On Whether to Prefer Pain to Pass
Author(s): by Tom Dougherty
Source: *Ethics*, Vol. 121, No. 3 (April 2011), pp. 521-537
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/10.1086/658896
Accessed: 05/09/2013 05:18

On Whether to Prefer Pain to Pass*

Tom Dougherty

Most of us are "time-biased" in preferring pains to be past rather than future and pleasures to be future rather than past. However, it turns out that if you are risk averse and time-biased, then you can be turned into a "pain pump"—in order to insure yourself against misfortune, you will take a series of pills which leaves you with more pain and better off in no respect. Since this vulnerability seems rationally impermissible, while time-bias and risk aversion seem rationally permissible, we are left with a puzzle.

I. TIME-BIASED PREFERENCES: RATIONALLY PERMISSIBLE?

There were warning signs. Alfred's date's musical tastes seemed, well, different from his own. As he looked around the audience of bohemian intellectuals, he started to worry that he was not nearly hip enough to appreciate a performance like this. If truth be told, he had never really enjoyed experimental music. Nonetheless, he had gamely decided to give the avant-garde violin ensemble a go. After only a few erratic and discordant screeches, his worst fears had been confirmed: he had hopelessly blundered. By now, he was trapped, and all he could do was to grit his teeth, longing for his misery to be over.

It is not just that Alfred would have liked the performance to be shorter. He glumly read from his program that, for artistic reasons, it was to be precisely fifty-four minutes long. Even knowing this, Alfred preferred those fifty-four minutes to be in the past. Many of us are like Alfred in that we also prefer painful experiences to pass, whether

* For helpful comments and discussions, I would like to thank Lara Buchak, Tyler Doggett, Ryan Doody, Andy Egan, Adam Elga, Daniel Halliday, Daniel Hausman, Richard Holton, Sophie Horowitz, Vann McGee, Alejandro Pérez-Carballo, Agustín Rayo, Damien Rochford, Miriam Schoenfield, Seana Shiffrin, Bradford Skow, Paulina Sliwa, Robert Stalnaker, Henry Swift, Judith Jarvis Thomson, two anonymous reviewers for *Ethics*, the editors of *Ethics*, and the audiences of the 2009 MIT-ing of the Minds Conference and the 2010 Bellingham Summer Philosophy Conference. I would like to thank Caspar Hare in particular for invaluable feedback on multiple drafts of this essay.

Ethics 121 (April 2011): 521–537

these are cacophonies, dentist appointments, or tax returns. Likewise, we prefer to have fun in the future: we all know the wistful tinge of regret that comes with realizing an enjoyable weekend break is over. We have preferences like these:

> Prefer Pain Over: You prefer a painful experience to be in the past rather than the future.
>
> Prefer Pleasure Not Over: You prefer a pleasurable experience to be in the future rather than the past.

Let us say that if you have these preferences, then you are "time-biased."

Most of us think that you are rationally permitted to be time-biased. Derek Parfit makes this point as part of his sustained critique of what he calls the "self-interest theory."[1] This theory "gives to each person this aim: the outcomes that would be best for himself, and that would make his life go, for him, as well as possible."[2] The rational permissibility of time-bias would pose a problem for such a theory. According to the self-interest theory, you should only be concerned with how a musical experience affects how well your life goes. Whether it is in the past or future should not matter.

Parfit supports the claim that time-bias is rationally permissible by appealing to our intuitions about a case much like this:

> Operations 1: On Monday, you are admitted into a hospital. You are told you will have one of two operations, but you are not told which. If you have the early operation, then you will have a painful, four-hour operation on Tuesday. If you have the late operation, then you will have a painful, two-hour operation on Thursday. After either operation, you will have amnesia for several days, and so you will not be able to remember if you have just had the operation. There is a calendar next to your bed, and so you always know what day it is.[3]

Parfit asks you to imagine that you wake up in the hospital. You can see from the calendar that the day is Wednesday but cannot for the life of you remember whether you had an operation yesterday. He asks you to consider whether you would prefer to have already had

[1] Derek Parfit, *Reasons and Persons*, reprinted with further corrections (1984; repr., Oxford: Clarendon, 1987), 167.

[2] Ibid., 3.

[3] Richard Kraut describes a similar case in "The Rationality of Prudence," *Philosophical Review* 81 (1972): 351–59, at 355.

the longer operation rather than to be about to have the shorter operation. Parfit claims that most of us would prefer to have had the early operation and that this preference is rationally permissible.

I share Parfit's intuitions about this case and his intuition that time-bias is rationally permissible. However, I have come to doubt that these intuitions are well founded. I will explain why.

II. TIME-BIASED PREFERENCES: INCONSISTENT OVER TIME!

There is an unsettling feature of being time-biased: it leads to having different preferences at different times. Suppose you are time-biased in Operations 1. On Monday, you would prefer to have the late operation rather than the early operation because both operations are in the future and the late operation is shorter than the early one. However, on Wednesday, the early operation is in the past, and so you would prefer to have had the early operation rather than to be about to have the late operation. In virtue of being time-biased, you would change your preferences about the early and late operations. Your preferences at different times would be inconsistent with each other. As economists would say, your preferences would be "time-inconsistent."

I imagine that this is not a novel observation. It may not have been one that has particularly worried people. Understandably so. Plausibly, the reason why an inconsistency in preferences is a rational defect is that this inconsistency will lead to problems when acting. Take a case of intransitive preferences. If Jones prefers chocolate ice cream to strawberry, strawberry to vanilla, but vanilla to chocolate, then she will be a dreadful bore when ordering dessert. Moreover, if Jones's preference for, say, strawberry over vanilla is strong enough that she would swap a pint of vanilla and some small amount of money for a pint of strawberry, then she can be turned into a "money pump." Suppose Jones has a pint of vanilla ice cream. An ice-cream trader announces that she will offer Jones the following series of trades. First, the trader will offer Jones a pint of strawberry in return for her pint of vanilla and any amount of money—Jones gets to name her price, as long as she pays something. Next, the trader will offer Jones a pint of chocolate in return for the pint of strawberry. Finally, the trader will offer back to Jones the original pint of vanilla in return for the pint of chocolate. Jones can see that the entire series of deals will leave her worse off monetarily and better off in no respect. Still, Jones will take each individual deal because this deal best satisfies her preferences. Just as someone draws water from a water pump, the trader draws money from Jones: she is a money pump. Jones's vulnerability does not seem indicative of rational preferences.

However, in contrast to ice-cream preferences, it would seem that time-biased preferences will not lead to problems when acting. This is because you cannot change whether an experience is past or future. Suppose you wake up on Wednesday in Operations 1 and prefer to have already had the longer surgery rather than to be about to undergo the shorter surgery. Unfortunately for you, you cannot do anything to satisfy this preference, since you cannot control whether you have already had surgery or not. It would seem that the cross-temporal inconsistency in time-biased preferences is immune from being a practical defect.

Or so we might have thought. For, despite appearances otherwise, there are cases where time-biased preferences are action-guiding. I will present a case where a risk-averse person's time-biased preferences lead her to take certain types of insurance. In this way, these preferences will be pragmatically problematic in a way that is analogous to the way in which Jones's intransitive ice-cream preferences are problematic.

III. THE RISK AVERSE LIKE INSURANCE

The easiest way to introduce risk aversion is with an example. Suppose you have a lottery ticket that has an equal chance of paying out nothing or $20. It has an expected value of $10. If you are risk averse, then you would prefer to exchange this ticket for another ticket that has the same expected value but involves less risk. Ideally, you would like to eliminate the risk altogether by exchanging it for a ticket that could be directly redeemed for a crisp $10 bill. But you would also prefer to reduce the risk by, for example, exchanging the original ticket for another ticket that has an equal chance of paying out $9 or $11. An intuitive way to think about risk aversion is to think of the fifty-fifty gamble as having a "good" outcome of paying out $20 and a "bad" outcome of paying out nothing. If you are risk averse, then you would prefer to *reduce the gap* between the good and bad outcomes by worsening the good outcome to $11 and improving the bad outcome to $9. By hedging your bets, you would be insuring yourself against the bad outcome obtaining.

Indeed, some people are sufficiently risk averse that they are willing to lower the expected value of a gamble by a suitably small amount in order to reduce the gap between the good and bad outcomes. For example, someone may prefer to exchange the original ticket for another ticket that has an equal chance of paying out $8.50 or $10.50—this ticket has an expected value of $9.50. And some people like to buy fire insurance, even though they know that their insurance company is not a selfless guardian of the vulnerable but instead expects to profit from their buying this insurance. For simplicity, let us restrict ourselves to fifty-fifty gambles. Let us say that if you are "robustly risk averse," then you dislike risk enough that you would always prefer to reduce the amount of risk you face even if this means decreasing the expected value of the gamble by a suitably small amount:

> EVERY RISK REDUCTION HAS ITS PRICE: If a robustly risk-averse person faces a fifty-fifty gamble, then for any reduction of the gap between the good and bad outcomes of the gamble, there is some decrease of the gamble's expected value that this person would accept in return for this reduction of the gap.[4]

In intuitive terms, if you are robustly risk averse, then you are always willing to pay something to insure yourself, however little this amount you are willing to pay is.[5] A fortiori, you are always willing to insure yourself for free.

Let me illustrate robust risk aversion in terms of the lottery ticket example. The status quo is that you have the original ticket with a good outcome of $20 and a bad outcome of $0. The gap between the good and bad outcomes is $20. The expected value of the gamble is $10. The principle says that if you are robustly risk averse, then for any reduction in the gap, there is some amount you are willing to pay for this reduction. To take an arbitrary risk reduction, let us consider reducing the gap from $20 to $2. The principle says that there is some amount you would be willing to pay for this risk reduction. Let us suppose that this amount is $0.10. In this case, you would exchange the original ticket for another ticket that has a 50 percent chance of paying out $8.90 and a 50 percent chance of paying out $10.90. The gap between the gamble's good and bad outcomes is $2, and it has an expected value of $9.90. The reduction of the gap from $20 to $2 is worth to you the reduction of the expected value from $10 to $9.90.

I will only be concerned with robust risk aversion. With this in mind, from now on I will simply talk of "risk aversion" for concision.

[4] To be maximally precise about the order of the quantifiers, the principle says that for all people, (if someone is robustly risk averse, then [for any gamble G, if the gap between G's good and bad outcomes is x and G's expected value is v, then (for any gap y such that $y < x$, there is some amount of expected value w such that $w < v$, and this person would prefer a gamble, H, with gap y and expected value w, to G)]).

[5] In characterizing risk aversion in this intuitive way, I am not taking a stand on the appropriate decision-theoretic characterization of risk aversion. The most common way of doing so is to model risk-averse people as having diminishing marginal utility for the goods in question. An alternative approach is Lara Buchak's. Buchak models risk-averse people as having utility functions that are directly sensitive to risk. Lara Buchak, "Risk Aversion and Rationality," unpublished manuscript available at http://philosophy.berkeley.edu/people/files/208 (as of December 17, 2010).

Many of us think that risk aversion is rationally permissible. I will not defend this claim here, and I will return to the issue of the rational permissibility of risk aversion in my conclusion. For our purposes, we should note that it is prima facie plausible that risk aversion is rationally permissible. As such, it would be an interesting result if it turns out that the combination of risk aversion and time-bias is irrational. Now it is time to see the case for its being so.

IV. HOW TO TURN THE RISK AVERSE AND TIME-BIASED INTO "PAIN PUMPS"

Earlier we saw that if someone has intransitive preferences, she can be turned into a "money pump"—she would take a series of deals that leaves her financially worse off and better off in no respect. The inconsistency that arises from time-bias makes a risk-averse person vulnerable in an analogous way. She would take a series of pills that leave her with more pain and better off in no respect. She would take this series even if she knew throughout that she is being offered it. We might say that she is being turned into a "pain pump"—while money can be pumped out of someone who is a money pump, pain could be pumped into someone who is a pain pump.

Let us assume throughout our discussion that you are risk averse and time-biased. Next, we will need to modify our earlier case, Operations 1, so that you have one of two courses of surgery. Let us add an extra hour's surgery on Thursday to both courses as follows:

> Operations 2: On Monday, you are admitted into a hospital. You are told you will have one of two courses of operations, but you are not told which. If you have the early course, then you will have a painful, four-hour operation on Tuesday and a painful, one-hour operation on Thursday. If you have the late course, then you will have a painful, three-hour operation on Thursday. After any operation, you will have amnesia for several days, and so you will not be able to remember if you have just had an operation. There is a calendar next to your bed, and so you always know what day it is.

On Monday, you are facing a fifty-fifty gamble. You judge that it is equally likely that you will have the early course or the late course. The outcomes of this gamble vary according to how much pain you will experience. The good outcome is that you have three hours' pain on Thursday. The bad outcome is that you have four hours' pain on Tuesday and one hour's pain on Thursday—a total of five hours' pain. The gap between these outcomes is two hours' pain.

None of us would be enthusiastic about the prospect of future surgery. But you would be especially anxious because you are risk averse. You are facing a gamble that exposes you to risk. You would prefer to reduce the amount of risk you face by shortening the gap between the good and bad outcomes.

Imagine you were offered a pill that did just this! The pill has two possible effects. If you undergo the Thursday operation of the early course, then the pill works as an anesthetic and reduces the amount of pain you experience in this operation by just less than thirty minutes. But if you undergo the Thursday operation of the late course, it has no anesthetic effect and instead has a nasty side effect: your pain continues after this surgery for a little more than thirty minutes. This pill would decrease the gap between the good and bad outcomes. Let us suppose that it reduces this gap by sixty minutes. We know from Every Risk Reduction Has Its Price that a risk-averse person is willing to pay for insurance to reduce this gap; that is, there is some increase in expected pain that a risk-averse person would accept in order to reduce this gap by sixty minutes. It will help simplify our discussion by picking a fixed increase that you are willing to accept: let us arbitrarily suppose that you are willing to accept one additional minute of expected pain. This supposition will make our discussion run more smoothly, and we are making it without any loss of generality.[6] So let us suppose that you are offered a pill that reduces the gap by sixty minutes but increases your expected amount of pain by one minute. This pill has the following effects:

> Help Early: If you will have the early course, then the pill decreases the length of the pain you experience on Thursday by just less than thirty minutes: it reduces it by twenty-nine minutes. If you will have the late course, then the pill increases the length of the pain you experience on Thursday by just more than thirty minutes: it increases it by thirty-one minutes.

Let us assume that this pill and all the other pills we come across later in this essay are effective only if they are taken on the day that they are offered. (The pharmaceutical company is notorious for the appallingly short shelf life of its products.) On this assumption, this pill would be a form of insurance that, on account of your risk aversion, you would like to take.

Now suppose that you wake up on Wednesday knowing the date but not knowing whether you have just had an operation. If you are time-biased, then you face another gamble. This gamble varies both

[6] The rest of our discussion could easily be altered so that “one” is replaced with any small numeral, or indeed a general placeholder for a small numeral, “Δ.”

in terms of how much pain you feel and when you experience it. The good outcome is that you have just had the four-hour operation and still have a short operation to come. The bad outcome is that a long operation is still to come. If you are risk averse, then you would like to make the good outcome worse in order to make the bad outcome better.

What luck: you are offered a second pill that does just this! Its effects counteract the effects of the first pill. If you undergo the Thursday operation of the early course, then the pill has no anesthetic effect and, instead, has a nasty side effect: your pain continues after the surgery. But if you undergo the Thursday operation of the late course, it works as an anesthetic and reduces the amount of pain you experience in the operation. The pill has the following effects:

> Help Late: If you are having the early course, then the pill increases the length of the pain you experience on Thursday by thirty minutes. If you are having the late course, then the pill decreases the length of the pain you experience on Thursday by thirty minutes.

If you are risk averse and time-biased, then you would take Help Late. Why so? Taking it does not affect your expected amount of pain. But it reduces the risk you face: it reduces the gap between the bad outcome and the good outcome. Moreover, it does so without affecting the expected amount of pain you will experience: it is a form of insurance for free! We know from Every Risk Reduction Has Its Price that a risk-averse person would always pay something for insurance. A fortiori, she would always take insurance for free.

What happens if you take both pills? You are sure to have one minute's more pain on Thursday and gain nothing. I will represent the additional minutes of pain you would have by taking these pills in the following table:

	Effect of Help Early	Effect of Help Late	Effect of Both Pills
You have the early course	−29	30	1
You have the late course	31	−30	1

Regardless of whether you have the early course or the late course, the net effect of taking both pills would be that you experience one more minute of pain on Thursday. In virtue of being risk averse and time-biased, you would take both pills, ending up with more pain and better off in no respect; by doing so, you would be a pain pump.

V. WHY STRATEGIES TO REFUSE PILLS FAIL

I presented the case for why a risk-averse and time-biased person would take these pills rather quickly. It may help to illuminate it by considering more carefully the question of whether you can be both risk averse and time-biased and yet avoid being a pain pump. Unfortunately, the answer is that you cannot. Avoiding being a pump would involve refusing at least one pill. But we will see that if you are risk averse and time-biased, then you would not do so. Instead, it turns out that taking each pill is the "strictly dominant" option: you would prefer to take it whether or not you take the other pill.

A. Why a Strategy to Refuse Both Pills Fails

In light of the sure loss that you would suffer if you took both pills, we might think that you should reason, "if I take both pills, then I will have a little extra pain and gain nothing. If I refuse both pills, then I lose nothing. Therefore, I should refuse both pills."

If you had to make one single decision about which pills, if any, to take, then this strategy would work. The problem is that you never make a single decision to reject both pills or take both pills. Instead, you make two independent decisions—one on Monday and one on Wednesday. At the point at which you make each decision, you prefer taking the pill you are currently offered and rejecting the other to rejecting both.

To see this, let us consider your decision about Help Late. Suppose that you have refused Help Early. You should reason as follows: "I face a gamble. Because I am risk averse, I would like to insure myself by reducing the gap between the gamble's good and bad outcomes. Taking the current pill would let me reduce the gap with no change in my expected amount of pain. Therefore, I ought to take this pill." Thus, a strategy to refuse both pills fails because if you have refused the earlier pill, you would prefer to take the later pill.

In passing, let us notice another point. You should use similar reasoning when deliberating about whether to take Help Early on the assumption that you will later refuse Help Late. Suppose you will refuse Help Late. You should reason, "I face a gamble. Because I am risk averse, I would like to insure myself by reducing the gap between the gamble's good and bad outcomes. I would like to reduce the gap by an hour, even if this would involve a suitably small increase in my expected amount of pain. Taking the current pill would let me reduce the gap for a suitably small increase. Therefore, I ought to take the pill." For the sake of exposition, I assumed earlier that the "suitably small increase" in expected pain that you were prepared to accept was one minute. In light of this, I stipulated that Help Early would reduce the gap by an hour and increase your expected pain by one minute.

But this was an arbitrary choice. We know from Every Risk Reduction Has Its Price that, being risk averse, there is some amount that you are willing to accept in order to reduce the gap by an hour. We can stipulate that Help Early will reduce the gap by an hour and increase your expected pain by this amount.

So, whichever pill you are currently deciding whether to take, you would prefer to take this pill if you refuse the other pill. We will return to this point shortly.

B. Why a Strategy of Refusing Just One Pill Fails

We just saw that a "refuse both pills" strategy fails. It does so because the decision about each pill is independent from the decision about the other pill. Interestingly, a "refuse exactly one pill" strategy fails for the very same reason. To execute this strategy, you would have to refuse one of the pills while taking the other. But at the time at which you are offered each pill, you prefer taking this pill and taking the other pill to refusing this pill and taking the other pill. Let us see why.

Let us start by supposing that you did take Help Early. On Wednesday, you should reason, "There is a good outcome in which I have the early course and a bad outcome in which I have the late course. By taking Help Early two days ago, I made the bad outcome worse and the good outcome better. . . . I lengthened the gap! Since I am risk averse, I now prefer that I had not done this. Too bad. Still, I would like to reduce this gap. By taking Help Late, I could. This pill would be a form of insurance that does not affect my expected amount of pain."[7]

Indeed, parallel reasoning applies to your decision about whether to take Help Early on the assumption that you will take Help Late. You should reason, "The good outcome is that I have the late course, and the bad outcome is that I have the early course. What a shame that I will accept Help Late and lengthen the gap between these outcomes! Still, I can reduce this gap by taking Help Early. It would slightly increase my expected amount of pain. But the reduction in risk would be worth it. So I ought to take Help Early." Again, we know from Every Risk Reduction Has Its Price that, being risk averse, there is some amount that you are willing to accept in order to reduce the gap by an hour. We can stipulate that Help Early will reduce the gap by an hour and increase your expected amount of pain by this amount.[8]

[7] Thanks to Caspar Hare for making this point.

[8] Indeed, we could even allow that the amount of extra pain that you would accept to shorten the gap by an hour varies according to whether or not you later take Help Late. We can see this by looking at a general recipe for the pain pump. Let Help Early (x) be a pill that increases your Thursday pain by $(30 + x)$ minutes if you have the late course and decreases your pain by $(30 - x)$ if you have the early course. We can show

So we know that a “take just one pill” strategy will not work. This is because you prefer to take each pill if you have taken the other pill. Recall that in the previous subsection (V.A), we saw that you would take each pill if you have refused the other pill. Thus, we can conclude that you would take each pill whether or not you take the other pill. In other words, taking each pill is the strictly dominant option.

C. The Significance of the Independence of Both Deals

We can summarize what we have observed by considering the four possible outcomes of your choices on Monday and Wednesday:

	Your Monday Self	
	You Take Help Early	You Refuse Help Early
Your Wednesday self:		
You take Help Late	A	B
You refuse Help Late	C	D

In Section V.A, we saw that on Monday you rank C > D and on Wednesday you rank B > D. In Section V.B, we saw that on Monday you rank A > B and on Wednesday you rank A > C. At all times, you rank D > A. This is because A is just like D but for the fact that you have one minute’s more pain in A. Therefore, on Monday you rank these outcomes C > D > A > B, and on Wednesday you rank these outcomes B > D > A > C.[9]

that if you are risk averse and time-biased, then there is some Δ such that you would prefer to take Help Early (Δ) whether or not you later accept Help Late.

We will assume throughout that you are risk averse and time-biased. Note that it follows from Every Risk Reduction Has Its Price that there is some number of minutes, Δ_1, such that if you will accept Help Late, then you prefer to take Help Early (Δ_1). This would improve the bad outcome from five hours and thirty minutes of future pain to five hours and Δ_1 minutes. And it would worsen the good outcome from two hours and thirty minutes of pain to three hours and Δ_1 minutes. Since you are risk averse and time-biased, we know from Every Risk Reduction Has Its Price that there is some number of minutes, Δ_1, such that you would accept Δ_1 more minutes of expected pain in order to reduce the gap between the good and bad outcomes in this way.

Similarly, it follows from Every Risk Reduction Has Its Price that there is some number of minutes, Δ_2, such that if you will refuse Help Late, then you prefer to take Help Early (Δ_2). This would improve the bad outcome from five hours of future pain to four hours and ($30 + \Delta_2$) minutes and worsen the good outcome from three hours of future pain to three hours and ($30 + \Delta_2$) minutes. Again, since you are risk averse, you would accept some Δ_2 extra minutes of pain to reduce the gap in this way.

Finally, let Δ be the minimum of Δ_1 and Δ_2. It follows that you would prefer to take Help Early (Δ) whether or not you later accept Help Late.

[9] Game theorists may be interested to note that these are, of course, the preferences that characterize the generalized form of the classic cooperation problem of a so-called prisoner’s dilemma.

In light of this, if you have taken Help Early, we should resist the temptation to reason with you, "If you take Help Late, then you will end up with a minute's more pain than you would have had in your original position before any pills are offered to you. Why is it not perfectly rational to refuse Help Late in order to avoid this loss?"[10] With reference to the matrix above, we can see that this would be to reason with you, "You know that if you take Help Late, then you will end up in A where you have a minute's more pain than you would have in D. Why is it not perfectly rational to refuse Help Late in order to avoid this loss?" You should reply, "Unfortunately, I cannot bring about A by refusing Help Late. I am not choosing between A and D, and so my preference for A over D is irrelevant. On Monday, I took Help Early, and so I am now choosing between C and D. I prefer D to C. Therefore, I should take Help Late." This is the significance of the fact that you make several independent choices and you are never able to make a single choice between A and D.

Similarly, we should be hesitant to advise you, "You should set aside your risk-averse tendencies and let it ride this one time by refusing Help Late in order to avoid a little extra pain." This advice might be meant in one of two ways. First, it might be meant as advice not to act on your risk-averse preferences in order to avoid an overall loss. Construed this way, we would be offering you bad advice. It is advice to take C rather than D, even though you prefer D to C simply because you prefer A to D.[11] If you followed this advice, you would be acting contrary to your preferences. Second, we might be advising you not to be risk averse. On this reading, this advice would not pose an objection to the pain pump argument. The argument only concludes that you would be a pain pump on condition that you are risk averse and time-biased. Maybe the moral to draw from the pain pump argument is that you should not be risk averse. Next I will discuss whether this is so.

[10] Thanks to an anonymous referee for highlighting the importance of responding to this objection and the related one that follows shortly in the text.

[11] Similarly, when Jones has traded a pint of vanilla and a nickel for a pint of strawberry and has traded this pint of strawberry for a pint of chocolate, we should not advise her to refuse trading this pint of chocolate for the original pint of vanilla on the grounds that this would lead her to make a loss of a nickel overall. Jones should respond that she is not currently able to decide whether to have a pint of vanilla or a pint of vanilla and a loss of a nickel. Instead, she is only able to decide whether to have a pint of chocolate or a pint of vanilla. She prefers the pint of vanilla to the pint of chocolate, and so she ought to make the trade.

VI. THE PUZZLE THAT WE ARE LEFT WITH

The argument that I have offered concludes that a risk-averse person with time-bias could be turned into a pain pump. This causes trouble because it means that we cannot accept both that you are rationally forbidden from having preferences that allow you to be turned into a pain pump and that you are rationally permitted to be time-biased and risk averse.

What to do? In effect, we have four options. (1) We could deny that you are rationally forbidden from having preferences that allow you to be turned into a pain pump. Alternatively, we could deny that you are permitted to be time-biased and risk averse. Notice that there are several ways in which we could deny this: (2) we could claim that separately time-bias and risk aversion are rationally permissible but the combination of the two is impermissible,[12] (3) we could deny that risk aversion is rationally permissible, or (4) we could deny that time-bias is rationally permissible. The main conclusion of this essay is that we have to embrace one of these options. I see no decisive grounds for picking one of them. In what remains, I will briefly consider reasons for taking each option and point out some further issues that would need to be addressed if we did.

A. Can Pain Pumps Be Rational?

Could we reject the claim that you are rationally forbidden from having preferences that allow you to be turned into a pain pump? Let us consider two reasons why we might reject this claim.

First, we might hold that the argument merely shows that agents who are unable to coordinate decisions at different times can be exploited.[13] Suppose you were able to bind yourself into making future decisions in certain ways. For example, on Monday you could bind yourself into refusing Help Late on Wednesday. Then you could avoid being turned into a pain pump: you could make a single decision to refuse both pills, for example.[14]

[12] Thanks to an anonymous referee for pointing out this option, which I had overlooked.

[13] Recall that in n. 9 we saw that a time-biased and risk-averse person's preferences at different times characterize a prisoner's dilemma. Part of the reason why different agents do not reach the collectively optimal outcome in this game is that they are unable to coordinate their actions. This might suggest that coordination is the root of the present problem.

[14] This response may get independent motivation from Frank Arntzenius, Adam Elga, and John Hawthorne's argument that certain decision-theoretic puzzles involving infinity show that "rational individuals who lack the capacity to bind themselves are liable to be punished, not for their irrationality, but for their inability to self bind." Still, it is not clear how general a moral we can draw from puzzles that involve infinity. Frank

However, if the self-binding response is correct, then it is a powerful one.[15] It would follow that a money pump argument cannot show that intransitive preferences are irrational. For example, ice-cream-loving Jones could avoid being turned into a money pump if she could bind her future decisions about the relevant ice cream trades. Denying that money pump arguments show intransitive preferences are irrational would arguably fly in the face of orthodoxy, but maybe orthodoxy is wrong.

Second, we might claim that a money pump argument can show that intransitive preferences are irrational while the pain pump argument does not show that time-biased and risk-averse preferences are irrational. We could do so by pressing on a disanalogy between the arguments. The intransitivity argument targets synchronic preferences —preferences someone has at a particular time. We might say that a "time-slice" of the agent has these preferences. But the time-bias argument targets diachronic preferences—at different times, a time-biased agent has conflicting preferences. The preferences of her time-slices are inconsistent with each other. Perhaps there are only rational constraints on our synchronic preferences and not our diachronic preferences.[16] After all, an agent can predictably change her preferences without this being irrational. For example, someone could predictably have liberal political preferences as a youth and conservative political preferences in middle age.

There is an important difference between changing political preferences and time-biased preferences, though. When someone changes her political preferences, she no longer endorses her earlier preferences. I suggest that this is why this change does not seem irrational. At no point does she endorse both preferences. By contrast, a time-biased person endorses the conflicting preferences that she has at different times: being time-biased, she will judge that these are the appropriate preferences to have. Now, there is some plausibility to thinking that there are diachronic rational constraints on the pref-

Arntzenius, Adam Elga, and John Hawthorne, "Bayesianism, Infinite Decisions, and Binding," *Mind* 113 (2004): 251–83, at 269. For criticism of their view, see Christopher Meacham, "Binding and Its Consequences," *Philosophical Studies* 149 (2010): 49–71. An anonymous referee made the interesting suggestion that we consider whether money pumps could be generated in light of similar puzzles described in the "Dynamic Epistemic Paradoxes" section of Roy Sorensen, "Epistemic Paradoxes," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Stanford, CA: Stanford University Press, 2009), http://plato.stanford.edu/archives/spr2009/entries/epistemic-paradoxes/.

[15] Christopher Meacham notes that this response also has surprising implications for several other issues in decision theory (Meacham, "Binding and Its Consequences").

[16] Thanks to Agustín Rayo for suggesting this point.

erences you endorse throughout.[17] As Lara Buchak nicely puts it, the intuitive idea would be that an "agent who endorses the preferences of her time-slices ought to act like a unified entity over time."[18]

This response rests on our seeing ourselves as agents who persist over time. It could be that this view is wrong. It may be that different time-slices are really different agents (admittedly different agents who stand in a special relationship to each other). This position strikes me as implausible, but to fully evaluate it, we would have to take a stand on some controversial and complex issues in the metaphysics and ethics of personal identity. This is not the place to do so. Still, we should concede that the more we think of different time-slices as different agents, the less inclined we should be to think that there are diachronic rational constraints.[19] If different time-slices are different agents, then these constraints would be constraints on agents to harmonize their preferences with the preferences of other agents. But failing to harmonize with other agents' preferences does not seem a rational defect in an agent.

B. Is the Combination *of Time-Bias and Risk Aversion Irrational?*

Could we claim that the combination of time-bias and risk aversion is rationally impermissible, although each is permissible by itself? Certainly, there are examples of some preferences that are separately permissible but jointly impermissible. Intransitive preferences are like this: preferring chocolate to strawberry is permissible, preferring strawberry to vanilla is permissible, and preferring vanilla to chocolate is permissible. However, it is rationally impermissible to have all three preferences at the same time.

The advantage of this option is that it would not require giving up any strong intuitions. Apart from an intuition that time-bias and risk aversion are separately rationally permissible, I suspect that we have little intuitive purchase on the permissibility of the combination of the two. As a default, we might expect the combination to be rationally permissible, but I doubt we have strong feelings about this. So this option is attractive in that it comes with little intuitive cost.

That said, our lack of intuitive purchase concerning the combi-

[17] Here I am drawing on an interesting suggestion of David Christensen's in the epistemology literature. Christensen discusses whether there are diachronic rational constraints on our credences at different times. He suggests that the most promising rationale for thinking that there are such constraints is by appealing to the credences we endorse having over time. David Christensen, "Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers," *Journal of Philosophy* 93 (1996): 450–79.

[18] Buchak made this point in her comments on this essay at the Bellingham Summer Philosophy Conference 2010, Bellingham, WA.

[19] This was inspired by a similar point in Buchak's aforementioned comments.

nation is also this option's drawback. In the case of the ice-cream preferences, we already have an intuition that intransitivity is a rational defect. But we lack any similar intuition about the combination of time-bias and risk aversion. As a result, this option may look ad hoc unless we could give it an independent motivation, and I cannot see what this motivation could be. What is special about the combination of risk aversion and time-bias?

C. Is Risk Aversion Irrational?

Could we deny that risk aversion is rationally permissible? Recall that the argument only concerns robust risk aversion, which is a willingness always to pay something, however little, to insure oneself against risk. There is a broader debate about which forms of risk aversion, if any, are rationally permissible.[20] As things stand, I think that this debate has not run its course. Still, without resolving this broader debate, we can stake out some available options.

We might hold that robust risk aversion is rationally impermissible, but a more minimal risk aversion is permissible. Perhaps you are rationally permitted only sometimes to be willing to pay to reduce the amount of risk you face. To make good on this option, we would like to know what principle would govern when it is permissible to so pay. And to avoid this principle seeming ad hoc, we would need to motivate it.

Alternatively, we might find this middle ground untenable and retreat to the claim that only a minimal risk aversion is rationally permissible: you are only permitted to insure yourself against risk when you do not have to pay for this insurance. This would be surprising. It would mean that we are acting irrationally when we buy fire insurance from companies that expect to make a profit from our doing so.

D. Is Time-Bias Irrational?

Finally, could we reject the claim that time-bias is rationally permissible? Somewhat hesitantly, this is the response that I find myself driven to. If you are time-biased, then you have preferences at different times that are inconsistent with each other, even though you endorse them. This inconsistency leads to acting counterproductively. Moreover, I think of all of us as agents who persist over time, and so I think we should act cohesively when we act on preferences that we endorse. This suggests to me that time-bias is irrational.

The reason why I am hesitant about this conclusion is that it is

[20] For a more thorough defense of the rational permissibility of a particular decision-theoretic account of risk aversion, see Buchak, "Risk Aversion and Rationality."

highly counterintuitive: are you really rationally required to be indifferent as to whether you are about to have root canal surgery in one minute's time or have already had it? I doubt any of us could actually stick to this preference as the dentist's drill approaches. Still, it is hard to say why it is rationally permissible to be time-biased. If we try to give a reason for why it is fine to prefer a fixed amount of pain to pass, it is tempting to say something circular like "well, it would be over and done with, then." But clearly this is circular. This leaves us to wonder whether there is an explanation of why time-bias is rationally permissible. If there is, what is it? It is partly because I cannot see any grounding for this claim that I can countenance the view that it is simply an unjustified and misleading intuition. Still, I would feel more comfortable taking this counterintuitive view if I had an error theory that explains why we have this misleading intuition. Sadly, what this explanation would be is beyond me.[21]

Instead, I will end by noting that there would be a silver lining to not being time-biased. Suppose that you are on your deathbed, having lived a long and fulfilling life. Parfit points out that if you were not time-biased, then you "should not be greatly troubled by the thought that [you] shall soon cease to exist, for though [you] now have nothing to look forward to, [you] have [your] whole [life] to look backward to."[22] Parfit's insight is that one of the reasons why you might fear death is that it would constitute the end of your future experiences. What is particularly distressing is the thought that you only have a few more days left to live. But this is simply a manifestation of time-bias—it is a worry about what lies in the future. If you were not time-biased, then you would not have this concern. You would be more at peace with your mortality.

[21] Daniel Hausman speculated to me that we might have this intuition because the memory and anticipation of pain have different phenomenologies: anticipation makes us anxious. It may be that we mistake the rationality of reducing this anxiety for the rationality of time-bias. This is an interesting suggestion, but the case would have to be made for why the anxiety is not simply symptomatic of time-bias but rather is essentially linked to anticipation in a way that it is not linked to memory. One possible explanation would be that we are typically uncertain about future pains, and another would be that we do not remember past pains well. But speaking from my own case, I am no less anxious when I am certain that future pains will occur, and I think that I am equally able to imagine past pains as I am future pains.

[22] Parfit, *Reasons and Persons*, 176. For discussion of other ways in which death can have disvalue, see, e.g., Kai Draper, "Disappointment, Sadness, and Death," *Philosophical Review* 108 (1999): 387–414.