of-hand envelope will be twice the value of the contents of the in-hand envelope is .5 and the probability that the value of the contents of the out-of-hand will be half the value of the contents of the in-hand envelope is also .5. (We might also have supposed that the contents of the out-of-hand envelope shift in a way that makes it true that the probability of the value of the contents of the out-of-hand envelope will be twice the value of the contents of the in-hand envelope is, say, .8; but we don't.) If that is the supposition we make—if that is the way the game works—the standard formula applies quite nicely. The referent of “m” continues to be fixed (we are not supposing that the value of the contents of the in-hand envelope will vary); and we can then calculate that  $EV(S) = 1.25m$ ; with  $EV(H)$  approaching m, we then obtain the result that it is better for the subject to switch.

But—on this new and unnatural supposition—it *is* better for the subject to switch! As McGrew et al. put it, in that case the standard calculation is “correct, but not at all paradoxical.”<sup>46</sup> Now, as between our two candidates for  $EV(S)$ , it is n we must set aside. For we have now abandoned the supposition that the value of the contents of the out-of-hand envelope remains fixed over the course of the game; there is now no basis on which to think that, as the game progresses, what we call “n” will not fluctuate. But that means that the standard formula will not generate the result that  $EV(H) = 1.25n$ . For that, we would need a fixed “n.”

Whether we stay with the original, natural supposition, or jettison it in favor of the new and unnatural supposition, we find ourselves arguing to inconsistency *only if* we fail to see the inconsistency between the two suppositions and make the mistake of making inferences on the basis of both in constructing the problem.<sup>47</sup> We run into trouble, in other words, only when, having made the more natural supposition, and introduced the terms “m”

and “n” in accordance with that supposition, we then bring to bear the less natural supposition and calculate as though the value of the contents of the out-of-hand envelope may shift over the course of the game. When we keep our inconsistent suppositions apart from one another, as we naturally do and always should, we avoid the inconsistency.

Still another scenario can also help us see just how inapt the particular application of the standard formula—the one that generates the result that  $EV(S) = 1.25m$ —in fact is in the context of the original problem. This scenario involves, not an unnatural supposition, but simply a distinct supposition that we would quite naturally make had we happened to be playing a quite distinct game. We can call it the “four envelope game.”<sup>48</sup> In this new game, four amounts of money are covertly placed in distinct envelopes. But just two envelopes are displayed to the subject—the “middle” two. The subject is reliably told that one amount is twice the other and that he may choose one of the two envelopes to keep. He arbitrarily chooses an envelope. A third envelope is then added to the game. The envelope that is added to the game is the envelope that contains the largest amount of money *if* the subject has selected the envelope that, between the two middle envelopes, contains the larger amount; and the envelope that is added to the game is the envelope that contains the smallest amount of money *if* the subject has selected the envelope that, between the two middle envelopes, contains the smaller amount. Where we let “m” designate the value of the contents of the in-hand envelope (and understand the referent of “m” to be fixed for the course of the game), the subject is reliably told that one of the two out-of-hand envelopes contains  $2m$  and one contains  $.5m$  and is then given the option of switching the in-hand envelope for either one of the (now) two out-of-hand envelopes. What, then, is the  $EV(S)$ ? As the subject understands quite well, independent of how the value of the contents of the in-hand envelope relates to the

value of the single envelope that it was paired with when the subject made his initial selection, the probability is now .5 that the subject, if he switches, will switch to the envelope containing  $2m$ , and .5 that the subject, if he switches, will switch to the envelope containing  $.5m$ . But that means that the condition established by original problem's application of the standard formula is now met—and we can thus calculate that  $EV(S) = 1.25m$ . So it is better to switch than to hold.

But now it *is* better to switch. Moreover, we avoid the result that it is also better to hold than to switch. That is so, since, at the moment just before choice, the value of “ $n$ ” is not fixed—“ $n$ ” simply abbreviates “what the subject gets if he switches.” There is thus no “one thing” that “ $n$ ” can be thought to stand for—and hence no basis on which to argue that it is also better for the subject to hold rather than to switch.

## 10.5 CONCLUSION

Both the can't-expect-better problem and the two-envelope problem have us determine betterness by reference to a comparison between values that we have been beguiled into haphazardly selecting from a potpourri of actual and expected values that can be—under different suppositions about what is settled and what is not—attached to the acts under scrutiny. The truth is that calculating betterness, and harm, is a delicate matter. When we confuse what is and what is not settled for the agent at a given time, our expected value calculations become unreliable. The can't-expect-better problem imports into its scenarios critical claims that the future will unfold in a particular way when a proper calculation of expected value will set those same claims to the side. The two envelope problem reverses things. There, we are urged to think that the future has *not* unfolded in any particular way

when in fact it already has. And a proper calculation of expected value will take that fact into account. It will yield not that it is better to switch, but rather that the two expected values that we can legitimately calculate— $m$  and  $n$ —provide us with no basis for determining whether it is better to switch or not.

When we take care to select our values in a discriminating way from an orderly array—as we naturally do and always should—we come to betterness results that are not disconcerting at all. Interestingly, our results then coincide with the “naive” impressions we might have had about the problem cases when we first examined them (long ago): the slave child has been harmed, and it doesn’t matter a whit whether the subject switches envelopes or not. I think in the end our best theories and best thinking will very likely lead us to exactly the same results.<sup>49</sup>

*robertsm@tcnj.edu*

---

<sup>1</sup> Parfit (1987), p. 363 and generally pp. 351-79.

<sup>2</sup> Traditional, impersonal, aggregative forms of consequentialism, such as totalism and averagism, thus face riveting population problems, including the repugnant conclusion, the mere addition paradox, the infinite population problem and extreme inequality problems. See Parfit (1987), pp. 381-90 and 419-41; Vallentyne and Kagan (1997), pp. 5-26; and Roberts (2002), pp. 322-23. Temkin's remains the best general introduction to this set of problems and to a view, which I will here call "pluralism," that is intended to address the population problems while avoiding the nonidentity problem. See Temkin (1992). Pluralism can be viewed as a form of consequentialism that emphasizes a plurality of values or ideals, including the maximization of aggregate wellbeing as well as individual human flourishing and autonomy, equality and improving the lots of the least well off. See Temkin (1992), pp. 221-27. Pluralism seems plausible on its face. However, while the articulation of totalism and averagism is well underway, the articulation of pluralism is more challenging. We must identify the relevant values and provide an account of how those values are to be balanced against each another. For that reason, pluralism is difficult to test and remains hard to apply in any practical setting—for example, by judges working their way through hard cases with respect to which the text of the law is indeterminate or by women thinking through the ethics of early abortion or even contraception.

<sup>3</sup> See e.g. Dworkin (1986).

<sup>4</sup> Hanser, as well, distinguishes among types of nonidentity problems. See Hanser (2009).

---

<sup>5</sup> See Kavka (1981), pp. 98-101; and Parfit (1987), pp. 361-66 and 371-74. See also Smolkin (1999), pp. 195-96; and Sher (2005), pp. 185-200 (discussing the nonidentity problem in the context of transgenerational compensation, the African slave trade and aboriginal land appropriation cases). See also Shiffrin (2009).

<sup>6</sup> See generally Broome (1992). I borrow from Jamieson here, who describes global warming as the “world’s biggest collective action problem.” Jamieson (2008).

<sup>7</sup> The “fallacy” I elsewhere describe in this paper is presented in more detail in connection with Parfit’s depletion example. See Roberts (2007). One main aim of this present paper is to provide a clearer and more well-grounded account of just why an accurate expected value calculation will not in fact yield the “no harm done” results that are usually attributed to the can’t-expect-better problem. The case I focus on here is Kavka’s slave child case, which I also discussed in Roberts (2003b). See generally Roberts (1998), ch. 3.

<sup>8</sup> Parfit (1987), pp. 366-71. For a brief discussion of wrongful disability, see note 27 below.

<sup>9</sup> Kavka (1981), p. 93.

<sup>10</sup> Parfit (1987), p. 361.

<sup>11</sup> Kavka (1981), p. 98.

<sup>12</sup> Extraordinary circumstances would include those in “The Negotiator,” in which the Kevin Spacey character shoots the Samuel L. Jackson character in the shoulder, thereby saving the latter from certain death at the hands of the true villains. “Harm” is open to distinct conceptions. I believe, however, that there is an ordinary, comparative sense of “harm” in which the shooting in this case does *not* constitute harm. In this sense, the fact of

---

impairment (or, e.g., serious bodily injury) alone is not sufficient to establish harm.

Similarly, I will say that (1) the meticulously performed open-heart surgery that is necessary to save a person's life does not impose a harm at all, notwithstanding the suffering, disability and substantial period of recuperation, in the case where the life saved is worth having, whereas (2) exactly that same meticulously performed open-heart surgery does impose a harm, in the case where an aspirin alone would have done the patient just as much good as the surgery.

<sup>13</sup> See note 12 above.

<sup>14</sup> I thus will not appeal to a “counterfactual,” or “but for,” account of harm. The notion that an act harms a person only if “but for” that act that person would have been better off has been clearly refuted. Suppose I shoot you in the arm, and that (I was so angry that) had I not shot you in the arm I would have shot you in the heart. I still *harm* you when I shoot you in the arm. A better account is to say that an act performed by an agent (or group of agents) harms a person if and only if that agent (or group) has in fact created less wellbeing for that person through the performance of that act when they could have (by performing an alternative act) created more. For that account of harm to be plausible, we need to recognize that the fact that an act harms a person does not, on its own, mean that the act is wrong.

<sup>15</sup> The term “agent” must be understood to include both agents acting individually as well as groups of agents acting collectively (though not necessarily collaboratively or in concert). We will otherwise miss important instances of harm. See Roberts (2007).

<sup>16</sup> We do face here an asymmetry, but one that seems untroubling in view of the underlying person-based principles. But see Persson (2009) and McMahan (2009).

---

<sup>17</sup> For a description of pluralism, see note 2 above.

<sup>18</sup> A handful of person-based principles are stated in more detail in Roberts (2003a), Roberts (2003b) and Roberts (2002).

<sup>19</sup> See note 18 above.

<sup>20</sup> This is Caspar Hare's example. Hare (2006), pp. 498-511.

<sup>21</sup> Hare concedes that PBA should not be interpreted as a form of what he calls *strong* moral actualism. But, contrary to Hare, there is also no reason to think that PBA should be interpreted as a form of what he calls *weak* moral actualism, a view that would assess as impermissible the act that fails to bring the additional persons into existence in the circumstances described here. But that reading of PBA seems unnecessary. PBA should, instead, be understood to take into account the highly person-affecting fact that the additional person will suffer—in an entirely avoidable way—if brought into existence as a slave or an organ donor. If failing to increase wellbeing for an already relatively well-off person p is necessary to avoid bringing another person q into existence who will then be treated very badly, PBA, to remain credible, should be understood to imply that failing to increase wellbeing for p is permissible. Tradeoffs, including trans-world tradeoffs, are going to be an important part of any plausible form of PBA.

<sup>22</sup> Narveson (1976), p. 73. See also Narveson (1967), p. 65. See too Heyd (2009).

<sup>23</sup> Narveson (1976), p. 73. See also Narveson (1978), pp. 55-56 (adopting impersonal principle in response to the nonidentity problem).

---

<sup>24</sup> Her refusal to take the vitamin harms the child, as well, where the woman accurately claims that, if she had been required to take the vitamin, she would never have had the child at all. See note 14 above (on the counterfactual account of harm).

<sup>25</sup> Philip Peters agrees that tort law embraces a person-affecting approach but argues that constitutional privacy law is best understood to include both person-affecting and impersonal values. See Peters (2009). In taking the position that constitutional privacy law is best understood as person-affecting in nature, I am adopting a view that John Robertson has described in substantial detail. See Robertson (2004) and Robertson (1994), pp. 22-42, 75-76 and 168-71.

<sup>26</sup> Another example of the can't-do-better problem is Parfit's "two medical programmes" example. Parfit (1987), pp. 366-71.

<sup>27</sup> Sometimes, of course, the impaired child exists as a result not of parental choice but rather of health care provider negligence. This happens, when the provider fails to diagnose or inform couples of their elevated risk of producing a genetically or congenitally impaired child. For the same reasons, however, that the couple can harm their own offspring by choosing to produce the impaired child in place of the healthier child, so can the provider harm the impaired child or that child's older or younger siblings or both. We must, in other words, take into account not just the *procreative* effect of the provider's negligence on the impaired child, but also the many *distributive* effects of that negligence. On those grounds, I argue that the impaired child—in conjunction with any siblings—may have a valid cause of action against the provider even in the case where the existence itself is clearly worth having. See Roberts (2008). Using the terminology of Buchanan et al., we perhaps most accurately

---

call this claim “wrongful disability” rather than “wrongful life.” See Buchanan et al. (2000), pp. 222-57.

<sup>28</sup> I have discussed the version of the nonidentity problem elsewhere as well. See note 7 above.

<sup>29</sup> I assume here that no interesting reading of the nonidentity problem will rely on the narrower, counterfactual (or “but for”) account of harm. See note 14 above.

<sup>30</sup> Kavka (1982), p. 93.

<sup>31</sup> Thus, the fact that p would have never existed at all had the couple not chosen A is not relevant to the question of whether the couple, in choosing A, harms p.

<sup>32</sup> See Kavka (1982), p. 100 n.15 (“It is enormously improbable that the couple . . . could succeed in producing the same child . . . even if they had tried. For it is unlikely they could arrange conditions of conception similar enough to ‘what would have been’ to insure that the very same sperm would have fertilized the same egg.”).

<sup>33</sup> An actual value consequentialist (whether he or she adopts a total, average or person-based approach) can thus take the following position: B\* is better for p than A\* since it produces, for p, more actual value than A\* does; A\* thus harms p; and, finally, that that harm (given the tradeoffs that are involved; given, e.g., that p’s being born into slavery does not somehow save the world but merely enables the couple to buy a “yacht”) is itself a wrong. That account assumes that, if B\* were performed, then p would exist and be better off than p in fact is given A\*. On that basis, B\* is said to produce more “actual” value for p than A\* does. Since B\* is just like A\* in all its identity-influencing, spatial-temporal-mechanical features, that assumption has strong support.

---

<sup>34</sup> We should concede that the probability of p's coming into existence, given A\* (or B\*), may well be substantial—or at least greater than the very low probabilities we must restrict ourselves to in calculating expected value—in some objective sense. That increase, in effect, is a function of the fact that as the future unfolds in one way rather than another, p's chances of coming into existence will (ordinarily) increase.

<sup>35</sup> Of course, the couple cannot, at t0, “know who p is”—and someone might think that that means that there is no possible act the couple could perform at t0 that could harm, or wrong, p. But it is implausible that a condition on harming a person—or on acting in a way that is morally impermissible in respect of a person—is that we know “who that person is”—or that we have any *de re* attitudes in respect of that person at all. Suppose, e.g., a man shoots into a bustling crowd of shoppers at a mall. He may not know in advance who he will end up shooting, but he plausibly knows that he will end up shooting someone or another. Call that person “p.” Whoever p happens in the end to be, it will be true that the man has harmed p, in virtue of the fact that he could have created more wellbeing for p and has instead created less.

<sup>36</sup> I put this point differently in Roberts (2007). There was some suggestion there that we needed to adopt a concept according to which the relevant probability (for example, the probability of p's coming into existence, given B\*) would itself change over time. The better view, which I have described here, is that what may change over time is simply the particular basis on which the probability is calculated. What changes over time, in other words, is the information that is within the grasp of the couple. The upshot is that the probability we calculate on the basis of what the couple grasps at t0 may be very low even if it is in point of

---

fact true that the probability of p's coming into existence, given B\* and the fact that B\* mimics A\* in all its identity-influencing details, is very high.

<sup>37</sup> In addition to actual value consequentialism and the “action-guiding,” or expected value, consequentialism, still a third account of betterness, and harm, makes use of what we can describe as “objective” probabilities—probabilities, that is, that the agent just prior to choice may well have no way of grasping and thus cannot rely on for action-guiding purposes. On this view, we have no need for the assumption that, if B\* were performed, then p would exist. What is relevant, rather, is that A\* and B\* have the same identity-influencing, spatial-temporal-mechanical features, such that, whatever the agents think, A\* and B\* are equally likely to bring p into existence, and that that is so even if (given that much else about the future also remains unsettled, at the moment just prior to choice, in addition to whether A\* or B\* is to be performed) p's coming into existence in point of fact remains highly improbable. This third view still leaves us unable to reach the result that A\* is at least as good for p as B\* is or that A\* does not harm p. Whatever the probability (relative to a particular time and world) that p will come into existence, given A\*, that is the probability (relative to that time and world) that p will come into existence, given B\*. Calculating value, then, on the basis of such “objective” probabilities, we again never reach the result that A\* is at least as good for p as B\* is. We obtain, instead, just the reverse result—that B\* is better for p than A\* is, and that (given that the risks eventuate; that p is born a slave) that A\* harms p.

<sup>38</sup> If, as above, 1.25m represents the *expected* value of switching, m obviously must represent the *actual* value of not switching. (Otherwise, EV(S) cannot be 1.25m since that amount is calculated by reference to the outcomes associated with switching. If m is not the

---

actual value of not switching, those outcomes will be something other than  $\frac{1}{2}m$  and  $2m$ , and  $EV(S)$  something other than  $1.25m$ .) It is thus an expected-against-actual comparison that has generated the betterness result in this case.

The relevance of this value— $1.25n$ —has been widely noted, often for the purpose of underlining that, once we determine that it is better for the subject, having chosen, to switch, we then determine that it is better for the subject, having switched, now to *switch back*, and so on ad infinitum. Gjelsvik, however, explicitly derives the contradiction. See Gjelsvik (2002). Chase, as well, relies on the symmetrical status of the two-envelopes in his description of what he considers a non-probabilistic version of the two-envelope problem to derive a contradiction. He writes: “Since  $n > n/2$ , it follows . . . that the amount you will gain, if you gain on the trade, is greater than the amount you will lose, if you lose on the trade. But an exactly parallel argument, which begins by dubbing the amount of money in the *other* envelope  $\$n$ , leads to the contrary conclusion that the amount you will gain, if you gain on the trade, is less than the amount you will lose, if you lose on the trade.” Chase (2002), p. 158.

<sup>39</sup> Suggestions that the comparison on which the two-envelope problem is based is expected-against-expected are scattered throughout the discussion of the two-envelope problem. However, if we do rewrite the problem in this way, we must understand “ $m$ ” to designate both an expected and an actual value. For if  $m$  is not construed as an actual value in the initial construction of the problem, then  $EV(S)$  would have to be something other than  $1.25m$ .

---

<sup>40</sup> It's just that, as in the bird-in-hand case, the probability that switching will yield the value of  $n$  is very high. For all practical purposes, it is 1.

<sup>41</sup> See especially McGrew et al. (1997), p. 29 (under the assumption that the amount in the selected envelope remains fixed, but that the “total amount involved in the game” is not, then the standard calculation is “correct, but not at all paradoxical”; under the assumption that the total amount is fixed, the “amount in the selected envelope cannot be taken as fixed. If the (fixed) total amount is, say,  $3x$ , then the selected envelope contains  $x$  if it contains the smaller amount, but it contains  $2x$  if it contains the larger amount . . . . And this means that the [standard] calculation, which assumes that the selected envelope contains *the same fixed amount* whether it is the higher or the lower envelope, is illegitimate”). See, also, Cook (2002), pp. 47 and 49.

Alternative resolutions of the two-envelope problem are grounded in mathematical considerations—relating, e.g., to the calculation of expected utility in the case where there exists an upper limit to the value of the contents of the two envelopes and to the question of whether cases in which there exist no upper limit are possible. See, e.g., Clark et al. (2003), pp. 691-98; Meacham et al. (2003), pp. 685-87; Clark et al. (2000), pp. 415-28; Arntzenius et al. (1997), pp. 42-45; Scott et al. (1997), pp. 37-38; and Broome (1995), pp. 6-10. For purposes here, however, we may set these discussions aside.

<sup>42</sup> Having ruled out  $1.25m$  as a candidate for  $EV(S)$ , we might alternatively adopt the following widely-respected account of why switching is not better than holding: let “ $z$ ” designate the amount in the lower-valued envelope. Then, there are two possible outcomes,  $n=z$  and  $n=2z$ , each having a probability of .5.  $EV(S)$  is then just  $1.5z$ . The parallel

---

calculation yields the identical result for  $EV(H)$ . We find no contradiction and no basis to switch. Gjelsvik, e.g., sets out this line of reasoning, which seems plausible. Gjelsvik (2002), p. 354. See also Schwitzgebel (2008). Of course, a full resolution of the problem also seems to require an understanding of why the standard expected value calculation does not also yield that  $EV(S) = 1.25m$ .

<sup>43</sup> A number of theorists have suggested related grounds for setting aside the results of the standard formula in the two envelope context. See note 41 above. I, however, owe my own appreciation of why the standard formula is inapplicable—and, indeed, the entire contents of the paragraph in the text that contains this note—to a constructive dilemma suggested (though not endorsed) by Ed Gettier.

<sup>44</sup> See note 43 above.

<sup>45</sup> That multiple suppositions are (arguably) consistent with the original description of the problem has been observed by various theorists. See Chase (2002), pp. 159-80; McGrew et al. (1997), pp. 28-30; Cook (2002), p. 47; and Markosian (2005). There are still other suppositions to consider as well—e.g., that the amounts in both envelopes can change as the game progresses. Working with the problem under each of those alternative suppositions—one at a time, since only one can be true in a given case—we also avoid contradiction.

<sup>46</sup> McGrew et al. (1997), p. 29. Markosian makes a similar point. See note 48 below.

<sup>47</sup> “In sum, the ‘paradox’ arises simply from a conflation of assumptions (a) and (b). Where assumption (a) is appropriate, the above calculation is legitimate, but its result is straightforwardly, and quite unparadoxically, correct. Where assumption (b) is appropriate, on the other hand, the above calculation is illegitimate.” McGrew et al. (1997), p. 30.

---

<sup>48</sup> Markosian has previously usefully contrasted the set-up we see in the original two envelope case with alternate games (e.g., what he calls “doubles or halves”). Seeing clearly just how appropriate the standard expected value calculation is in the latter case helps make vivid just why the calculation is inappropriate in the former case. See Markosian (2005).

<sup>49</sup> I owe many people thanks for their extremely helpful comments on earlier versions of this paper, including, in particular, David Wasserman, Nils Holtug, Fred Feldman, Peter Vallentyne, John Sisko, Pierre LeMorvan, Wlodek Rabinowicz, Larry Temkin and Alan McMichael. I presented an earlier version of one part of this paper at a DeCamp Seminar (University Center for Human Values, Princeton University, Oct. 2006). I owe Jeff McMahan, Peter Singer, Elizabeth Harman and other participants in that seminar thanks for their extremely helpful comments. I presented another version at the University of California at Riverside (Feb. 2008), and owe members of the philosophy department there a large debt for their useful and stimulating discussion.