SUNK COSTS

Robert Bass, PhD
UNC-Pembroke
Robert.Bass@uncp.edu
478.238.3944

ABSTRACT

Decision theorists generally object to "honoring sunk costs" – that is, treating the fact that some cost has been incurred in the past as a reason for action, apart from the consideration of expected consequences. This paper critiques the universal doctrine that sunk costs should never be honored on three levels. As background, the rationale for the doctrine is explained. Then it is shown that if it is always irrational to honor sunk costs, then other common and uncontroversial practices are also irrational. Second, it is argued that there is no satisfactory way to make a distinction between the two kinds of cases. Third, a proposal is developed which would explain how it could be rational, under certain circumstances, to honor sunk costs.

SUNK COSTS

Consider the following case:

> You and your companion have driven half-way to a resort. Responding to a reduced rate advertisement, you have made a nonrefundable $100 deposit to spend the weekend there. Both you and your companion feel slightly bad physically and out of sorts psychologically. *Your assessment of the situation is that you and your companion would have a much more pleasurable weekend at home*. Your companion says it is "too bad" you have reserved the room because you both would much rather spend the time at home, but you can't afford to waste $100. You agree. Further, you both agree that given the way you both feel, it is extraordinarily unlikely you will have a better time at the resort than you would at home. Do you drive on or turn back? If you drive on, you are behaving as if you prefer paying $100 to be where you don't want to be [rather] than to be where you do want to be. (Dawes 1988, p. 22)

Suppose for the moment that you and your companion proceed to the resort because you cannot afford to waste the hundred dollars. If you do, it will be a paradigm case of what economists and decision theorists refer to, and object to, as honoring sunk costs. Their objection will be that it is irrational to go on rather than returning home. The hundred dollars have been spent in any case. Since the deposit was nonrefundable, going on to the resort will not result in any refunds – and so, all that should enter into the decision is how well you and your companion will (probably) feel, how much you will enjoy yourselves, etc., at the resort or at home. If, all things considered, you expect to feel better at home, then you should turn back. You gain neither the deposit nor any compensation for having made the deposit by going on to the resort; you only end up feeling worse than you would at home. Generalized, we get what I shall call *the economist's doctrine* – that it is never rational to honor sunk costs.

When matters are put this way, the economist's doctrine may seem obviously correct. Yet, that very obviousness suggests a problem. For the fact is that people frequently do appear to honor sunk costs. People sit through movies they are not enjoying because they paid for tickets, carry out

costly projects solely because of what has already been invested in them, and so on. If it is really so obvious that they should not, it is puzzling that they (we!) so often overlook the obvious.

Now, I think that, often, sunk costs should not be honored. I shall not be arguing that you and your companion are, after all, right to go on to the resort. However, I am less confident that the economist's doctrine holds for all cases. There may be cases in which it is rational to honor sunk costs.

I shall approach this in several steps. As background, I shall try to examine more directly the general reasons that can be given for the economist's doctrine. Then, I shall claim that, if the economist's doctrine is correct for the reasons given, it also shows that some types of actions that almost everyone is willing to allow to be rational (or, at least, not irrational) fail to be rational in just the same way that honoring sunk costs fails to be rational. Second, I will argue that there is no satisfactory way, consistent with the reasons given for denying the rationality of honoring sunk costs, to distinguish the two kinds of cases. The adjustments needed in order for these actions not to fall afoul of the canons of rationality would also allow us to salvage the rationality of honoring sunk costs. Third, I shall make a speculative proposal about one way in which it can sometimes be rational for us to honor sunk costs.

Why is it (thought to be) irrational to honor sunk costs?

For it to be the case that something can properly be called irrational, there must be some conception of what rationality requires. That conception need not amount to as much as a complete theory of rationality, but it does have to be sufficiently definite in what it requires that we can recognize (sometimes, anyhow) sufficient conditions for the irrationality of action. To put matters the other way around, the conception will have to enunciate at least some necessary conditions for

rationality. The conception that is relied upon – call it the Standard Theory[1] – in making the case for the economist's doctrine about sunk costs is, however, often *thought* to be a complete theory of rationality so far as it pertains to the selection of actions, specifically, to what it is rational to do, given certain information and objectives.

The Standard Theory sets out certain requirements thought to define the rational pursuit of goals or actions directed at the satisfaction of preferences given relevant constraints. The requirements do not include any substantive restriction on the goals or preferences themselves; any preferences, whether those would ordinarily be thought to be wise or foolish, virtuous or vicious, may be included without calling into question the rationality of a person's actions. The constraints may include conflict with other goals or preferences, limitations upon the resources available for deployment in such action, and possibilities of conflict or cooperation with other goal-seeking agents. What the theory claims to do is to say what considerations are relevant to acting so as to achieve one's goals or satisfy one's preferences, as well as can be expected.

Avoiding formalities, the conditions can be summed up this way:

> A *rational choice* can be defined as one that meets three criteria:
> 1. It is based on the decision maker.'s *current* assets....
> 2. It is based on the possible consequences of the choice.
> 3. When these conditions are uncertain, their likelihood is evaluated without violating the basic rules of probability theory. (Dawes 1988, p. 8)

An additional condition is needed to the effect that these considerations (alone) are to be used in ranking possible outcomes of one's actions with respect to the extent to which the outcomes satisfy one's preferences. Then, the action correlated with the most highly ranked outcome (or, of course, one tied for highest ranking) is to be selected. These conditions attempt to spell out what is involved in the intuitive idea of selecting the best means to one's ends.

---

[1] The Standard Theory is just expected utility theory, which owes its formal statement to the work of Von Neumann and Morgenstern and to many subsequent decision theorists.

How do these conditions bear on the correctness of the economist's doctrine with respect to sunk costs? The answer seems straightforward: If the conditions exhaust the considerations that it is rational to attend to in making a decision, there seems to be no place for sunk costs to appear so as to make a difference to what one ought to decide (or in what one would rationally decide). That may be too abstract to be clear, so let us spell it out a bit further. A sunk cost, by definition, is one that has already been incurred. The fact that it has been incurred may have various causal impacts on the considerations that are allowed by the Standard Theory. The action through which the cost was incurred may affect one's current assets: They are less than they would be if everything else had gone the same, but one had not incurred the sunk cost. The sunk cost may have affected what the possible consequences of one's current actions are. It also may have affected the probabilities associated with those consequences. However, it will not affect what one ought to do, *given* a particular set of preferences, assets, prospective consequences and associated probabilities.

One way to make this clear is to compare two choosers who are faced with the same choice situation – with the same preferences, the same assets, the same information and facing the same set of alternatives – except that one of them has not incurred a relevant sunk cost while the other has. Imagine someone in almost the same situation as you and your companion in the story with which we started. She and her companion are also half-way to a resort when they realize they would enjoy the weekend better at home. They have (let us suppose) exactly the same assets, preferences and expectations as did you and your companion. However, she and her companion did not incur any sunk cost in the form of a nonrefundable deposit. They had a hundred dollars less in assets to begin with. She and her companion will be better off if they turn back, just as you and your companion would be. The fact that different causal sequences led up to the two pairs' having the same preferences, etc., and that one of those sequences did, while the other did not, involve incurring a sunk cost makes no difference. If only the considerations allowed by the Standard Theory as relevant to decision making are admitted, she and her companion will turn back. Since precisely the same

considerations bear on whether you and your companion should turn back, you also, provided you are guided only by what the Standard Theory allows, will turn back. Since both pairs will make the same decision but differ as to whether a sunk cost has been incurred, the presence of a sunk cost makes no rational difference to the decision.

Generalizing, for any case in which a sunk cost has been incurred, we can construct a parallel case in which preferences, assets and expectations are precisely the same, but no sunk cost has been incurred. What it is rational to do in such a situation can be worked out for that parallel case using all and only the considerations allowed by the Standard Theory. If a decision maker nevertheless does not make that decision because of some sunk cost, then the decision maker must have violated one or more of the conditions on rational choice proposed by the Standard Theory.

The Problem of Promises

People often make and not quite so often keep promises. Very few would hold that it is irrational to keep promises. Yet most of us would say, I think, that part of the reason for keeping a promise has to do with a past event – with the fact that one made the promise.

It appears that that fact, however, is not one that the Standard Theory will allow us to take into account in considering whether or not to keep a promise. The argument for this is analogous to the argument against honoring sunk costs, so we can be brief. Having made a promise may have a causal impact on what one's current assets are, on what the possible consequences of one's actions are and on the probabilities associated with those consequences. Here, too, we can imagine parallel cases in which preferences, assets and expectations are held constant but in which no promise has been made. The Standard Theory would prescribe the same action for both cases, whether a promise has been made or not. Thus, *given* a set of preferences, assets and expectations, having made a promise should make no difference to what is done. Accordingly, if it *does* make a difference, the rationality conditions of the Standard Theory must have been violated in some way.

But surely, it is counter-intuitive to say that it is irrational to keep a promise unless, with the same preferences, assets and expectations, one would have performed the same action even if one had not made a promise. To sharpen the point, consider the following: Suppose I promise to take in a neighbor's mail while he is on vacation. I like my neighbor and the effort involved is minimal. However, while he is on vacation, I get sick. Relative to how much energy I have, the effort is greater and with my current preferences, etc., I find that I am indifferent to whether I take in his mail or not. If I do not, my neighbor may be mildly annoyed and I can take that consequence into account in deciding whether or not to take in his mail. Offsetting that is the fact that I will feel slightly better if I do not take the trouble. Since, in terms of my current preferences, assets and expectations, I am indifferent whether I take in the mail or not, taking it in would be effectively *costless* for me. I face one set of costs and benefits if I do take in his mail and a different but equivalent set of costs and benefits if I do not. Since I am, in terms of my assets, preferences and expectations, no better or worse off on account of taking in or not taking in my neighbor's mail, the Standard Theory has to say that the two options are rationally indifferent. Is it really plausible that the fact that I've made a promise should make *no difference* to whether or not I bring in the mail even though there would be no net cost to me in bringing it in? If the Standard Theory requires us to say there is no rational difference, many of us will be suspicious that it is the theory that has gone wrong somewhere rather than the practice of promise-keeping.

Intuitions may mislead, however, and there are complications to be considered. So let's set that case aside and consider what can be said on behalf of promise-keeping within the confines of the Standard Theory. One consideration that can readily be accommodated by the Standard Theory is the likely effects upon reputation that will flow from breaking a promise. If someone considering keeping or breaking a promise cares about his reputation, either directly or because of the effect that having a reputation as a promise-keeper is likely to have upon the future willingness of others to enter into beneficial agreements with him, then that may certainly be included as a relevant

consideration. There are also, no doubt, more complicated ways in which keeping promises may causally affect what options will be open to the promise-keeper in the future, and considering those presents no problems for the Standard Theory.

Such concern with the future consequences of promise-keeping seems inadequate to meet the initial question, however, for surely there are cases in which, with a high degree of certainty, we know that the future positive consequences of promise-keeping or negative consequences of promise-breaking are negligible. A standard example is the death-bed promise, known only to the promisor and the promisee – one of whom, the promisee, dies shortly thereafter without informing anyone else. If keeping the promise looks as though it will be costly or if breaking it appears advantageous (even slightly), then it seems that, given the absence of normal reputation effects, the Standard Theory would counsel breaking the promise. Yet most of us and, I suspect, most proponents of the Standard Theory, would take death-bed promises to be especially serious and be especially reluctant to break them for the sake of small gains.

Where else, if not in future consequences, can the Standard Theory find lodging for *rationally* taking promises seriously? The obvious alternative is to say that the agent may regard being a promise-keeper as a current asset which will be lost if he breaks a promise. This *may* give us what we want.[2] If I have made a promise and treat being a promise-keeper as a current asset that I am loath to relinquish, that will provide a sense in which my preferences, current assets and expectations *cannot* be paralleled by anyone whose situation is otherwise identical but who has not made a promise. If I take an action which amounts to breaking my promise, I will lose the asset of being a promise-keeper; if a person in the parallel case who has not made a promise takes a similar action, he will not lose the asset of being a promise-keeper. In that respect, our situations necessarily

[2] If this alternative is adopted, we need to be careful how the position is stated. It will not do – because it would land us in vitiating circularity – to say that the reason for promise-keeping in the cases we are considering is to avoid the regret that would attend the loss of the asset of being a promise-keeper: If one did not value the asset for some other reason, one would not regret its loss.

differ, so we are not in a position to construct our by-now familiar argument that, if preferences, assets and expectations are the same, but the choice differs, there must be a violation of the Standard Theory's conditions upon rational choice.

I think we should be very suspicious. First, this supposed solution sounds very much as if a past difference – between one who has made a promise and one who has not – has simply been redescribed as a difference in consequences for one's current assets. Surely, it is difficult not to suspect trickery here.

Second, it seems that if the solution works for this case, we can easily apply an analogue to argue that there is nothing irrational about honoring sunk costs. We only have to redescribe the person who honors sunk costs so that he is protecting a current asset. The person who drives on to the resort for which he has made the nonrefundable deposit (though he would feel better at home) may just say that he treats being someone who honors sunk costs as a current asset which he is loath to relinquish. No one who had not incurred a sunk cost *could have* the same preferences, assets and expectations. If *he* turns back, he will relinquish being someone who honors sunk costs; if someone who has not incurred a sunk cost turns back in an otherwise similar situation, that person will not relinquish being someone who honors sunk costs. Since the cases differ, the argument that the one who honors sunk costs must be violating some condition on rationality fails.

Something seems to have gone wrong. If promise-keeping turns out to be rational, in the range of cases in which we want to regard it as rational, so, it appears, does honoring sunk costs. The conclusion, however, is not easy to avoid for, as noted earlier, the requirements of the Standard Theory include no substantive restrictions on the goals or preferences by which an agent may rationally guide his action. For an agent to treat being someone who honors sunk costs as part of current assets is just as respectable, so far as the theory goes, as for some agent to treat being a promise-keeper as part of current assets.

At this point, there seem to be four options:

- We could try to find some further difference, respectable within the Standard Theory, between promise-keeping and sunk-cost honoring that would enable us to regard the former as (typically) rational (even when there appear to be gains to be had from promise-breaking) and the latter as irrational. I think there's very little hope that that search will succeed.
- We could accept the proposed solution, in terms of protecting current assets, for both promise-keeping and for the honoring of sunk costs. We'll be able to defend the rationality of promise-keeping, but at the price of admitting that honoring sunk costs is rational as well.
- We could reject the proposed solution and hold that neither promise-keeping nor the honoring of sunk costs is acceptable unless justified in other ways. That will mean that we have no reason to keep a promise unless, in a parallel situation, with the same preferences, assets and expectations, we would find it rational to perform the same action even if we had not promised.
- Last is the option I recommend: we could conclude that the Standard Theory is incomplete and that work needs to be done towards working out a richer theory, capable of making subtler distinctions between promise-keeping (which we often take to be rational even if there are some gains to be had from promise-breaking) and, at least some cases of the honoring of sunk costs (where it seems that doing so is something that, on reflection, we take to be irrational).

When might it be rational to honor sunk costs?

I am not in a position to offer anything like a general theory to replace or supplement the Standard Theory. I do not believe that anyone has worked out a satisfactory and complete account.[3] Nonetheless, if we reject the completeness of the Standard Theory (as I think we should), it is worthwhile to explore what some other relevant considerations might be, even if we do not yet have a fully adequate alternative.

---

[3] Actually, I am inclined to think that it will not be possible to work out a satisfactory and complete account without venturing beyond consideration of what best serves whatever goals or preferences we happen to have into the domain often reserved for moral theory, consideration of the goals and preferences which it is better to have. But that would be a very large project, and I won't begin to argue further for it here.

Let us suppose that it may sometimes be rational to honor sunk costs or, more generally, to take into account in our decision-making facts about what has occurred in the past and to treat them as reasons that make a difference to current decisions, at least partially apart from the ways in which those past facts affect the set of possible future consequences of our actions.

Consider goods that are produced by human action that have the following properties: First, they must be produced (if at all) over an extended period of time. Second, they are complex in that they have multiple distinguishable parts. Third, they possess what some philosophers have called *organic unity* – their value as goods depends to some extent on the way that their parts fit together, not just upon the value of the parts separately considered. Fourth, the action which produces the good has the same kind of structure delineated in the three conditions above in that it is temporally extended, composed of multiple distinguishable parts and the parts are organically related to the completion of the productive activity – that is, they are valued more highly as parts fitting together into the production of the good than they would be separately. This may seem excessively abstract, and perhaps it is, but it is meant to provide a passable portrait of what often happens in the production of works of art.[4]

Now, suppose that you are in the process of producing such a work. For illustrative purposes, I shall stipulate that it is a symphony and that you are a talented composer. You have written part of the symphony and are faced with the question whether to finish it. Many factors may enter into this decision. One is whether you might write something better if you laid this one aside. Another has to do with how other demands on your time may compare to the value of this completed work (should you complete it). But it seems that one thing that may properly enter into your decision – and may make a difference if other considerations are sufficiently close to being balanced – has to do not with the future consequences of completing the symphony but with the fact that effort that you

[4] Though I do not follow him closely, I got the idea for this line of thought from Thomas Hurka. (Hurka 1993, pp. 110f.)

have put into it *in the past* may be more valuable if the symphony is completed. For remember, we were assuming that your work on the project has an organic structure such that its parts are more valuable *as parts of the whole* than they are considered separately. So, in a certain sense, as strange as it sounds, a current action may contribute not only to a stream of benefits to be realized in the future, but may have as a "consequence" that something that has *already* occurred is better than it would otherwise be. If you do not complete your symphony in A#, it will not be true that six weeks ago you wrote the opening chords of that symphony. At most, you will have written the opening chords of an uncompleted symphony.

I chose to present this possible pattern of motivation and evaluation with an aesthetic example because the pattern is especially clear there. But once we have noticed the pattern – where an outcome of a process of action over an extended period of time has the structure I delineated and where the activity that produces that outcome has the same structure and is also valued, we can see many other examples of the pattern. It shows up in friendships, in loving relationships, in the pursuit of satisfying careers and elsewhere. In all such cases, what we care about is more than just a future stream of benefits. The person who does not treat past actions as part of such a pattern – and therefore worth continuing in part because of the value that the whole pattern confers upon its parts, including parts that are in the past – is always open to a better deal, a better friend, a better career, a better mate. However, his continued openness to change on the basis of expected future consequences of the change actually (tends to) prevent him from finding and enjoying a good friendship, a satisfying career, a successful relationship.

Two points should be noted about this. It might appear that we have appealed after all, ultimately, only to things that can be countenanced by the Standard Theory. In effect, I have pointed out that a person is likely to be worse off in terms of the future consequences about which he cares if he only pays attention to those future consequences. In a sense, that is right, but it doesn't salvage the Standard Theory, for that theory imposes the requirement that he not consider anything but future

consequences. What it shows instead, and this is the second point, is illustrative of something more general, that a method of reasoning about what to do may be *self-undermining*. A person who considers only future consequences may discover that the future consequences of considering only future consequences are worse than those that can be achieved by taking other things into account.[5]

References

Dawes, Robyn M. 1988. *Rational choice in an uncertain world*. Fort Worth: Harcourt Brace College Publishers.

Hurka, Thomas. 1993. *Perfectionism*, Oxford ethics series. New York: Oxford University Press.

[5] It may be useful to distinguish between consequentialism in a decision-making procedure, which refers to the view that nothing but consideration of consequences should shape one's decisions, and consequentialism as a moral theory, which holds that the comparative value of actions or events is determined by nothing but valuable or disvaluable consequences. The moral theory does not necessarily prescribe a decision procedure.