Prudence, Sunk Costs, and the Temporally Extended Self

When we make major life choices, we sometimes look to our past as well as our future. But can doing so be in our self-interest? Is it prudentially rational? Set aside cases in which past actions have a *causal* effect on the future – no doubt I should bear in mind the fact that I've taken a big mortgage when considering downshifting, since the bank won't let me forget about it. But here it's really the fact that the bank will come after me if I don't pay that counts. Set aside, too, cases in which what happened before is an *indication* of what is likely to happen in the future. It's surely a good idea to bear in mind that my next diet will probably not be any more successful than the previous ones, unless I do something quite different. But here the past matters only insofar as it helps predict the future. In this paper, I'll investigate and defend the more interesting thesis that it can be prudentially rational to give weight to past actions and events in choosing between possible futures even when they don't play such causal or informational roles.

How could the past merely as such affect what it is in our rational self-interest to choose? Several philosophers, such as David Velleman (1991), Elizabeth Anderson (1993), and Thomas Kelly (2004) have argued that while we can't change what happened in the past, our choices can change the *significance* of past events, and past events can change the significance of future choices. I think this is the right way to go. But this approach still remains underdeveloped. Here, I focus on two main questions. First, why isn't taking our past actions into account simply committing the fallacy of honoring sunk costs, that is, allowing ourselves to be influenced by sentimental considerations that have no bearing on the expected utility of our options? I argue that when (and only when) we can genuinely change the value of past investments, this objection doesn't apply. This is where significance comes in: I hold that the prudential value of events for us depends in part on what I call their teleological significance, roughly their contribution to our excellence as a temporally extended agents who

pursue long-term aims. And what we do now can change the teleological significance of past events, and therefore make our past better for us.

This brings me to my second main question: why does prudence require us to care about our past good, rather than just present and future fortune, as many assume? In his recent work, Dale Dorsey (2017a) argues, roughly, that insofar as I now normatively expect my future self to cooperate in realizing my present projects, I owe it to myself to treat my past self's projects likewise. I argue that Dorsey's answer faces a dilemma: either we treat intrapersonal relations as analogous to interpersonal relations, in which case his proposal fails to generate prudential or rational requirements, or we take seriously the notion that we are temporally extended agents, in which case the puzzle he addresses does not arise. I suggest we should take the latter option, and show how caring about the teleological significance of the past is linked to attitudes like self-respect and pride.

1. Commonsense Data

Let's start with a few brief scenarios in which it's natural to think that what happened earlier matters for choosing well, setting aside moral considerations. Here's the first one:

> *Lawyer in Recovery*
>
> Jerry is a graduate of a prestigious law school. He is determined to dedicate himself to a job that really makes a difference. He gets offered a position at two non-profits, one dedicated to assisting inner-city youth, and the other to helping immigrants. He believes with good justification that these are both equally worthy causes and that he'd be equally successful and happy doing either. There is, however, one relevant difference: many of the inner-city youth he'd be working with suffer from a variety of problems related to substance abuse. As it happens, Jerry himself has a dark history of alcoholism and addiction, which almost dragged him under in his mid-20s, until he

caught a lucky break with a former football coach, who convinced him it's not too late to join a support group and go back to college.

(For more scenarios like this, see e.g. Velleman 1991 and Dorsey 2015.) The way I've described the case, Jerry's going to do just as well whichever way he chooses. Let's suppose he'll also do just the same amount of good for others. We may add that he knows that if he chooses to work with immigrants, he'll never give another thought to his past, so it's not going to be something that will bother him in that case. Still, there appears to be a relevant difference between the two scenarios. Helping kids who struggle with substance abuse isn't just another worthy cause for Jerry. It's a way of *redeeming* his own past struggles to some extent.[1] He won't have to look back on his downward spiral as a total loss any more, if he does look back. Instead, it will have been a way to gain insight into something important. He'll make it into something meaningful, as it is natural to say. It's not that he'll necessarily feel good about it. He may or he may not, and in any case I'm stipulating that he'll feel just as good in the future if he works with immigrants instead. (If it helps, imagine both offers come with a selective amnesia pill that will make him forget the relevant part of his past.) But he will have *good reason* to feel a bit better about his past mistakes if they turn out to serve something valuable. It seems reasonable for him to take this into account as one consideration in making up his mind. (Whether it actually *is* reasonable is what I'll discuss in this paper.)

For a real-life case of this sort, we might think of Monica Lewinsky, who was publicly shamed and ridiculed when her affair with Bill Clinton become public. After a long time trying to run away from these events, she has recently begun to work and speak for victims of

[1] The notion of redemption is introduced into well-being literature by Velleman (1991/2015) and developed by Portmore (2007). As will become clear, I cash it out in a different way.

online bullying. In an interview with *The Guardian* in April 2016, she herself said that "To be able to give a purpose to my past, if I'm stuck with my past, feels meaningful to me".

Here's a case that is a mirror image of the first:

Sell-Out Scientist

Kathy is a well-known research biologist at a large state university. One day she gets an offer from an R1 University, offering her a state-of-the-art laboratory with enough funding to hire the best talent to work with her on whatever she wants. She also gets a similar offer without any teaching or admin responsibilities from a major agriculture corporation developing genetically modified vegetables. The catch is that she has spent the last nine years of her life leading an ultimately successful campaign against using GMO corn sold by the same agriculture corporation as cattle fodder, having accidentally discovered it significantly lowers cattle welfare. Alas, she realizes that if she takes the job with the corporation, she'll have to give up her advocacy, and the changes she has wrought will quickly be reversed by a skillful, celebrity-studded PR campaign.

Once again, Kathy is set to do equally well in either scenario, if we focus on self-interest and look to the future – let's assume that the generosity of the corporate funding compensates for any drawbacks the option might otherwise have. If that's difficult to believe, you may be underestimating the human capacity for self-deception in the sell-out option. Her past will not come back to haunt her if she sells out. We may stipulate that were it not for her activist past, Kathy might even prefer having a private lab with no admin duties. Alas, the past is there, and she realizes that accepting the new offer would mean undermining her signature achievement. Again, it seems this is a consideration that should bear on Kathy's decision, even if she leaves moral considerations aside. And again, this is not, or not only, because she would feel bad

about it – it would be a bad idea, even if she also got a pill that would wipe the events from her mind (and the minds of those around her). If she were to ask for my advice, I'd tell her not to tarnish her own achievement by selling out. To be sure, everyone whose choice would have the same causal consequences would have a moral reason not to work for the corporation, but in her case there is an *additional* self-interested reason not to do so.

Finally, consider a different type of scenario:

Two Awards

Sally studies at a prestigious architecture program. Every year, the students compete for two prizes, the Classical or the Modernist one, either one of which will open a lot of doors for the future. Sally would love to have either one just as much, but one student can only get one, so she flips a coin and immerses herself in a project on rococo architecture, and comes up with a novel thesis on its unifying features. The Classical judges are delighted, and she is asked whether she would accept the Classical award. But before she does so, she is also offered the Modernist prize! Baffled, Sally asks the Modernist judges if they know what she's been working on. They reply: "Yes, we know, but we saw how hard you worked, and we've never given a prize to anyone from your minority group before. And just a year ago, we heard you say you would love to have either award just as much!"

Once again, if what she has done in the past didn't matter to Sally – imagine she took the amnesia pill – she would be indifferent between the two awards. Of course, she might now anticipate feeling bad about accepting the Modernist prize, but if rationality were exclusively forward-looking, such feelings would be irrational, so that she should welcome a drug that gets rid of them. Yet it seems clear that it's *not* irrational for her to have a preference between the awards, given her past. This time, the issue isn't about wasting or contaminating her past

efforts – after all, if she hadn't worked so hard, she wouldn't have been offered either prize. It is rather that in the case of the Modernist award, her success would be related to the past in the wrong way, due to irrelevant factors rather than what she's accomplished. And this, again, seems on the face of it like the kind of backward-looking consideration it is reasonable for us to take into account in making decisions.

2. The First Concern: Committing the Sunk Cost Fallacy

In the scenarios I just gave, what one has done in the past intuitively makes a difference to what it is rational for one to choose in one's self-interest, apart from its causal consequences or informational value. Nevertheless, the idea goes against some very mainstream views about prudentially rational choice, according to which we should ignore any past investment, lest we commit the sunk cost fallacy. In this section, I'll argue that while there really is such a fallacy, it is only irrational to give weight to past costs when they are genuinely 'sunk' – that is, when we can no longer change their value.

According to standard decision theory, to choose rationally, you need to work with both the utility and the probability of all possible outcomes. The utility of an outcome for you is derived from your preferences between outcomes, provided they meet constraints like completeness and transitivity. So if you consistently prefer a Big Mac to a Whopper, eating a Big Mac has higher utility for you than eating a Whopper. When you put together the probabilities and utilities of all possible outcomes of an act, you get its expected utility, and the rational thing to do is to choose the act that has the highest expected utility. So given your preferences, rationality tells you to go to McDonald's, assuming all you care about is burgers.

What I've just said is widely accepted. But many theorists also take it for granted that when we form preferences, we should be *exclusively forward-looking*. It would be a mistake to let past choices influence your decisions now (except insofar as they affect the future). If

you did so, you'd be guilty of the Fallacy of Honoring Sunk Costs, or "a greater tendency to continue an endeavor once an investment in money, effort, or time has been made" (Arkes and Blumer 1985: 124). Sunk costs are costs you can no longer recover, and therefore should ignore in your choices. For example, suppose that last week you bought a ticket to see Televisionhead tonight for $120, but now all of a sudden, someone gives you a ticket to The Fillers, whom you'd much prefer to see tonight. No one else wants your Televisionhead ticket. It's natural for you to think "I don't want the $120 I paid to go to waste, so I'll go to see Televisionhead". But I agree with the economist that doing so would be irrational, if you really prefer The Fillers. Whatever you do, you'll never see that $120 again. That cost isn't recoverable. But you can go see the band you prefer, so that's the rational thing to do.[2] I also agree with the economist who points out that it's foolish to stay in a doomed relationship *just because* you've already invested so much time and effort in it. It's better to cut your losses. It is tempting to think such cases mean that "Things that happened in the past matter only insofar as they affect future outcomes", as a recent textbook puts it (Angner 2016: 43). Call this view The Principle of Future Exclusivity.

If the Principle of Future Exclusivity is true, the intuitions I started with are cognitive illusions of some sort. But is it true? *Why* exactly is it irrational to care about sunk costs? Officially, decision theory is fully neutral regarding the content of your preferences, as long as they are coherent. In the ticket case, you prefer The Fillers to Televisionhead. But if you choose to see the latter instead, this reveals that you prefer the *combination* of seeing Televisionhead and making use of the ticket you bought for $120 to having paid $120 for Televisionhead and seeing The Fillers for free. If the utility of an outcome is simply a function of coherent preferences, using the ticket you bought to go see Televisionhead has the

[2] If, however, you would feel very bad about wasting your ticket, you should see Televisionhead after all. But, as Kelly (2004) emphasizes, that would not be a case of honoring sunk costs, but forward-looking choice based on anticipated (irrational) future feelings.

highest utility for you. (There's nothing incoherent about preferring to see The Fillers to seeing Televisionhead while simultaneously preferring [seeing Televisionhead + using the ticket I bought] to [seeing The Fillers + throwing away the ticket].) So going to the Televisionhead show seems like the rational choice according to the standard decision-theoretical picture, rather than a paradigm of irrationality! Indeed, *whenever* you prefer the combination of past investment + inferior future outcome to the combination of the investment with a superior future outcome, the latter has a lower utility for you by the standard definition of utility, and you're irrational if you choose the better future.

Now, I don't want to say, as some philosophers do, that caring about genuine sunk costs isn't irrational. I think it *is*, even if having a *reputation* for taking past investments into account can be beneficial in some contexts, as Robert Nozick (1993) argues. But this needs explanation, which is *not* provided by the standard view, or indeed any view on that is strictly neutral on the *content* of our preferences. This is parallel to the argument that John Broome (1999) makes regarding the transitivity of preferences: we can always individuate options in a choice situation in a way that allows a seemingly intransitive set of preferences to come out as coherent, so when we make judgments of (subjective) rationality, we need to go beyond the agent's actual preferences and ask whether they have sufficient reason to have them. The lesson he draws is that "it is not rational to have a preference between two alternatives unless they differ in some good or bad respect" (Broome 1999: 75).

We can give a parallel explanation in the case of sunk costs. In the ticket case, you don't have sufficient grounds for thinking that [seeing Televisionhead + using ticket] is *better for you* than seeing Televisionhead alone, and or that [seeing The Fillers + wasting Televisionhead ticket] is worse for you than seeing The Fillers. The money you spent on the Televisionhead ticket is gone either way. What this suggests is that you should rationally ignore what happened in the past when it is irrelevant to the *value* of your options (in the light

of your evidence), where the value of an option is not a function of your actual preferences, however coherent (since otherwise we'd lose the critical distance required to say that preferring the sunk cost option is irrational). That is, we must identify sunk costs in accordance with the following:

> *The Sunk Cost Principle*
>
> A past investment is a sunk cost if and only if nothing you can now or in the future do will make a (noteworthy) difference to its value for you.

The Sunk Cost Principle explains why the money you paid for the Televisionhead ticket is a sunk cost: whether you now use it or not, it won't make buying it a better or worse an investment, at least to an extent that would make it noteworthy in deliberation. When you have access to sufficient evidence for this fact about value, it is irrational to prefer the combination of the past investment and a worse future to the combination of past investment and a better future.

This explanation of the Sunk Cost Fallacy clearly departs from neutrality regarding the content of your preferences, since it says that caring about past investments is irrational only when it does no good for you in the light of what you know. My claim is thus that we must amend decision theory to explain why it is irrational to care about sunk costs. But there is in any case reason to do so, even if we don't buy Broome's point about transitivity. After all, decision theory is supposed to provide us with a *normative standard* for guiding and evaluating our choices, and it is difficult to see how it could be normative for us unless the inputs (the preferences that are the basis for constructing a utility function) at least fallibly track genuine reasons. Here I follow L. A. Paul, who proposes a simple amendment: in order to be rational, "we must assign our values and credences based on sufficient evidence" (Paul

2014: 22) – that is, rational agents don't just base their credences on evidence, but *also* their preferences (Paul's 'values').

Importantly, on the kind of view that does link rationality and value, if we *can* now change the value of past investments, they *won't* amount to sunk costs. If you're sure that your relationship will not get any better, then you should break it off right now, and forget about whatever you've done in the past to fix things. But if you think there's a chance that those efforts won't have been in vain, that you may turn a corner soon, it's less clear what you should do, even if a new option arises by dint of luck. Maybe you can make your past investment count for something. And if so, what happened before is *not* a sunk cost (or, more precisely, not all of the cost is sunk). It might be reasonable to prefer hard-won happiness to a stroke of luck with a new partner, other things being equal. If you can change the (preference-independent) value of your past investments, you can to some extent 'recover' them. That's why one can consistently think that honoring genuine sunk costs is irrational while denying that it is irrational to care about the significance of the past, and thus rejecting the Principle of Future Exclusivity.

3. Prudential Value and Teleological Significance

So, the question now is this: can we really change the value of past efforts and investments? A defender of the Principle of Future Exclusivity might argue that preferring a past with a certain significance is irrational, because how well or badly things went for us depends on something we can no longer change. This objection may well seem initially plausible, given that we can't change what happened. The past is fixed. That's no doubt true, when it comes to *causal* difference-making. But consider Arthur Danto's well-known observation that in 1618, even an ideal chronicler with all the information about the events at the time could not have known that they amounted to the beginning of the Thirty Years War (1962: 154–155). The

core idea here is that some truths about earlier events depend on later events. This isn't causal dependence. Rather, it is a kind of non-causal dependence that we sometimes try to characterize in the language of meaning or significance. As Danto's case shows, it should be no more controversial that later events *can* change the *significance* of earlier events than that they can't *causally* affect them.

So *if* (and plausibly *only if*) the prudential value of past events for us depends on their significance, it can be changed by what happens later.[3] It is no wonder that those who believe in the rational importance of the past do use the language of significance in this context (e.g. McMahan 2002 and Kelly 2004). But what exactly does 'significance' mean in this context, and why should we think it matters for prudential value? There has been a lot less work spelling this out. In the rest of this section, I will summarize and develop the account I have proposed in past work on these issues (Kauppinen 2012; 2015a; 2015b).

The common intuitions about redemption, contamination, and merit I introduced in the first section suggest that changing the significance of the past can indeed affect its value, and consequently what it is rational to choose. What I'll argue now is that such intuitions can be explained by principles that have a good independent rationale. The framework within which I'll formulate my principles is based on the idea that if we want to know what's fundamentally good for us, it's a good idea to begin by asking who we fundamentally are. This is a broadly Aristotelian way of approaching things. I don't claim that it's the only way – indeed, the argument I'm going to make could fairly easily be formulated in terms of a subjectivist value realization view of the sort developed by Valerie Tiberius (2008) and Jason Raibley (2013).

[3] I've addressed formal objections to relational conceptions of final value by Bradley (2009) in earlier work (Kauppinen 2015a), so I will assume here that the crucial questions are matters of substantive value theory.

Nevertheless, I believe the Aristotelian approach offers an appealing and simple way to make sense of the intuitions. Aristotle thought that different things are good for different kinds of beings in virtue of their different natures (e.g. *NE* 1098a), and held that "what properly belongs to each thing by nature is most excellent and most pleasant for each of them" (*NE* 1178a). For example, plants have the potential to grow and reproduce, and they flourish when they do so successfully. This explains why it's good for them to get water and sun. There is no need to buy wholesale into Aristotelian teleological metaphysics to agree that there's something importantly right about what he says about what's good for living things. If we want to know what it is for us to flourish, it makes sense to ask what kind of beings we are.

As such, this approach isn't necessarily partisan at the level of first-order theories of well-being. Both hedonism and various subjectivist views, such as preference-satisfaction accounts, can be seen as answers to the question of what it is for beings like us to flourish. They both highlight an important fact about us. One thing we fundamentally are is subjects of experience. This is a really deep fact about us. It's part of what makes us human. We have consciousness – we're not just *there* like a thing, but there's *something it's like* for us. In part because of this, we also have subjective preferences between possibilities and value some ways of life. We'd like to have certain kind of experiences and avoid others. But as the experience machine considerations reveal, we also value and prefer *doing things* rather than just having experiences (Nozick 1974, Lin 2016).

These deeply rooted values and preferences points to our other fundamental aspect, active agency. We are the kind of creatures who can change the world to fit better with our conception of how it should be. When we exercise our agency, we conceive of a goal or an end, think of a means to realize it, and then perhaps take those means. This results either in the realization of our goal, or it doesn't. What is distinctive of human as opposed to animal agency is that we can think about the distant future in the light of our values, and form long-

term goals, ends that shape the more immediate goals we pursue and the way we pursue them. Often the end is distinct from the actions themselves, such as when you aim to cure a disease, but it is also possible to aim just at doing something well, such as when you aim to be a good father (see *NE* 1140b).

Now, my quasi-Aristotelian thesis is this: exercising agency well is a crucial part of what it is for us to fare well. After all, this is what we can't do inside the Experience Machine. But what kind of exercise of agency is good for us? Traditionally, Aristotelians have adopted *perfectionist* views, according to which what's basically good for us is the development and exercise of our natural capacities (Kraut 2007, Bradford 2015). But I want to take a different tack that avoids the difficult challenge of identifying capacities worth developing and exercising in an evaluatively neutral way. So instead of traditional perfectionism, I'll take it as my starting point that from the perspective of agency, our life consists of a series of events and actions that pertain to the realization of our aims. In an important sense, our life goes well when these things go our way. A little more precisely, my general thesis is the following:

> *Teleological Significance*
>
> The non-instrumental prudential value of a subject's life for her at time *t* is determined, in part, by the teleological significance of the actions and events that constitute her life at *t*. The teleological significance of an action or an event is its contribution to the subject's excellence as the kind of goal-directed agent she is – in the case of humans, temporally extended and fallibly reasons-responsive.

What I'll do in the rest of this section is suggest two simple diachronic principles for teleological significance, which are motivated by this conception of our nature, and explain the common intuitions I started with.[4]

[4] Again, I'm drawing and elaborating on earlier work, especially Kauppinen 2012 and 2015a.

The first and most obvious principle is motivated by the thought is that since agency is about pursuing aims, something contributes to excellence as an agent when it contributes to or constitutes success in realizing our aims. Compare the following brief scenarios:

Success/Failure

Alex studied hard for the entrance exam, learned many new things, and was accepted. Before he found out about the results, he was killed in an accident.

Bert studied hard for the entrance exam, but wasn't able to grasp the material, and was rejected. Before he found out about the results, he was killed in an accident.

While both Alex and Bert met a tragic end, it seems to me that Alex's life went in one respect better than Bert's, even if their experience was the same (maybe both came away from the exam thinking they had nailed it). It may have been because she had more skill than Bert, or simply better luck. Either way, contributing to success in our aims is one way in which something can benefit us as agents. So cases like Success/Failure suggest the following simple principle:

Instrumentality

Other things being equal, an exercise of agency is finally prudentially good for you to the extent it contributes positively to reaching your aims.

I emphasize that we're talking about *final* (or non-instrumental) prudential value here: it is finally good for us (other things being equal) for our actions to serve our purposes rather than be wasted. Or perhaps better, it's finally good for us as agents that we *make progress* towards our aims. It is, of course, also instrumentally good for us, but my claim is that the value of making progress isn't reducible to its instrumental value. Instead, whenever we make progress, we are actualizing the potential we have as goal-directed agents, and thus realizing

our nature, to put it in Aristotelian terms. Note also that the Instrumentality isn't restricted to present aims – progress towards future aims also counts, which will be important in the big picture.

Instrumentality captures the aspect of intrinsic value that is highlighted by simple achievementists like Simon Keller (2004). But I'm not persuaded by simple achievementism. One important difference is that the Instrumentality Factor focuses on *progress* towards the aim rather than realizing the aim as such, and also allows contribution to future aims count, as I just emphasized. In addition, I join many critics in thinking that when we exercise our agency in pursuing aims, *direction* matters as well as effectiveness, when it comes to excellence. It is better for me to be successful at solving the problem of mass-producing photovoltaic cells cheaply than at making a handwritten copy of the celebrity gossip website tmz.com, so as to have something to peruse in case I lose my Internet connection. So I take it that other things being equal, it is better for us to pursue goals that are objectively valuable. I'm not going to take a stand here on exactly which things are objectively worth pursuing, but presumably they include things like justice, artistic excellence, and scientific knowledge.

The second way in which actions at other times can affect teleological significance is *raising the stakes* of success or failure. Consider the following pair:

Diligent/Lucky

Diligent Ellie undertakes an expedition to Peru to find a full skeleton of a brachiosaurus. She puts together a detailed plan, informed by reading a vast literature on the topic, and goes through a laborious fund-raising process. After a painstaking search and two weeks of intense, backbreaking digging, she is in possession of a full skeleton of a brachiosaurus.

Lucky Florence undertakes an expedition to Peru to find a full skeleton of a brachiosaurus. Without a more specific plan, she hitches a ride to the airport. In the

lounge she meets a billionaire who gives her a lift in his private plane. Once in Peru, she heads straight for the closest bar. As she falls asleep on the patio, a mutt of indeterminate breed befriends her, and brings her a bone. Florence gives the dog a sausage left behind by another customer, and throws the bone in the back of an abandoned pick-up truck. This scenario repeats itself for two weeks. At the end of it, she is in possession of a full skeleton of a brachiosaurus.

Other things being equal, success in reaching the goal matters more to Diligent Ellie's well-being than it does to Lucky Florence's. It boosts the value of her life more than it does Florence's, and failure would be worse for her than for Florence. There is more at stake for her excellence as an agent. People have tried to cash out this sense in various ways – for example, for Gwen Bradford (2015), Ellie's success is more of an achievement than Florence's. My preferred idiom, however, is that of *merit*: since Ellie's success results from her developing and exercising her capacities and skills to a greater degree than Florence's success, which is due to sheer luck, her success is more merited. Merit, on this conception, is non-moral, gradable and relative to the realization of a specific aim. Plausibly, several different factors combine to determine the level of merit relative to an aim-realization, including the amount of effort the agent makes and the extent to which success manifests the agent's abilities rather than external factors. (There is a lot to say on this challenging topic, but I cannot go further here.)

The Diligent/Lucky case suggests the following principle:

Merit

Other things being equal, an agent's degree of merit intensifies the final value of success or failure in pursuit of her aims.

While Instrumentality shows how our present actions can affect the teleological significance of *past* actions and events, Merit shows how past actions can affect the teleological significance of *future* actions and events (beyond their causal role). Roughly, the harder you try, the more you have to gain or lose in the future. I don't claim that Instrumentality and Merit exhaust the factors that affect teleological significance – for example, the degree to which an agent identifies with an aim will also raise the stakes of success or failure – but they will do for my purposes here.

It is worth highlighting what's *not* on the list: the satisfaction of past preferences as such. This is a good thing, since it is very implausible that it would in any way benefit me to go on a rollercoaster ride on my future 50th birthday, if I will then dread the unpleasant prospect, even if I really *wanted* to celebrate the birthday that way when I was 11 (cf. Parfit 1984: 157). Things are otherwise if I've actually *pursued* a worthwhile aim, even if I no longer have it – if I've given up trying to publish my poetry collection, and you successfully bring it to the attention of Farrar, Strauss, and Giroux, there is something to be said in favor of what you're doing from the perspective of my self-interest, given Instrumentality. This distinction helps, in part, to clarify what's going on in genuine cases of sunk costs. Simply buying a ticket to a Televisionhead show is a borderline case of pursuing the aim of seeing the band, and the stakes of success are low, given Merit. Essentially, you just had a preference for spending tonight in a particular entertainment, even if you did something minimal to make its satisfaction more likely. That's why it won't make a noteworthy difference to the value of your past whether you use the ticket or not, which makes it a sunk cost.

4. Explaining the Common Intuitions

Armed with these simple principles for teleological significance as well as the Sunk Cost Principle I defended in Section 2, we can account for the prudential rationality of caring for

the past in the right circumstances. Roughly speaking, my claim about value was that beyond their effect on our experience, things we do and things that happen to us are good for us to the extent that they contribute to bringing about merited success in worthwhile aims we identify with.

So let's go back now to Jerry, our lawyer in recovery. Focus on the things he did in his wasted youth. Many times, he decided to take another swig or head for another dive. At the time, beyond the ephemeral pleasure that they gave him, they contributed to nothing but his spiral towards self-destruction. What happened, happened. Presumably, if he takes an unrelated job working for immigrants, it won't make a difference to the significance of the past. So if we look at the value of the different segments of his life, the picture looks like this: there's a bad part in the past, and good ones later. But since he has the options he does, he can now change the significance of the past actions and events. If he chooses to work with people who are now making the mistakes he left behind, he can draw on his experiences to serve them better than he would otherwise have. He can empathize more deeply and advise more effectively. It may be painful for him occasionally to be reminded of what he's ashamed of. But if those bad choices turn out to have been necessary for him to learn something important, he can take some pride in having made the best of them, and regret them less.

What makes his life story a redemption story is that his earlier mistakes turn out to contribute positively to future success. This makes his past better for him than it would have otherwise been. Here it seems clear that what it is for him to change the significance or meaning of his past *just is* for him to change its teleological significance. Redemption is, to be sure, only one example. But I think we can cash out other relevant kinds of meaning or significance in the same way. The sell-out scientist Sally's contamination story is a mirror image of Jerry's story. By undermining her own achievement, she makes it the case that her earlier efforts did not after all contribute to the realization of a valuable goal, except for a

brief and relatively insignificant while. It would be bad enough if someone else annulled her success. But the fact she does it herself intensifies the significance of her failure.

Does it matter for changing the value of past activity whether we make use of it to promote the original aim? On Dale Dorsey's (2017a) account, it does, since we can only benefit past selves (of our own or of others) by contributing to the success of past projects or satisfying past pro-attitudes. In contrast, Gilbert Harman observes that we do, as a matter of fact, sometimes adopt new ends that "help to rationalize and give significance to what we have been and are doing" (1976: 462), and seems to endorse this as a form of practical reasoning. I side here with Harman, as my treatment of Jerry's case shows. In his wasted youth, Jerry's project was getting wasted. If he makes use of what he learned in the course of that activity, he's *not* in any way completing that past project. Nevertheless, he can make his past meanderings worth something. Mutatis mutandis, the same goes for Monica Lewinsky. When she engaged in an affair with Bill Clinton, her aim wasn't to collect materials that would help in future anti-bullying campaigns. Nevertheless, by making use of her experiences in the service of later goals, she was able to partially redeem her earlier choices. This is evidence in favor of the Aristotle-inspired view over Dorsey's more subjectivist account. (I'll come back to other issues with his view below.)

How about the Two Awards case, then? Instrumentality won't suffice to account for such cases, since Sally's efforts will pay off either way. But only receiving the Classical award constitutes merited success, since only it is based on achievements that manifest Sally's skilled exercise of her capacities. Here her past exercises of agency increase the value of one outcome, making it better for her than an initially equally good one. The case is, to be sure, a somewhat contrived one, since we don't often get to choose between merited and unmerited success. It's perhaps not quite as rare, however, to choose between a demanding path and a shortcut to the same goal. In such situations, the demanding option typically has

opportunity costs (we could be promoting some other worthwhile aim, too), so it is overall not worth choosing. But in the absence of such opportunity costs, or just opportunity costs that are outweighed by increased merit, it may well be in our best interest to take on a challenge.[5]

5. Second Concern: Why Care About My Past?

Someone who accepts that I can change the significance and value of my past might still wonder why prudence requires (or even permits) caring about the past. Isn't it in my best interest to focus just on the present and future? In Dale Dorsey's (2018) terms, even if I *can* benefit my past self, why *should* I do so? Dorsey proposes an ingenious answer for the special case of project-related goods, with respect to which he argues prudence requires temporal neutrality. He starts with the thesis that "to achieve a project-related good at t, the success conditions of which occur at times later than t, requires cooperation between one's t-self and selves at times other than t" (2018, 1916). Given the need for cross-temporal cooperation, if I now embark on a project like knitting a sweater, prudential rationality requires being committed to taking the necessary steps in the future – taking a *cooperative attitude* toward my future self. According to Dorsey, this means that I now expect that my future self will "recognize the effort one's past self has put in, and cooperate for the sake of the success of the project" (2018, 1918).

Dorsey next observes that every present self is some future self's past self, so that "in adopting the cooperative attitude now, I am explicitly committed to my future self granting normative status to my present self—that is, my future self's *past self*—in rendering my current efforts a success rather than failure" (2018, 1919). Combining this attitude with a bias towards the present and future – in other words, giving little or no weight to my past self's projects – is "normatively unsavory" (ibid.). After all, it is parallel to normatively expecting

[5] As Bradford (2015: 95-7) observes, in a utopia in which no activity is instrumentally necessary, we're better off inventing games that require the exercise of our capacities.

someone to return a favor to me while at the same time refusing to return the very same favor to someone else.

While Dorsey's approach is characteristically inventive, I'm going to argue it suffers from two major problems: first, it does not satisfactorily answer the question, and second, the question itself is a bad one. First, on Dorsey's picture, as he frames it, when you take the cooperative attitude, you normatively expect your future self to give weight to your present project, and if you combine this with giving no weight to the projects of your past self, you have a set of attitudes that is "incompatible with what one owes to oneself" (2018, 1921). The problem is that Dorsey's argument trades on equivocation between the self as temporally extended and a momentary self. When he talks about relations among past, present, and future selves, which in his terms "mirror" interpersonal relations, he is committed to such selves being distinct entities – how else could attitudes like faith or normative expectation be possible? But when he talks about owing something to oneself, he tacitly switches to a temporally extended notion, since this "oneself" turns out to be the very same self now and in the future, as we'll see in a moment.[6] He can't have it both ways. This means his approach faces a dilemma: either the self is momentary, in which case the normative unsavoriness he identifies disappears, or it isn't, in which case the very puzzle he addresses dissolves, or so I'll argue.

To begin with the first horn, consider what his argument sounds like if we stick strictly to the terminology of distinct selves. The present-and-future biased t_0 self expects the t_{10} self to treat her investment in her project as a reason to complete it, but ignores the projects

[6] The same ambiguity or equivocation can be seen in Dorsey's talk of commitment. For example, he says "the prudentially rational agent will in fact be committed to one's future self taking the necessary steps to render one's currently-embarked-upon projects successes rather than failures" (2017a: 17). It certainly makes sense for a temporally extended agent to be committed to doing something in the future. But how could my *present self* be committed to my *future self* doing anything? This makes no more sense than my being committed to you doing something.

of the t_{-10} self. What does the t_0 self owe to herself? One thing is for sure: she is *not* her t_{-10} self (or, depending on your metaphysics, isn't her t_{-10} self any more), so she doesn't owe *herself* anything regarding the t_{-10} self. There's nothing *prudentially* unsavory about the combination of attitudes that the t_0 self has (assuming, charitably, that talk of prudence has a sense in the first place when it comes to time-slice selves). It is only if the present and past selves are one and the same that it makes sense to talk about the present self owing it to herself (and not to another self) to take her past self's – which is to say, her own – interests into account. Or, to put the point differently, if we conceive of intrapersonal relations as mirroring interpersonal ones, as Dorsey does, we lose the unity that is needed to identify a prudential tension among attitudes at or toward different times.[7]

Suppose, then, that we start instead with the assumption that your self is temporally extended – that your past, present, and future 'selves' are not really selves in the plural. It's just you at different times. What does this mean for the rationality of caring about the past? To get a grip on this, it's instructive to consider a parallel issue regarding intentions for the future. One distinctive role of intentions is intrapersonal coordination over time, which makes possible actions whose completion takes time by ruling out options that clash with one's plans (Bratman 1987). However, this key feature of intention leads some philosophers of action worry about what is sometimes called 'the problem of diachronic autonomy': how can we make decisions regarding our future actions without enslaving our future selves? (Ferrero 2010) How can it be rational for a later self to abide by the decisions of a past self rather than consider each situation anew (which would, of course, defeat the very purpose of forming intentions for the future)? If we formulate the problem like this, we see the same pattern as

[7] Could there be a *moral* tension instead? Although Parfit (1984) makes some suggestions along these lines, the idea is highly implausible. Time-slice selves that go in and out of existence instantaneously are far from the kind of subject to whom we have obligations, or who can be bearers of duties.

with Dorsey's account: there's an earlier self and a later self, and a question about rational relations between them.

However, as Julia Nefsky and Sergio Tenenbaum (forthcoming) argue, it is this very conceptualization of the issue that makes it appear that there is a problem. While many of those who worry about diachronic autonomy or self-governance over time officially recognize that our selves are temporally extended, they nevertheless slip into a time-slice conception of the self when they formulate the question. The problem with this, Nefsky and Tenenbaum point out, is that "On a time-slice conception, your relation to your past selves is of the same type as your relation to other people. So, your past self can only decide – or, settle a practical question – for you in the sense that someone else can settle a practical question for you." (forthcoming, no page number) They argue that once we articulate the question in these terms, we're faced with a 'puzzle' that is unsolvable, because there is no analogue between future-directed intention and its execution in our relations to others. But, most importantly for my purposes, they also emphasize that no puzzle arises, if we take the notion of a temporally extended self seriously. There's no philosophically interesting difference between deciding to call my grandmother right now and doing so, on the one hand, and deciding to call my grandmother after finishing this paragraph and doing so then, on the other. In both cases, the self who decides and the self who acts are the very same. As Nefsky and Tenenbaum put it, "To think that being moved to act directly by the intention I formed earlier would be a case of lacking autonomy is, again, to mistakenly treat my future self and past self as two different agents." (forthcoming, no page number)

My view is that the situation is just the same when it comes to the prudential rationality of caring for the past. Again, I am the same agent, the numerically same self I was before. There is no more a puzzle about why I should care about making my past actions more successful, say, than there is about trying to succeed in my present and future endeavors. I

should care for the significance of my past actions just as much and for the same reason as I should care about my present or future ones, unless there is some significant discontinuity in my life that does warrant serious talk of a 'past self' (or 'future self' for that matter). Once we recognize that we're temporally extended selves, the problem that Dorsey addresses does not arise. Asking why I should care about the good of my 'past self' is just asking why I should care about my own good at a past time – which should not generate any particular puzzle if we grant, as Dorsey does, that we can affect our good at a past time.

At this point, some might worry that this gives too much weight to the past. Surely we shouldn't be indifferent between equal past and present benefits to ourselves, but prefer the latter! Here, the first thing to bear in mind is that in the actual world, it's rare for us to face such a choice, since we can usually do much less now to affect the teleological significance of past actions than the significance of present or future ones, and we can't affect other aspects of the value of the past. But suppose we must choose between past and future benefits on a particular occasion. Maybe I was brought up in a weird religious cult, as a result of which my core project as a teenager was filling a lake with trash. However, as I came into contact with the outside world, I abandoned the project halfway through, and got into engineering recyclable photovoltaic cells. By some weird coincidence, I could now just as easily make either my current project or my past one successful (but not both). Other things being equal, isn't there a big prudential difference between completing the abandoned project and completing the one I'm currently committed to?[8]

My answer is that there is a difference by default, but it's because other things are typically not equal, not because past-directed benefits count for less. First, in scenarios like the cult case, the challenge gains intuitive force from the past aim being misguided or worthless. So imagine it's the other way around: I used to develop photovoltaic cells, but on

[8] I thank a reviewer for this journal for posing this challenge.

the verge of breakthrough, got converted by the cult, and now have the aim of filling the lake with trash. My claim is that it would be better for me to contribute to the success of my past actions by completing the project I abandoned for bad reasons, assuming the benefits of doing so aren't outweighed by ancillary costs (such as being ostracized by cult members).[9] If I chose success in the present project when it meant failure of the past one, it would be fitting for me to regret it. What this suggests is that in the original cult case, the explanation for why I shouldn't contribute to the past project is that there is *more worthwhile* aim I could now serve, and not the mere fact that the other aim is a *past* one.

What if the projects are on a par when it comes to value, however? Doesn't prudence still require me to work for the present aim? Yes, by default. First, it's clear that there are indirect benefits to completing the present project – for example, realizing a rejected aim won't bring me joy, and in fact might displease me (Dorsey 2018, 1921). But second, present plans may have a more direct prudential significance. As I noted, intentions can only play their role in temporally extended agency if they rule out conflicting plans unless the agent reconsiders. It has not proven easy to explain just why and when our aims have this sort of authority (see Ferrero 2010 for discussion). But as long as they do, I can't rationally be indifferent between the goal I *actually* have and a different goal I *could* have, even if they're equally worthwhile. (If I have evidence of a clearly better alternative, I should reconsider, and perhaps return to a project abandoned for a bad reason.) By definition, my past goals are no longer my actual goals. So by default, I display excellence in temporally extended agency by working for present rather than past aims, when there isn't a notable difference in value, merit, or likelihood of success. This explanation, again, doesn't appeal to the *pastness* of past aims, but simply to the default priority of my *actual* aims over other possible ones, whether past,

[9] Bear in mind that we're assuming there's sufficient psychological continuity to ensure that it is still clearly *myself* who would receive a past-directed benefit, not a past self from whom I'm now estranged.

present, or future. So even if present projects by default have prudential priority over past ones, this doesn't show that I shouldn't care equally about the teleological significance of my past actions, other things being equal. It just shows that by default, other things are not equal if there is a clash between present and past aims.

Next, let's come back to the question of why we should care about our own good *at any time* in the first place. I won't attempt a full answer, but I do want to point to a link between caring about the teleological significance of our actions in particular and various self-directed attitudes. I'll begin with *self-respect*. At least since Kant, philosophers have linked respect with rational agency and the special dignity it gives us. If we respect ourselves, we won't allow others to use us for their purposes, but insist on making up our own minds about which ends to pursue, and try to live up to these commitments (Dillon 1997). This is not compatible with undermining one's own past (or present) achievements for some future good that is not of comparable significance. (Would it be respectful towards someone else to undermine *their* achievements, past or future?) Someone who is willing to sully her own past manifests a kind of servility, unless she has changed her mind for sufficient reason about which ends are worth pursuing. More subtly, if we fail to take advantage of an opportunity to make our past pursuits count for more when it would amount to a net benefit, we distance ourselves from who we were earlier. I think this can also betray a lack of self-respect – my past self is after all just me at an earlier point in time, so if I have a sense of self-worth, I can't be indifferent to the success of my earlier pursuits and commitments any more than I can be indifferent to the success of my current ones. At least to this extent, then, respecting ourselves requires taking teleological significance into account.

Second, the teleological significance of our actions is what makes many self-directed attitudes warranted or fitting, so insofar as don't just care about how we feel, but also about having *good grounds* for our feelings, we must be concerned with it. Consider in particular

attitudes of pride, regret, and shame. They are evidently backward-looking, so whether they're fitting depends on the significance of our prior actions. It should be clear that pride and related third-personal attitudes like admiration track precisely the significance of what we've done or are about to do. You can be proud of your past pretty much to the extent that you've achieved merited success in realizing valuable goals. On the converse side, things are more complicated. If you ended up wasting your talents on something worthless, or never had a chance to pursue a worthy aim, you might rightly feel useless or sad, and possibly ashamed. Regret seems most clearly warranted when you had a better option available to you, but didn't take it. If you don't have reason to wish you had done otherwise or that your actions would have had different consequences, you don't have reason to regret. As I've argued, we can now change the significance of our past actions, and thereby either increase the strength of reasons to be proud of them or weaken the reasons we have to regret them. That means that insofar as we care about whether we can be justly proud of ourselves or avoid regret or shame, we have reason to care about past success or merit.

This relates to Derek Parfit's (1984) observation that we tend to be biased for the future when it comes to experiences like pain and pleasure: I'd rather have already gone to the dentist yesterday than face the prospect of going there tomorrow. Future pain counts for more than past pain. But he also noted that this bias isn't there when it comes to pride and shame: I'm just as mortified by the thought of having done something shameful in the past as by the thought that I will do something shameful in the future. Future shame doesn't count for more than past shame. Some philosophers argue that future bias is rational when it comes to experiential goods (Dorsey 2017b, Kauppinen 2018), while others have made a strong case it is never rational (Greene and Sullivan 2014). Either way, it's difficult to see why future bias would be rational when it comes to *non-experiential* goods, including agential success. If

that's the case, you rationally should care about having a better past just as much as a better future, if you can make a difference.

Conclusion

In this paper, I've argued for three main claims. First, what makes some past investment a sunk cost is simply the fact that we can no longer make a (noticeable) difference to its value for us. Insofar as practical rationality is not just a matter of having coherent preferences but also having preferences that are sensitive to what is genuinely good, it is irrational for our decisions to be influenced by sunk costs. However, second, some past investments are not sunk costs, since what we do now can make a difference to their teleological significance or contribution to our success as agents, which is one aspect of final prudential value according to both Aristotelian and sophisticated subjectivist views. And finally, I've just argued that since our selves are temporally extended, there is no special puzzle about why we should care about making our past better for us, when we can do so.[10]

References

Anderson, Elizabeth (1993) Value in Ethics and Economics. Cambridge, MA: Harvard University Press.

Angner, Erik (2016) A Course in Behavioral Economics. London: Palgrave Macmillan.

Aristotle, Nicomachean Ethics. Translated and edited by C. D. C. Reeve. Indianapolis: Hackett, 2014.

[10] I owe a debt to many people for useful comments on and challenges to various versions of this paper over the years, including (but not limited to) Erik Angner, Nomy Arpaly, Gwen Bradford, Donald Bruckner, Dale Dorsey, Dave Estlund, Tobias Fuchs, Jennifer Hawkins, Arto Laitinen, Julia Nefsky, Lilian O'Brien, L.A. Paul, Jason Raibley, Connie Rosati, Nicholas Smyth, Daniel Star, and Jussi Suikkanen.

Arkes, Hal R. and Blumer, Katherine (1985) 'The Psychology of Sunk Cost'. Organizational Behavior and Human Decision Processes, 35, 124-40.

Bradford, Gwen (2015) Achievement. New York: Oxford University Press.

Bradley, Ben (2009) Well-Being and Death. New York: Oxford University Press.

Bratman, Michael (1987). Intention, Plans, and Practical Reason. Stanford: CSLI.

Broome, John (1999) Ethics Out of Economics. Cambridge: Cambridge University Press.

Danto, Arthur (1962) 'Narrative sentences'. History and Theory 2 (2), 146–179.

Dillon, Robin (1997). 'Self-respect: Moral, emotional, and political'. Ethics 107 (2), 226–49.

Dorsey, Dale (2015) 'The significance of a life's shape'. Ethics 125 (2), 303–30.

Dorsey, Dale (2017) 'Future bias: a defense'. Pacific Philosophical Quarterly 98, 351-73.

Dorsey, Dale (2018) 'Prudence and past selves'. Philosophical Studies 175, 1901–1925.

Ferrero, Luca (2010) 'Decisions, diachronic autonomy, and the division of deliberative labor'. Philosophers' Imprint 10, 1–23.

Greene, Preston and Sullivan, Meghan (2015) 'Against time bias'. Ethics, 125(4), 947-70.

Harman, Gilbert (1976) 'Practical reasoning'. Review of Metaphysics, 29(3), 431-63.

Kauppinen, Antti (2012) 'Meaningfulness and time'. Philosophy and Phenomenological Research 84 (2), 345–377.

Kauppinen, Antti (2015a) 'The narrative calculus'. Oxford Studies in Normative Ethics 5, 196–220.

Kauppinen, Antti (2015b) 'What's so great about experience?' Res Philosophica 92 (2), 371–388.

Kauppinen, Antti (2018) 'Agency, experience, and future bias'. Thought 7 (4), 237-245.

Keller, Simon (2004) 'Welfare and the achievement of goals'. Philosophical Studies, 121, 27-41.

Kelly, Thomas (2004) ‘Sunk costs, rationality, and acting for the sake of the past’. Noûs, 38(1), 60-85.

Kolodny, Niko (2005) ‘Why be rational?’ Mind, 114, 509-63.

Kraut, Richard (2007) What Is Good and Why. Cambridge, MA: Harvard University Press.

Lin, Eden (2016) ‘How to use the experience machine’. Utilitas, 28(3), 314-32.

McMahan, Jeff (2002). The Ethics of Killing. Oxford: Oxford University Press.

Nefsky, Julia and Tenenbaum, Sergio (forthcoming) ‘Extended agency and diachronic autonomy’. In Carla Bagnoli (ed.) Time in Action: The Temporal Structure of Rational Agency and Practical Thought. New York: Routledge.

Nozick, Robert (1974). Anarchy, State, and Utopia. New York: Basic Books.

Nozick, Robert (1993) The Nature of Rationality. Princeton: Princeton University Press.

Parfit, Derek (1984) Reasons and Persons. Oxford: Oxford University Press.

Paul, L.A. (2014) Transformative Experience. New York: Oxford University Press.

Portmore, Douglas (2007) ‘Welfare, achievement, and self-sacrifice’. Journal of Ethics and Social Philosophy, 2(2), 1-28.

Raibley, Jason (2013) ‘Values, agency, and welfare’. Philosophical Topics, 41(1), 187-214.

Tiberius, Valerie (2008) The Reflective Life. New York: Oxford University Press.

Velleman, David (1991/2015) ‘Well-being and time’. Reprinted in Beyond Price: Essays on Birth and Death (Cambridge: Open Book Publishers), 141-73.