THOUGHT
A JOURNAL OF PHILOSOPHY
THE NORTHERN INSTITUTE OF PHILOSOPHY

Thought ISSN 2161-2234

ORIGINAL ARTICLE

A Deluxe Money Pump

Tom Dougherty

University of Sydney

So-called money pump arguments aim to show that intransitive preferences are irrational because they will lead someone to accept a series of deals that leaves his/her financially worse off and better off in no respect. A common response to these arguments is the foresight response, which counters that the agent in question may see the exploitation coming, and refuse to trade at all. To obviate this response, I offer a "deluxe money pump argument" that applies dominance reasoning to a modified money pump case.

Keywords rationality; transitivity; preferences; money pump; rational choice; cyclic

DOI:10.1002/tht3.91

1 The standard money pump argument and the foresight response

Intransitive preferences seem irrational. For example, Jones seems irrational for preferring strawberry ice cream to vanilla, chocolate to strawberry, and vanilla to chocolate. Some of us think that the irrationality of Jones's preferences can be demonstrated by a so-called money pump argument.[1] Suppose Jones starts with a pint of vanilla ice cream. Mr Whippee announces the following series of trades. First he will offer Jones a pint of strawberry for the pint of vanilla and any amount of money—Jones gets to name her price, so long as she pays something. He notes that if Jones's preference for strawberry over vanilla is non-negligible, then she will prefer to make this trade. He then announces that if she accepts the first trade, then he will offer her a pint of chocolate for the pint of strawberry. Again, he points out that taking this trade will satisfy her preference for chocolate over strawberry. Finally, he announces that if she accepts the second trade, then he will offer her back the original pint of vanilla for her pint of chocolate. Once more, he points out that taking this trade will satisfy her preference for vanilla over chocolate. If Jones acts on her ice-cream preferences, then she will accept each deal, leaving her with her original pint of vanilla, but less money than she began with. But things get worse. Mr Whippee announces he will iterate this cycle of trades.[2] If Mr Whippee offers the series a sufficiently large, but finite, number of times Jones would end up losing *all* of her money by acting on her intransitive preferences. She would become a "money pump" for Mr Whippee: just as one can draw water from a water pump, so Mr Whippee could draw money from Jones, despite telling her exactly what he is doing. Call this version of the case the "standard money pump case." We should not be distracted by the fact that Mr Whippee is exploiting Jones. We might instead assume that Mr Whippee is a kindly old

Correspondence to: E-mail: tom.dou@gmail.com

gent, whose goal is to ensure people get the ice cream they want. (That's why he lets his patrons decide how small a fee they pay for his services, insisting on something merely as a token gesture.) All the same, Jones would make a series of trades that leave her worse off financially and better off in no respect. This behavior seems irrational, and appears to indicate that her underlying preferences are irrational.

A common response to the standard money pump argument is to argue that Jones would foresee the sure loss that would come from accepting the trades, and hence would refuse to make any trades.[3] Call this the "foresight response." According to the foresight response, at square one Jones should reason, "If I refuse the first trade, then I will have my pint of vanilla and all my money. If I accept, then eventually I will have my pint of vanilla and no money at all. Therefore, I should refuse the initial trade." As Frederic Schick puts it, "seeing what is in store for [her], [she] may well reject the offer and thus stop the pump . . . [She] need not act as if [she] wore blinders . . . [She] may see [she] is being pumped and refuse to pay for any further deals."[4] But if foresight allows Jones to avoid irrational behavior, then we cannot claim that her intransitive preferences are irrational on the grounds that they lead to such behavior.

2 Rabinowicz's backward induction counter to the foresight response

The foresight response depends on the assumption that if Jones declines the initial trade, then she gets to stick with the *status quo*. It is this feature of the case that allows her to reason, "if I refuse the first deal, then I will end up with the *status quo*, which I prefer to a sure loss." If the case is amended so this feature is absent, then the foresight response fails.

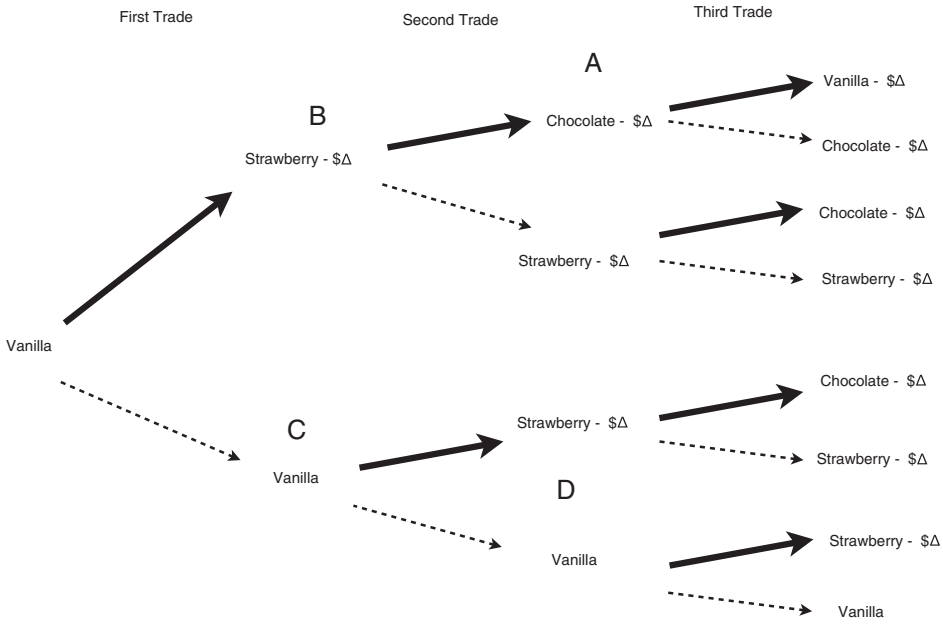
Wlodek Rabinowicz has proposed just such an amendment. His modification is elegantly simple. He supposes that the person offering the trades is persistent, by repeating an offer if it is refused. We can think of his case as follows. Suppose Mr. Whippee announces he will offer exactly three trades, and the terms of each offer depends on what Jones possesses:

(**a**) If Jones has a pint of vanilla, then Mr. Whippee offers a pint of strawberry for the pint of vanilla and any amount of money.
(**b**) If Jones has a pint of strawberry, then Mr. Whippee offers a pint of chocolate for the pint of strawberry.
(**c**) If Jones has a pint of chocolate, then Mr. Whippee offers Jones a pint of vanilla in return for this pint of chocolate.

Since this is an improvement on the standard money pump case, call it the "premium money pump case." Crucially, (a) and (b) guide Mr. Whippee's offers in the second and third trades. So, if Jones refuses the first deal and retains her pint of vanilla, then Mr Whippee offers the very same deal again. Because she might accept later trades, Jones does not ensure that she eventually ends up with the *status quo* by refusing the first trade.

Indeed, Rabinowicz shows that Jones will accept all three trades if she uses the method of backward induction. This method requires Jones to consider her final

decision, and then work backwards to her initial decision. The easiest way to illustrate this reasoning is by means of a decision-tree diagram:



In this diagram, upward and downward sloping lines lead to the consequences of accepting and rejecting each trade, respectively. Thick unbroken lines indicate choices that satisfy Jones's preferences, and thin dashed lines indicate choices that don't. The fact that all and only the upward sloping lines are thick unbroken lines represents the substantive claim that for any trade, accepting this trade best satisfies Jones's preferences at that time.

This claim is justified by Rabinowicz's backward inductive reasoning. This reasoning starts with the third choice. There are four possible scenarios that Jones could be in. To see this, let us start with the top scenario, labeled "A." Here Jones chooses between [Vanilla − $Δ] and [Chocolate − $Δ]. Because Jones prefers vanilla to chocolate, irrespective of her other holdings, she prefers [Vanilla − $Δ] to [Chocolate − $Δ], and so will accept the third trade. Similar reasoning tells in favor of Jones accepting the third trade in each of the other scenarios.

The second step in the backward inductive reasoning is to consider Jones's choice about the second trade. There are two possible scenarios. Let us consider the first scenario, labeled "B," in which she holds [Strawberry − $Δ]. Here we make use of the result that Jones can predict that she will always accept the third trade. Consequently, Jones can focus only on her preferred final outcomes of the third trade, which are reached by upward sloping lines. So at B if she accepts the second trade, then she will end up with [Vanilla − $Δ]; but if she rejects the second trade, then she ends up with [Chocolate − $Δ]. Since she prefers [Vanilla − $Δ] to [Chocolate − $Δ], she will prefer to accept the second trade at B. Again, similar reasoning tells in favor of Jones accepting the second trade in the other scenario.

The third step in the backward reasoning is to consider Jones's initial choice. Since Jones predicts that she will accept the second and third trades, she reasons that if she accepts the first trade, then she will end up with [Vanilla − Δ] and if she rejects the first trade, then she will end up with [Chocolate − Δ]. Again, she prefers the former outcome and so will accept the initial trade.

Rabinowicz's proof compellingly shows that in his case an intransitive agent would make a sure loss if she uses backward induction. However, his proof is vulnerable to a general challenge to backward induction. In the literature, the classical statement of this challenge begins with the claim that the method requires the agent to believe of any future choice that she will make this choice rationally.[5] Consequently, the method requires an agent to believe she will choose rationally in the future, even when she has chosen irrationally in the past. Yet it would seem that these past irrational choices can undermine her trust in her future rationality.

Rabinowicz claims that this challenge does not apply to his case:

> The objection in question does not apply to the short money pump described . . . There, it is only the final choice node that cannot be reached without violation of backward induction. But what is rational at the final choice node does not depend on what the agent expects to do in the future. On the other hand, at the non-final choice nodes in this problem, the agent lacks evidence about prior violations. (p. 138)

But Rabinowicz has made a slip in claiming that "only the final choice node . . . cannot be reached without violation of backward induction." To see this, consider the scenario labeled "C" on the diagram. Backward induction requires Jones to accept the first trade. However, C can only be reached by rejecting the first trade. Therefore, C is a "non-final choice node" at which the agent has "evidence about prior violations" of backwards induction. According to the challenge, this entails that Jones is not entitled to predict at C that she will act rationally after C.

It is open for Rabinowicz to counter that it is only at the third choice node that the agent can have evidence about *repeated* irrational behavior in the past, and that it is repeated irrational behavior that tends to undermine that agent's trust in her future rationality.[6] Indeed, while conceding that an agent may gather this evidence in longer sequences of trades, Rabinowicz states that his short three-round sequence may often be a special case for the following reason:

> . . . if a money pump is not too long, and if the sophisticated agent starts out with a firm conviction about his commitment to the backward-induction procedure, the evidence about his deviant past behaviour might never be extensive enough to shatter his initial conviction. He will be able to explain away his past deviations from the backward-induction path as isolated mistakes that would not recur in the future.

However, we should note the conditional nature of this response means that it is not fully general: it does not apply in the case of agents who have only a weak conviction

in their future rationality, or who have independent reasons for thinking that "isolated mistakes" are unlikely. It may take awareness of only one past rational mistake to undermine these agents' faith in their future rationality. And at the aforementioned scenario, C, such an agent will have gained this awareness. Therefore, if Rabinowicz relies on this less-than-general response, then the power of his argument is correspondingly weakened.

Moreover, the problem becomes more pressing in light of the fact that we can formulate the challenge to backwards induction in a simpler way than the way it is traditionally formulated.[7] Note that the backwards induction method requires the agent to predict *at the beginning* that he/she will rationally choose to satisfy his/her preferences in *every* scenario. Consider the scenario labeled "D" on the diagram. This scenario is reached by two repeated irrational choices. Now an agent who is otherwise reasonably confident in her rationality may judge that it would be quite a coincidence to have two repeated "isolated mistakes." Thus, such an agent may predict that if D obtains, then the best explanation is that he/she is partially disposed to act irrationally. Consequently, he/she may fail to predict that he/she will choose rationally at D. And yet the backwards induction method requires the agent to make this prediction.[8]

I do not claim that this is the final word on the matter, as there is a further debate about when, if ever, this challenge successfully undermines backward inductive reasoning.[9] However, while this debate continues, we should be interested in whether other methods can counter the foresight response. I will argue that there is one—dominance reasoning.

3 The dominance counter to the foresight response

Like Rabinowicz, I will offer a modified case that prevents Jones from securing the *status quo* by refusing the first trade. Suppose that instead of only having a pint of chocolate, Jones also has a pint of strawberry and a pint of vanilla. Let us assume that regardless of how much ice cream Jones has, she prefers a pint of strawberry to a pint of vanilla and some money $\$\Delta$, she prefers a pint of chocolate to a pint of strawberry, and she prefers a pint of vanilla to a pint of chocolate. (To anticipate, this assumption will limit the scope of my argument.) Mr Whippee proposes the same series of trades as before. First, he will offer to trade a pint of strawberry for a pint of vanilla. Second, he will offer to trade a pint of chocolate for strawberry. Third, he will offer to trade a pint of vanilla for a pint of chocolate. Call this the "deluxe money pump case."

For ease of exposition, let us start with Jones's last decision about whether to accept the third trade of vanilla for chocolate. Jones prefers to accept this trade, *whatever decision she has previously made regarding the first or second trades*. Accepting is the dominant option. The reason is simple. She prefers vanilla to chocolate, she has a pint of chocolate, and she is offered a pint of vanilla in exchange for this pint. The next step in my argument is that exactly the same reasoning applies to her decision about whether to accept the *second* trade of chocolate for strawberry. Regardless of whether she has accepted the first trade and regardless of whether she *will* accept the third trade, she prefers to accept this

second trade. Again, the reason is child's play—she prefers chocolate to strawberry. The temporal ordering of the deals does not affect this fundamental fact. It matters not a bit that the third deal is in the future with respect to the second deal. This is because dominance reasoning requires her to take the second deal, whether or not she takes the third. You will not be surprised to see that I make the same claim about the first trade: she will prefer to accept this trade, regardless of whether she accept the second or third trades, for the simple reason that she prefers strawberry to vanilla.

But what about the foresight response? Won't Jones see that accepting all trades leaves her with a sure loss? She will. But this gives her no reason to refuse the first trade. Let us assume that she predicts that she will act rationally in the future. (We will relax this assumption shortly.) Since accepting the second and third trades are the dominant options, Jones will predict that she will accept these trades. On this prediction, she is choosing between the following options:

(a) If she accepts the first trade, accepts the second and accepts the third, then she ends up with 1 pint of vanilla, 1 pint of strawberry, 1 pint of chocolate, but having lost some money.

(b) If she refuses the first trade, accepts the second and accepts the third, then she ends up with 2 pints of vanilla, 0 pints of strawberry and 1 pint of chocolate, having lost no money.

The crucial point is that even though Jones sees that (a) involves her making a sure loss overall, she still prefers (a) to (b). The reason why is that the two outcomes differ in the following respect: in (a) she has an additional pint of strawberry, whereas in (b) she has an additional pint of vanilla and some additional money. Since she prefers a pint of strawberry to a pint of vanilla and some additional money, she will prefer (a) to (b).

Now let us relax the assumption that Jones deliberates while predicting that she will act rationally in the future. I claim that Jones will prefer to accept the first deal, regardless of how she predicts she will act in the future. To see this, consider the following table, which illustrates all eight permutations of Jones's three choices.

We can apply dominance reasoning to Jones's decision about the first deal, by holding fixed various possible futures and pairing outcomes as follows. The pair of outcomes (1)

	Jones's Decisions			Jones's Final Possessions			
Outcome	Deal 1	Deal 2	Deal 3	Vanilla	Strawberry	Chocolate	Money
1	Accept	Accept	Accept	1	1	1	$-\Delta$
2	Refuse	Accept	Accept	2	0	1	0
3	Accept	Accept	Refuse	0	1	2	$-\Delta$
4	Refuse	Accept	Refuse	1	0	2	0
5	Accept	Refuse	Accept	1	2	0	$-\Delta$
6	Refuse	Refuse	Accept	2	1	0	0
7	Accept	Refuse	Refuse	0	2	1	$-\Delta$
8	Refuse	Refuse	Refuse	1	1	1	0

and (2) are the same with respect to her second and third decisions. The same goes for the pairs (3) and (4), (5) and (6), and (7) and (8). Therefore, holding fixed her future decisions about the second and third deals, we can see that there are four possible effects of accepting the first deal: accepting the first deal would bring about either outcome (1) rather than (2), or (3) rather than (4), or (5) rather than (6), or (7) rather than (8). Because Jones prefers strawberry to vanilla, she prefers outcomes (1) to (2), (3) to (4), (5) to (6) and (7) to (8). So regardless of which decisions she makes in the future, she prefers to accept the first deal. Accepting is the dominant option.

Symmetrical reasoning applies to her second choice. The outcomes can be paired so that Jones makes the same decision in the first and third deals in each member of the pair: <(1),(5)>, <(2),(6)>, <(3),(7)>, <(4),(8)>. If she accepts the second deal, then she brings about (1) rather than (5), or (2) rather than (6), or (3) rather than (7) or (4) rather than (8). Because she prefers chocolate to strawberry, she prefers the former of each of these paired options. Therefore, accepting the second deal is the dominant option.

Again, analogous reasoning applies to her third choice. The outcomes can be paired so that Jones makes the same decision in the first and second deals in each member of the pair: <(1),(3)>, <(2),(4)>, <(5),(7)>, <(6),(8)>. If she accepts the third deal, then for one of these pairs, she will bring about the former option rather than the latter option. Because she prefers vanilla to chocolate, she prefers the former of each of these paired options. Therefore, accepting the third deal is the dominant option.

My argument for the claim that Jones will accept all three trades is complete. It is important to note its scope is limited by an assumption that I made about Jones's reasoning. I assumed that Jones's preferences between individual pints of ice cream with different flavors remained the same, regardless of how much ice cream she has. In other words, I assumed that Jones's preferences over these components are (weakly) separable. Consequently, the argument neither targets intransitive preferences over indivisible alternatives nor nonseparable intransitive preferences over packages of these components.[10] For example, my argument fails to target Smith who has intransitive preferences about single pints of ice cream, (i.e., 1 pint vanilla > 1 pint chocolate > 1 pint strawberry > 1 pint vanilla), but whose preferences are transitive when she holds at least two pints of ice cream. Unfortunately, I cannot see any way to generalize my argument, which is unsatisfying, given that Smith's preferences are intuitively irrational, and we should like to have in hand an argument that demonstrates this.

In addition, dominance reasoning depends on the assumption that the agent's decisions are causally independent of each other. If this assumption fails to hold, then Jones would not be entitled to hold fixed her decisions about other trades when considering each permutation of possible choices. In other words, when making the first choice, she would not be entitled to compare the four pairs of (1) or (2), (3) or (4), (5) or (6), and (7) or (8). These outcomes are paired together because for each pair, the same decisions are made regarding the other two trades. These pairings are apt only if we assume each decision is causally independent of the others. Suppose instead that Jones's first choice causally influenced her future choices as follows. If she accepts the first trade, then this causes her to accept on the second and third trade; but if she rejects

the first trade, then this causes her to reject the second and third trade. Given these causal relations, when deciding whether to accept the first trade, Jones should compare outcome (1), which is reached by accepting three times, and outcome (8), which is reached by declining three times; since she prefers (8) to (1), she ought to reject the first deal. Thus, if the causal independence assumption fails to hold, then Jones may avoid becoming a money pump.

What could secure the causal independence assumption? A relatively weak sufficient condition would be that (i) Jones's choices do not change her subsequent preferences; and (ii) she makes each choice solely on the basis of her preferences. Condition (ii) is quite weak, as it does not require that the agent makes these choices rationally; it only requires that she reasons (rationally or irrationally) solely on the basis of her preferences. Moreover, assuming condition (i) holds is defensible given the broader dialectic: we are inquiring into the rational permissibility of intransitive preferences, and so it is appropriate to consider agents who stably maintain such preferences.

Let me end by comparing my argument to Rabinowicz's. Our arguments are similar insofar as they both appeal to a modified case, in which the other trader repeats each trade if it is declined. However, our arguments differ in two respects. First, my argument is more limited than Rabinowicz's insofar as it only targets agents with intransitive preferences that are weakly separable over components. (I.e. the agent's preference ordering remains constant, however much he/she possesses of the good in question). Second, our arguments employ different methods of reasoning that depend on different assumptions. Rabinowicz's argument uses backward induction, and so assumes that choosing irrationally does not shake an agent's trust in her future rationality. My argument uses dominance reasoning, and so assumes that each decision is causally independent of the others. I think there is a reasonable case that my assumption is less controversial than Rabinowicz's, but ultimately I propose that our arguments are complements rather than competitors. Assuming that either backwards induction or dominance reasoning is legitimate, the combination of our arguments shows that there is a way of turning an agent with intransitive preferences into a money pump.

Acknowledgments

For helpful comments, the author thanks Caspar Hare, Brian Hedden, and Wlodek Rabinowicz.

Notes

1 The idea is attributed to Norman Dalkey in the first paper proposing a money pump argument—Davidson, Donald, JCC McKinsey, and Patrick Suppes. "Outlines of a Formal Theory of Value, 1." *Philosophy of Science* 22.2 (1955): 140–60.

2 We should assume that once Jones has named her original price for trading her vanilla for strawberry, Mr Whippee continues to offer to trade strawberry for vanilla at this price. Otherwise, Jones could avoid losing all her money by paying $1 on the first round, $0.50 on

the second round, $0.25 on the third round, and so on. Thanks to Caspar Hare for making this point.

3 Schick, Frederick. "Dutch Bookies and Money Pumps." *The Journal of Philosophy* 83.2 (1986): 112–9 at pp. 117–118. See also Schwartz, T. *The Logic of Collective Choice*. New York: Columbia University Press, 1986 and Mongin, Philippe. "Does Optimization Imply Rationality?" *Synthese* (1999).

4 Schick, Frederick (1986:117–18).

5 Binmore, Ken. "Modeling Rational Plays I." *Economics and Philosophy* 3 (1987): 9–55; Bicchieri, Cristina. "Backward Induction without Common Knowledge." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1988, Volume Two: Symposia and Invited Papers, 1988, 329–43; Pettit, Philip, and Sugden, Robert. "The Backward Induction Paradox." *The Journal of Philosophy* 86.4 (1989):169–82; Reny, Philip. "Common Belief and the Theory of Games with Perfect Information." *Journal of Economic Theory* 59 (1993): 257–74; Binmore, Ken. "Rationality and Backward Induction." *Journal of Economic Methodology* 4.1 (1997): 23–41.

6 Thanks to Wlodek Rabinowicz for both the substance and much of the wording of this response.

7 The challenge is traditionally raised in the game-theory literature in the context of discussing interpersonal games. As Philip Pettit and Robert Sugden put it, "neither of the players can believe that the common belief in rationality will survive whatever moves the players make," Pettit and Sugden (1989:174). Thanks to Wlodek Rabinowicz for pressing me to clarify the difference between this formulation and the standard formulation.

8 In comments on this article, Wlodek Rabinowicz noted he considers the traditional formulation of the challenge the most serious one. But here our intuitions simply differ, as both challenges strike me as equally serious.

9 For general defenses of backward inductive reasoning against the challenge, see Sobel, Jordan. "Backward-Induction Arguments: A Paradox Regained." *Philosophy of Science* 60.1 (1993): 114–33; Aumann, Robert. "Backward Induction and Common Knowledge of Rationality." *Games and Economic Behavior* 9 (1995): 6–19. For defenses of backward inductive reasoning in games in which a player's move can terminate the game, see Rabinowicz, Wlodek. "Grappling With the Centipede: Defence of Backward Induction for BI-Terminating Games." *Economics and Philosophy* 14.1 (1998): 95–126; Broome, John, and Rabinowicz, Wlodek. "Backwards Induction in the Centipede Game." *Analysis* 59.4 (1999): 237–42.

10 I owe both the substance and the wording of this point to Wlodek Rabinowicz.