If There Are No Diachronic Norms of Rationality, Why Does It Seem Like There Are?

Ryan Doody

Abstract: I offer an explanation for why certain sequences of decisions strike us as irrational while others do not. I argue that we have a standing desire to tell flattering yet plausible narratives about ourselves, and that cases of diachronic behavior that strike us as irrational are those in which you had the opportunity to hide something unflattering and failed to do so.

1 Introduction

Suppose that you have a banana and I have an apple. You pay me a nickel to trade your banana for my apple. Then you pay me another nickel to trade back. You have the banana you started with and two fewer nickels. It seems like you've behaved foolishly. It looks like you've done something irrational.

Let's say that you *suffer diachronic misfortune* when you perform a sequence of actions resulting in an outcome that is worse, by your own lights, than some other outcome that would've resulted had you performed a different sequence of actions which was, in some sense, available to you. Suffering misfortune is unfortunate, but not necessarily irrational. That said, some cases of diachronic misfortune, like the example above, *do* strike us as irrational. The main objective of this paper is to explain *why*.

The explanation goes like this. Being practically rational, at the very least, involves being *instrumentally rational:* your preferences over actions (or, "means") should cohere with your goals (or, "ends").[1] Taking an action that does a worse job of furthering your ends is instrumentally irrational. I will argue that creatures like us—creatures who are deeply social—have, as

[1] This is *not* to say that being practically rational is *only* a matter of being instrumentally rational. It could be, for example, that certain ends are in themselves irrational irrespective of your beliefs and desires. Or it could be that certain combinations of attitudes are *ipso facto* irrational even if they don't lead to failures of instrumental rationality. I won't be taking a stand on that issue in this paper, however. All I am claiming here is that practical rationality requires instrumental rationality.

Res Philosophica, *Vol. 96, No. 2, April 2019, pp. 141–173*
https://doi.org/10.11612/resphil.1721

a matter of practical necessity, come to internalize a standing desire to construct flattering yet plausible autobiographical narratives about ourselves and our behavior. Constructing these sorts of narratives are, as a matter of psychological fact, important ends for creatures like us. Performing an action that doesn't serve this end as well as some other (when there are no overriding considerations) isn't instrumentally rational. And when you perform a suboptimal sequence of the sort that strikes us as irrational, there is some action that you've performed which didn't best serve your goal of constructing a flattering yet plausible autobiographical narrative.[2]

So, even if we are not rationally required to care about what we've done in the past or will do in the future, we often *do* care about these things. We care about what we've done and what we will do because we care about the kinds of stories that can be plausibly told about our diachronic behavior. And, furthermore, we can't help but care about this; our social nature has led us to internalize this desire, rendering it *inescapable* for creatures like us. And so the purely synchronic rational requirement to choose the available option that is prospectively best given *everything you now care about* gives rise to what appears to be a diachronic norm,[3] or so I will argue.

2 Time-Slice Rationality

There is a debate in decision theory and epistemology about whether or not there are fundamental, irreducible diachronic requirements of rationality.[4]

[2] Elsewhere (Doody Unpublished), I hypothesize that the cases in which we feel rational pressure to *honor sunk costs* are those in which it will be easier to integrate the action which honors sunk costs into a plausible autobiography according to which its protagonist has not suffered diachronic misfortune. In these cases, there will be an asymmetry in the prospects of spinning a plausible story that casts you in a good light; in the cases in which we don't feel pressure to honor our sunk costs, however, honoring sunk costs will make the prospects of telling an exonerating story just as dire as they would be were you to not honor sunk costs. The idea, which will be developed in more detail below, is that we have a standing desire to come across as the kind of people who would make good teammates. And the *ideal* teammate doesn't lose bets (because losing bets—even if they were rational bets to take—signals, perhaps unfairly, a vicious *rashness*) and doesn't have unstable preferences (because *fickleness* makes one's behavior problematically hard to predict).

[3] What about asocial creatures? Or what about agents about whom it is *stipulated* that they, for example, only care about money? Doesn't it also seem like these creatures can behave irrationally in virtue of falling afoul of a diachronic norm? Yes, but I think that our intuitions about such cases shouldn't be fully trusted. I'll hold off on discussing this point more fully until section 6.

[4] This debate bears, sometimes directly and sometimes indirectly, on several different issues in philosophy. In moral psychology, there are questions about the nature and normative status of *intentions* and other future-directed attitudes which govern the behavior of rational agents through time (e.g., Bratman 2010, 2012; Gauthier 1997; Holton 2009; Velleman 2000). In Bayesian epistemology, there are questions about the extent which various epistemic principles—like Conditionalization and Reflection—are motivated by Diachronic Dutch Book arguments (e.g., Briggs 2009; Christensen 1991; Levi 2002; Maher 1992; Schick 1986; Skyrms 1987, 1993; Teller 1976; van Fraassen 1995). Relatedly, there are issues in the foundation of

A requirement is *synchronic* if it tells you how things ought to be at a time, and a requirement is *diachronic* if it tells you how things ought to be across time. A requirement of rationality is *fundamentally* and *irreducibly* diachronic so long as someone can fail to satisfy it without violating any synchronic requirements.

Following Hedden (2015a, 2015b), let Time-Slice Rationality be the view that there are no fundamental, irreducible diachronic norms of rationality; all fundamental requirements of rationality are synchronic.[5]

There are compelling reasons on both sides of this issue. On the one hand, it certainly seems, in many cases, like there *are* irreducibly diachronic requirements of this sort. Take, for instance, the example which opens this paper: by trading your banana for my apple and then trading back, you suffer diachronic misfortune; but it's not obvious what, if any, synchronic requirement you've violated. Nevertheless, your behavior seems irrational. On the other hand, the existence of fundamental, irreducible diachronic rational requirements appears to conflict with a modest version of *internalism*, according to which behaving rationally is a matter of doing what makes the most sense to you, given your perspective. What I did, or believed, or cared about *last week* aren't facts about what I am *currently* doing, believing, or caring about. And insofar as rationality is concerned with how my actions, beliefs, and desires all hang together, what actually transpired in the past isn't relevant to what it's rational for me to do now. And so, if internalism about rationality is right, there cannot be any fundamental diachronic requirements of rationality.[6]

Bayesian decision theory about the extent to which its axioms can be justified by appealing to diachronic behavior in sequential decision problems (e.g., Davidson et al. 1955; Hammond 1976, 1988; Levi 1991; Machina 1989; McClennen 1990; Rabinowicz 1995; Ramsey 1926; Seidenfeld 1988; Steele 2010). There's an issue in game theory regarding the conditions under which a game in strategic form is equivalent to a game in extensive for (e.g., Seidenfeld 1994; Stalnaker 1999). Recently, several philosophers have addressed this question directly (e.g., Carr 2015; Ferrero 2012, 2009; Hedden 2015a,b; Meacham 2010b; Moss 2015).

[5] In addition, Time-Slice Rationality, as espoused by Hedden (2015a), holds that "your beliefs about what attitudes you have at other times play the same role as your beliefs about what attitudes other people have" (452). According to the view, the requirements of rationality are both synchronic and impersonal. This paper focuses more heavily on the former feature.

[6] See Hedden 2015a (and Carr 2015; Moss 2015) for more discussion on the motivation that internalism provides for Time-Slice Rationality. I think the most helpful way to see the point is to focus on the so-called action-guiding role of the rational 'ought.' Rationality, the thought goes, should provide us with some guidance about what to do. And if there are irreducibly diachronic requirements, rationality will have trouble offering us helpful advice, in some cases. Why? We have to decide what to do at a time. And in order for the rational requirements to be operationalizable—that is, for the advice they give to be useful—they have to make reference only to that which is, in some sense, accessible to me. For example, "Buy the winning lotto ticket" is good advice in the sense that, so long as I succeed in complying with it, I am guaranteed riches; but it is bad advice in that it is supremely unhelpful. I don't know how to succeed in taking the advice unless I know which ticket is the winner, and that's something that needn't be (and usually isn't) accessible to me. Moreover, that which is not encompassed in my current perspective will not be accessible to me at the time the decision is

There is a tension here. That being said, Time-Slice Rationality needn't be a *revisionary* thesis—that is, one that radically consigns many of our plausible first-order rational principles to the flames. Proponents of the view instead argue that much of the work done by diachronic requirements can be equally well done by synchronic requirements alone.[7] It's not my intention in this paper to argue for Time-Slice Rationality; maybe there are fundamental, irreducible diachronic norms and maybe there aren't. Instead, I will offer an explanation for why, in some cases, it *seems* like there are norms governing our diachronic behavior even if the proponents of Time-Slice Rationality are correct that there aren't any.

3 Sequential Choice and Diachronic Misfortune

I claim that the cases of diachronic misfortune that strike us as irrational are those in which one has the opportunity to act so as to *disguise* the fact that one has suffered diachronic misfortune but fails to do so. Let me bring this out by considering two structurally analogous cases of diachronic misfortune.[8]

> **Generous Game Show.**[9] You are on a very generous game show. There are two boxes before you: box *A* and box *B*. Box *A* contains an all-expenses-paid alpine skiing vacation, and box *B* contains an all-expenses-paid beach vacation. (And you know which

made. But diachronic norms, insofar as they are *irreducibly* diachronic, make reference to features—namely, features about the past or the future—that might not be accessible to my current perspective. For example, while the norm "Follow through on the plans you made yesterday" is genuinely diachronic, the quasi-diachronic norm "Follow though on the plans you currently believe you made yesterday" needn't be. The latter is, in the relevant sense, synchronic; it only makes reference to features (e.g., what you currently believe about what you previously did) that are accessible to your current perspective. (Thanks to an anonymous referee for suggesting this example.)

[7] For example, Hedden (2015b) argues that diachronic principles in Bayesian epistemology, like Conditionalization and Reflection, can be replaced with purely synchronic analogs without much loss.

[8] The phenomenon of diachronic misfortune is more general than what Hedden calls *diachronic tragedy*: cases in which you have attitudes that "lead you to act over time in a manner that is to your own acknowledged, predictable disadvantage" (2015a, 423). I'm interested in *misfortune*, not *tragedy*; your performance of a suboptimal sequence needn't be foreseeable. (In fact, several of the cases in Hedden 2015a are actually examples of what I call *diachronic misfortune*, and not diachronic tragedy.) Ending up in a suboptimal outcome, of course, is not ipso facto irrational; that happens every time we lose a bet, and it's certainly not always irrational to take bets. But as some of Hedden's examples bring out, it's not just *foreseeable* misfortune that strikes us as irrational. By focusing on the more general phenomenon, we can get clearer about which features our intuitions about diachronic rationality are sensitive to.

[9] This example is a variation on a case given in Hedden 2015a. In Hedden's version, you suffer diachronic misfortune because you have "imprecise preferences": your preference-ordering is incomplete.

box contains which prize.) The game has two rounds. In Round 1, you get to decide to place a $50 voucher in one of the two boxes. In Round 2, you get decide which box to take. The rounds happen quickly; as soon as you decide what to do in Round 1, you have to make your Round 2 decision.

Suppose that you'd be happy with either vacation, but you slightly prefer the beach vacation to the ski vacation. You decide to place the $50 voucher in box *B*. Round 1 ends and Round 2 begins. You have a change of heart—you think about how fun it would be to ski the slopes—and come to prefer the alpine ski vacation to the beach vacation. You decide to take box *A*.

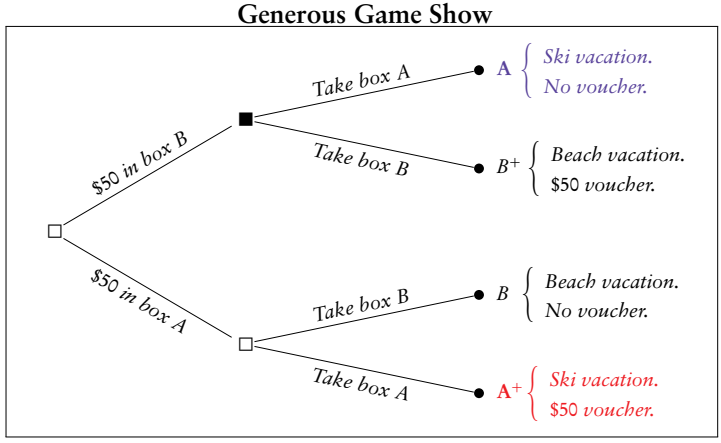

FIGURE 1. Generous Game Show tree-diagram.

There are a couple things to be said about this case. First, your diachronic behavior—putting the $50 voucher in box *B* during Round 1, then choosing box *A* during Round 2—seems *irrational*. There are, of course, ways of filling out the story so that your behavior no longer seems irrational. Imagine, for example, that after putting the voucher in box *B*, you receive an emergency call from your physician informing you that you're allergic to salt water. It no longer seems irrational for you to choose box *A*. All I'm claiming is that, absent details like these, your diachronic behavior strikes us as irrational.

Second, this is a case in which you've suffered diachronic misfortune—you've performed a suboptimal sequence of actions

$$\langle \$50 \textit{ in box } B, \textit{Take box } A \rangle$$

—without (seemingly) doing anything synchronically irrational. Given your feelings about the two vacations during Round 2, it's rationally permissible

for you to take box A over box B. But was it rationally permissible for you to put the \$50 in box B during Round 1? That depends on what you, during Round 1, believed you would do at Round 2. (If, for example, you were 100% confident at time t_1 that you'd take box A during Round 2, it would be synchronically irrational for you to put the \$50 voucher in box B.) It's rationally permissible for you to place the \$50 voucher in box B just so long as you are, at time t_1, reasonably confident that you will take box B in Round 2.[10] It is perhaps more accurate, then, to represent your predicament with the tree-diagram in Figure 2, which makes explicit the role uncertainty regarding your future preferences plays in your decision during Round 1.

In Generous Game Show, you place the \$50 voucher in box B at Round 1 because you prefer the beach vacation to the ski vacation and you are reasonably confident that your preferences won't switch in Round 2. Then, at time t_2, you learn that your preferences have changed: you now prefer the ski vacation to the beach. Acting on these preferences, you decide to take box A (and thus forgo the \$50 voucher). It looks like you've acted rationally at each time. But it also looks like you've done something diachronically irrational.

And now consider the following case.

> **Gamble Game Show.** You are on a game show similar in many respects to the previous one. There are two boxes before you: box A and box B. One of the boxes contains an all-expenses-paid cruise vacation, and the other box contains an all-expenses paid dude ranch vacation. But you don't know which box contains which prize. You (as well as the studio audience and the viewers at home) *do* know, however, that the host has rolled a six-sided die: if the die rolled a six, then the dude ranch vacation was placed in box A and the cruise vacation was placed in box B; otherwise, the dude ranch prize is in box B and the cruise prize is in

[10] How confident is "reasonably confident"? At time t_1 you slightly prefer the beach vacation to the ski vacation. Let p be your credence at time t_1 that by the moment of choice at Round 2 your preference will have shifted in favor of the ski vacation. It's rationally permissible to put the money in box B just so long as the expected utility of doing so is just as great as putting the money in box A. So,

$$\begin{aligned} eu(\$50 \text{ in box } B) &\geq eu(\$50 \text{ in box } A) \\ p \cdot u(A) + (1-p) \cdot u(B^+) &\geq p \cdot u(A^+) + (1-p) \cdot u(B) \\ (B^+ - B) &\geq p \cdot ((A^+ - A) + (B^+ - B)) \\ u(\$50) &\geq 2 \cdot p \cdot u(\$50) \\ \frac{1}{2} &\geq p \end{aligned}$$

You have to think it more likely than not that your preference for the beach vacation over the ski vacation will remain stable up to the moment of choice at Round 2.

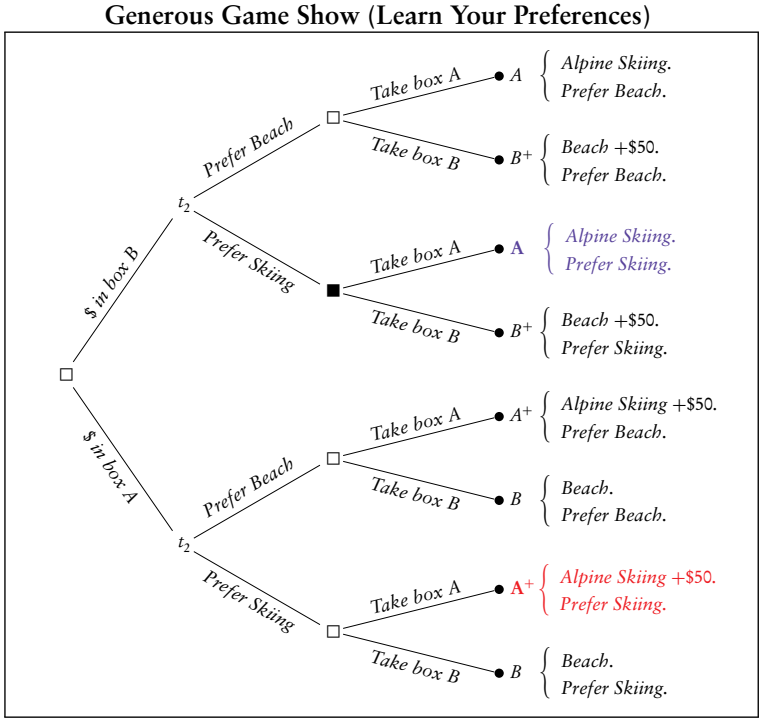

FIGURE 2. Generous Game Show tree-diagram (learn preferences at Round 2).

box *A*. Again, the game has two rounds. In Round 1, you get to decide to place a $50 voucher in one of the two boxes. Then, in Round 2, after learning which prize is in which box, you get to decide which box to take home.

Suppose that you slightly prefer the dude ranch vacation to the cruise vacation. Because you know that there is a five-sixths chance that the dude ranch prize is in box *B*, you decide to place the $50 voucher in box *B* during Round 1. You then learn—unfortunately for you—that the die didn't roll in your favor: the cruise vacation is in box *B* and the dude ranch vacation is in box *A*. You decide, in Round 2, to take box *A*—and, thus, forego the $50 voucher.

You've suffered diachronic misfortune, but it doesn't seem like you've acted irrationally in this case. What accounts for the difference?

In this case, just as in the previous one, you suffer diachronic misfortune by performing a sequence of actions resulting in an outcome that is worse, by your own lights, than the outcome that would have resulted had you

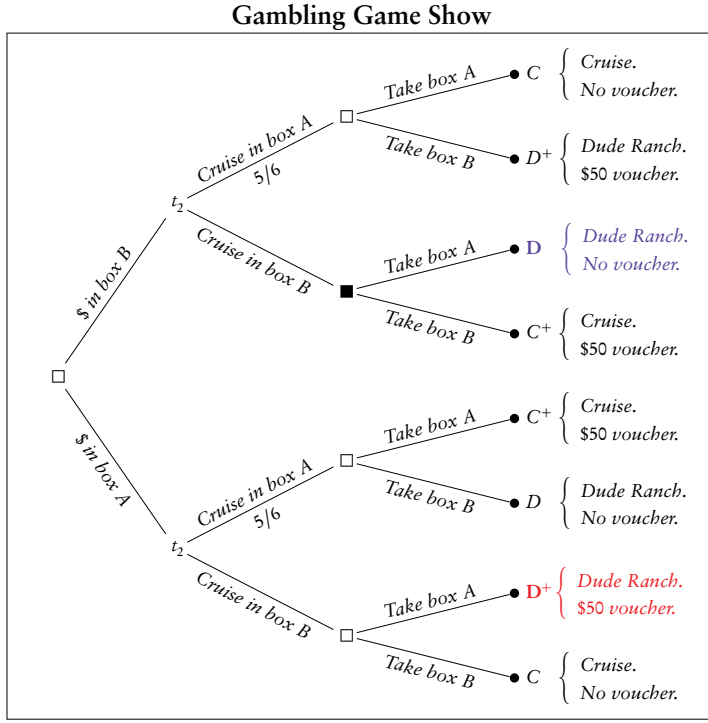

FIGURE 3. Gamble Game Show tree-diagram.

gone "down" instead of "up" at the first choice-node. In each case it was rational for you to go "up" at the first node given your beliefs and desires at that time. But we're inclined to judge your diachronic behavior in Generous Game Show more harshly than in Gamble Game Show. Why is that?

There are several potentially relevant disanalogies between the two cases that might account for the difference. One might think, for example, that suffering diachronic misfortune in Generous Game Show is irrational but not in Gambling Game Show because, in the former but not the latter, you've failed to follow through on an *intention* that you formed during Round 1 and it's irrational to fail to follow through on your intentions. But that needn't be the case. In both cases, you take a bet (broadly construed) and lose. In Generous Game Show, you haven't formed the *intention* to take box *B*; rather, you've made a *prediction* about what you will feel like doing in Round 2, and acted on the basis of that prediction. (In fact, we can imagine that you don't have strong feelings one way or the other about which vacation is better during Round 1. You place the $50 voucher in box *B not* because you right now prefer the beach vacation to the alpine skiing vacation, but rather because you right now *predict* that you will prefer

the beach vacation. When we *intend* to ϕ, generally, we prefer that our future selves ϕ whether or not our future selves *feel like doing so*, but your preferences in Round 1 needn't be like that.)

Here's a different thought. Although in both cases you've lost a bet, in Generous Game Show it's a bet that turns on what your preferences will be in Round 2, while in Gamble Game Show it's a bet that turns on which vacation prize is in which box. And one might think that your diachronic behavior in the former case, but not the latter, is irrational because beliefs about your future preferences are beliefs about *you*. And being wrong about yourself might seem closer to a rational failing than being wrong about some feature of the world—like how the die landed—that is entirely external to you. Maybe, one might think, this is because which box you take in Round 2 is something within your control, whereas which box contains which prize is not. But that doesn't seem right. Although which box is chosen in Round 2 is within your control *during Round 2*, it's at least not obvious that this is something under your control in Round 1.[11] (Furthermore, in this case, you haven't placed a bet on *what you will do* in Round 2; rather, you've placed a bet on *what your preferences will be* in Round 2. And it's even less obvious that we are able to exercise voluntary control over our future preferences.) More importantly, if Time-Slice Rationality is correct, it's unclear how a distinction like this could make a rational difference. On that view, facts about the preferences of your future self are external to you right now in much the same way as facts about which box contains which prize.

Instead, I claim that the relevant difference between these two cases—the difference that accounts for our inclination to judge your diachronic behavior more harshly in the former than in the latter—is that, in the former case but not the latter, you have at time t_2 the ability to disguise the fact that you've suffered diachronic misfortune. Taking the box in which you placed the $50 voucher is consistent with a story about you according to which everything is going your way (e.g., your preferences are stable, you understand yourself, you haven't lost any bets, etc.), whereas taking the *other* box is not consistent with a flattering story like this. On the other hand, in Gamble Game Show, there's nothing you can do to disguise the fact that you've suffered diachronic misfortune. Whether you take box *A* or box *B*, you reveal that you lost a bet (in the broadest understanding of the phrase); the studio audience and the viewers at home, given that they

[11] This brings up an interesting issue about *self-binding*. We've been assuming that your options during Round 1 only concern where to allocate the voucher. But if you have the ability to self-bind, your options during Round 1 might better be represented as follows: put-voucher-in-box-*A*-and-take-box-*A*, put-voucher-in-box-*A*-and-take-box-*B*, put-voucher-in-box-*B*-and-take-box-*A*, put-voucher-in-box-*B*-and-take-box-*B*. If these are your available options during Round 1, then opting for put-voucher-in-box-*B*-and-take-box-*A* is straightforwardly *synchronically* irrational. I will postpone further discussion of self-binding until section 5.

know it is reasonably likely that the dude ranch vacation is in box *B*, are able to infer from your decision in Round 1 that you must prefer the dude ranch vacation to the cruise vacation; and so, no matter what you do in Round 2, you reveal that you've suffered diachronic misfortune.

I am not claiming, however, that you can in one case but not the other *avoid* suffering diachronic misfortune by acting differently at time t_2; in both cases, you will suffer diachronic misfortune no matter what you do at that time. Rather, the difference comes down to whether or not you can act so as to *hide* your diachronic misfortune. In Generous Game Show, you can; in Gamble Vacation Game Show, you cannot. If you happened to care about hiding your diachronic misfortune—and I will argue that we *do* in fact care about this—that would give you a reason in the former case but not the latter to act differently at time t_2 than you did. And if this reason is sufficiently strong—strong enough to outweigh other relevant considerations—you've done something *synchronically* irrational at time t_2.[12]

If we have standing non-instrumental desire to hide our diachronic misfortune (when it's possible to do so), we can explain why, in some cases, it seems like there are irreducible diachronic requirements of rationality even if, in fact, there aren't any. But what exactly would such a desire look like? And why think this is something we care about?

4 Spinning a Flattering Social Story

I've claimed that we care about making it *appear as if* we've avoided diachronic misfortune when it's feasible to do so. (Of course, often a great way to ensure that it appears as if you avoided diachronic misfortune is to avoid diachronic misfortune in the first place.) The desire to *maintain plausible deniability* about having suffered diachronic misfortune can help to explain why we feel rational pressure to carry on with our past projects in some cases and not others.

First, allow me to explain what it is to desire to maintain plausible deniability about having suffered diachronic misfortune, and why it is that

[12] What about a variant of Gambling Game Show in which your decision about where to allocate the $50 voucher *doesn't* reveal your preferences over the vacations? Take, for example, a case in which no one, except for you, has any idea which box contains which prize. You have some private, personal evidence that box *B* contains the dude ranch vacation. And no one (including yourself) has any reason to think you have a strong preference for one of the vacations over the other. You slightly prefer the dude ranch vacation to the cruise vacation, so you put the $50 voucher in box *B*; it turns out that you were wrong about which box contained which prize. Would it be irrational for you to choose box *A* nonetheless? By taking box *B* you have a greater chance of hiding your diachronic misfortune. I contend that in cases like these, we would feel some rational pressure to take box *B*. It needn't be *irrational* to fail to do so, however. To put this differently, it's not irrational for you to *honor your sunk costs* by going home with the contents of box *B* (unless your desire to maintain plausible deniability about having suffered diachronic misfortune is outweighed by other considerations).

I think it's plausible to expect creatures like us—creatures who crucially rely on cooperating with one another—to, as a matter of psychological fact, have such a desire.

4.1 Signaling in the Social World

We live in a social world in which our choice-behavior is, very often the subject of examination by others. Navigating through the world involves interacting with each other. This, in turn, involves coming to understand what others believe, care about, and value and making sufficiently reliable predictions about their future behavior given this understanding. To get on with one another, we must construct rough-and-ready folk psychological theories of each other. These theories are based on our evidence about each other's choice-behavior.[13,14]

Often then, in addition to whatever else they do, our actions *signal* something about ourselves to others. Sometimes, in fact, the signaling-power of an action is so compelling that we're, ironically, disposed to perform it at the expense of undermining the thing we wanted to signal about ourselves.[15] Regardless of the power of the signal, all the decisions we make have the potential to communicate *something* about ourselves, no matter how weakly or defeasibly. When you opt for X over Y, you suggest—albeit defeasibly—that, all else equal, you prefer X to Y. And, if you've always opted for X over Y in the past, it wouldn't be unreasonable for an onlooker to predict that, all else equal, you'll opt for X over Y, again, in the future. If you care about what your choice-behavior signals about you, it's reasonable for you to take this into account when deciding what to do. Moreover, I think, as a matter of psychological fact, we *do* care

[13] Note that 'choice-behavior' is here being understood in its broadest sense so as to include, for example, *linguistic behavior*. What we say to one another is a major source of evidence—but not, by any means, the *only* source of evidence—about what what we believe and care about.

[14] The idea that rationality plays a crucial role in predicting and explaining behavior via attributing folk psychological states to each other has been developed and defended, among others, by Davidson (1973), Lewis (1974), Pettit (1991), and Ramsey (1926).

[15] I have in mind cases in which you, in some sense, want to signal that you care about X, but select an action that promotes X less effectively than another available action would because the former action increases the chances of reliably signaling what you want to signal than the latter. There are a number of interesting cases of this. There's the example in evolutionary biology of the male stalk-eyed fly, whose large eye span, it's been hypothesized, serves as a costly signal of fitness despite undermining it (Zahavi 1975). Another example is the *Prius Halo* (Sexton and Sexton 2014), which hypothesizes that the fact that the Toyota Prius dominates the hybrid car market is because of its distinctive (some might say 'unattractive') look. The idea being that environmentally conscious consumers choose to purchase a Prius rather than its competitors—even when those competitors are more attractive both financially and environmentally—because the Prius's unique look provides a stronger public signal of environmental consciousness. Robin Hanson is famous for analyzing a wide variety of large-scale social phenomena (e.g., Hanson 2008) in terms of signaling (see Simler and Hanson 2018).

about what we can sensibly expect our choice-behavior to lead a reasonable observer to conclude about us. *And*, I'll argue, this is something, given our nature as social creatures, we cannot help but care about.

4.2 Social Evolution and the Desire to Maintain Plausible Deniability

Social coordination is essential to our success as social creatures. Social coordination requires that I take you to be, and you take me to be, a good cooperator. In order to make myself appear like a good cooperator, I must present myself in a good light. Communities of successful cooperators are more successful than communities of unsuccessful cooperators. We can expect then that "traits" (broadly construed) conducive to successful cooperation will be "selected" for.[16] We've come to internalize the capacities, dispositions, and sentiments necessary for being decent cooperators in a social world.

I'm gesturing here toward a family of arguments familiar from evolutionary game theory.[17] A pattern of behavior is explained by, first, analyzing it in terms of a game-theoretic strategy, and then by showing that the strategy is evolutionarily stable under certain conditions. One way of interpreting these results is to understand the payoffs of the games plugged into the evolutionary dynamics *materially* and to understand the various strategies under consideration as corresponding to various *preference profiles* defined over those material payoffs. Consequently, we can understand the agents, who are the subjects of the evolutionary dynamics, as always acting *rationally* (i.e., they all perform the action that they most prefer from those available). Evolutionarily stable *strategies* will correspond to those preference profiles—or those ways of valuing material goods—that would be selected for (under the conditions specified elsewhere in the model). In this way, these sorts of argument in evolutionary game theory can be thought of as explaining how, and under what conditions, certain motivational features (e.g., certain desires, norms) can become *internalized* by agents.

In order to cooperate effectively—and, more generally, in order to successfully coordinate with each other—we must be able to reliably make fairly accurate predictions about both the future behavior of others and ourselves. We have to make these predictions, often, on the basis of somewhat meager evidence. Consequently, we have reasons to present to each other *coherent narratives* of ourselves—that is, we have reasons to act so that a

[16] I put 'traits' and 'selected' in scare quotes in order to indicate that the evolutionary mechanism at work here needn't be that of *biological evolution*—and so needn't involve phenotypic information transmitted reproductively—but, are more plausibly, the work of *sociocultural evolution*—in which norms, values, and general social information is transmitted culturally (McGeer 2001, 2007; Ross 2005). The characteristics under discussion here are more *memetic* than *genetic* (but could, of course, be both).

[17] See, for example, Axelrod 1986; Binmore 1998; Gintis 2000; Frank 1987; Maynard Smith 1982; Skyrms 2004, 1996; and Young 1998.

competent observer would be able to make fairly accurate predictions of our future choice-behavior on the basis of our past choice-behavior.[18]

Of course, making oneself *predictable* to oneself and others is not by any means the *only* characteristic that the social evolutionary pressure to successfully cooperate might inculcate. Maximally attractive prospective teammates, for example, are—in addition to being stable—not overly prone to taking losing bets. In short, to make oneself into an attractive candidate for social collaboration, one must avoid the stench of failure (Baumeister 1982; Baumeister 1999; Schlenker 1980; Trivers 2000).

Suffering diachronic misfortune, although not an infallible indicator of irrationality, *is* an indicator of failure. Here's why. There are two main ways to suffer diachronic misfortune: one, you take a gamble (in a broad sense), and lose; or, two, you exhibit diachronically unstable choice-behavior (as if in response to a preference shift).

Consider way two. By exhibiting diachronically unstable choice-behavior, you make yourself hard to predict.[19] If you're hard to predict, you're hard to coordinate with. If we can't coordinate with you, you will make a less-than-ideal teammate. There's pressure on us, then, to present ourselves in ways that uphold the *appearance* of consistency (Cialdini 2001; Stone et al. 1997; Swann 1985; Tedeschi et al. 1971).[20]

Consider way one. By taking a gamble and losing, you risk revealing that you made a bad prediction. Of course, it's not necessarily *irrational* to lose a bet—so, a team's *pro tanto* desire to not be associated with bet-losers might seem like a matter of superstition[21]—but given the meager amount of information we have about each other's behavior, it's difficult to determine

[18] The relationship between narrative, folk psychology, and the construction of "the self" has been explored in philosophy (e.g., Dennett 1992; Velleman 2005) and in cognitive science (e.g., Goldie 2012; Gazzaniga 1998; Hutto 2007; Ross 2005). A common theme throughout is the importance of the role that narrative plays in social coordination, which often requires presenting a unified account of our behavior.

[19] Diachronically unstable choice-behavior is difficult to rationalize as the product of coherent beliefs and desires had by a unified agent, who cares about things in ways that we around here find intelligible. It's not difficult, in general, to rationalize an agent's behavior if we are allowed to individuate the outcomes of the decision-problems the agent faces as finely as need be—which amounts to representing the agent's preferences as sensitive to those features individuating the outcomes (see, e.g., Broome 1993; Dreier 1996; Pettit 1991). But we rescue the unified agent's coherence at the expense of representing her as caring about things that we might find hard to understand. Either way, our ability to predict the agent's behavior suffers.

[20] What counts as diachronically consistent is a more complicated matter than I'm letting on. One can suffer diachronic misfortune as the result of diachronically unstable choice-behavior in a way that doesn't make one's future behavior hard to predict. For example, *predictable* preference shifts—like those that standardly occur as we mature, or like those that typically accompany significant life changes—in virtue of being predictable, needn't undermine our ability to coordinate with each other. More will be said about this in section 5.

[21] Whether or not this is a matter of superstition, it appears to be a real phenomenon. We are often judged by our success and failures, even when they are the product of chance. For example, dealers at casinos are sometimes removed from their posts, and even fired, after suffering a sufficiently long streak of bad luck (Goffman 1967).

whether your decision to take the gamble was a rational one. We want teammates who are good at assessing their evidence and who appropriately account for risk. As the number of bets you lose increases, the likelier it seems that you are failing on these fronts. This provides you with a reason to hide your losses when it's easy to do so, even if you lost a bet that was rational to have taken given what you knew at the time.

Moreover, it is *particularly* bad to reveal that you've lost a bet that turns on how you will feel, what you will do, or what your preferences will be, and the like. When making a prediction about yourself, it's presumed that you have a privileged position with respect to the relevant evidence, and it's often particularly opaque to others exactly what this evidence specifically is. The more private your evidence, the more vulnerable you are to charges that you failed to assess it correctly. And, furthermore, by revealing that you've made a bad a prediction about yourself, you reveal that you aren't predictable even to yourself. And, as prospective teammates might very well worry, if you aren't predictable to *yourself*, what hope is there for the rest of us? Someone who is bad at predicting what they themselves will do is someone for whom it's reasonable to think it will be difficult for the rest of us to predict as well.

If you've suffered diachronic misfortune, there is nothing you can do to change that. It might yet be possible, however, for you to avoid *signaling* to others that you have, and thus avoid acquiring the reputation as a subpar teammate. So, insofar as there is social evolutionary pressure to cooperate with one another, there is likewise pressure to present oneself as an attractive teammate. Maintaining plausible deniability about having suffered diachronic misfortune (i.e., acting so that your choice-behavior can be woven into a flattering self narrative) is instrumental in presenting oneself as an attractive teammate. And so it's not unreasonable to expect a process of social evolution to instill in social creatures like us a deep-rooted desire to maintain plausibility about having suffered a diachronic mistake. Because evolution doesn't paint with a fine-brush, we've come to internalize this desire as a *non-instrumental* one.

Here's an analogy. I have, as I'm sure you do too, a pro tanto desire for things that taste sweet. When pushed, I cannot offer a satisfying justification of the reasonableness of this desire. I don't, for example, desire sweetness as the means to some end. I simply like things that taste sweet. I'm hard pressed to say much more than that. It isn't, though, *mysterious* why I, and creatures like me, desire things that taste sweet. Most things that are sweet contain sugar. And sugar has fitness-promoting caloric properties. Creatures who desired sweet things did better than creatures who didn't. Although NutraSweet doesn't contain the fitness-promoting caloric properties of sugar, it still tastes sweet to me. And although (granting the evolutionary story I've sketched) the reason, in some sense, that I non-instrumentally desire sweetness has to do with the caloric properties of sugar, it isn't unreasonable to desire NutraSweet. I think, in some important

respects, our desire to maintain plausible deniability about having suffered diachronic misfortune is like my pro tanto desire for sweet foods.

In addition to this speculative story for why it might be that we'd come to internalize the desire to spin flattering yet plausible autobiographical narratives, there is a fair amount of empirical evidence that we do, as a matter of psychological fact, care quite strongly (albeit, not always consciously) about our self-presentation.[22] For example, Kurzban and Aktipis (2007) propose that humans have internalized a set of cognitive mechanisms, which they call the Social Cognitive Interface (SCI), that is "designed for strategic manipulation of others' representations of one's traits, abilities, and prospects" (131). These mechanisms, they argue, are the result of competition for partnerships, group memberships, and other positions of social value (Cosmides 1989; Kurzban and Leary 2001; Levine and Kurzban 2006; Tooby and Cosmides 1996). The mechanisms are designed to strike the optimal balance in self-presentation between favorability and plausibility (Baumeister 1982; Schlenker 1975). In particular, one primary function of these mechanisms is to maintain the appearance of consistency (Stone et al. 1997; Swann 1985; Tedeschi et al. 1971).[23] Furthermore, although these mechanisms serve a social function, there's evidence that the mechanisms exert motivational force on us even in private (Baumeister 1982; Briggs 2009; Shrauger and Schoeneman 1979; Tice and Baumeister 2001). We come to see ourselves as we imagine others see us. And so we don't, for example, stop caring about our (flattering yet plausible) social story when we know no one else is looking. In order to effectively convince others, we often must convince ourselves.

4.3 Plausible Deniability

For you to maintain plausible deniability about something, you have to construct a narrative about your behavior that's *plausible*. But what is it for a narrative to be plausible? And for whom are we constructing our narratives?

[22] For an accessible summary of the evidence from evolutionary psychology, see Kurzban 2010. See also Simler and Hanson 2018, which draws on work in microsociology, social psychology, primatology, and economics to argue that, because of our social nature, we've evolved to disguise various "ugly" motives as "pretty" ones (both to others and ourselves). We have, they argue, hidden motives to signal that we have certain socially valuable attributes. They then argue that a wide variety of large-scale social phenomena (e.g., charity, education, healthcare, religion, politics) can be fruitfully illuminated by taking these hidden motives seriously. My claim is consistent with theirs, but is more modest. All I'm claiming is that—for reasons similar to theirs—we've come to care non-instrumentally about our self-narratives.

[23] Kurzban and Aktipis (2007) say, for example, that "one important design feature of the SCI is to maintain a store of representations that allow consistency in one's speech and behavior that constitute the most favorable and defensible set of negotiable facts that can be used for persuasive purposes" (135).

4.3.1 Ways of Hiding Your Diachronic Misfortune

You will not be able to construct a plausible narrative about your behavior according to which you haven't suffered diachronic misfortune when it is *obvious* that you've taken an action that has resulted in an outcome *O* which is sub-optimal relative to an outcome that's diachronically accessible to you. For example, in Generous Game Show, the outcome in which you prefer the alpine ski vacation to the beach vacation and take the box containing the ski vacation plus the $50 voucher is *obviously* better than the outcome in which you prefer the ski vacation to the beach vacation and take the box containing *only* the ski vacation (and no voucher); and, in Gamble Game Show, the outcome in which you enjoy the dude ranch vacation plus $50 is *obviously* better than the outcome in which you go to the dude ranch without the money. When you bring about these outcomes, then, you reveal your diachronic misfortune.

If you want to tell a plausible story according to which you haven't suffered diachronic misfortune, there are two ways to do it. First, if it is obvious that *O* is sub-optimal, you might yet be able to maintain plausible deniability by misrepresenting *O* as some other outcome. This can be accomplished if the state of the world that partially constitutes *O* is suitably non-public. Take, for example, Generous Game Show. Given that you prefer a ski vacation to a beach vacation, it's *obvious* that you would prefer a ski vacation plus $50 to a beach vacation plus $50. But, because your preferences over vacation destinations are *non-public*, you might be able to hide your diachronic misfortune by hiding that your preferences have shifted by opting to take box *B*. A story according to which you place the $50 voucher in box *B* and then take box *B* is a story that's consistent with you being on the best-of-all branches of the decision tree.

Second, if it is obvious that outcome *O* is the outcome your actions have brought about, you might yet be able to maintain plausible deniability by disguising the fact that you prefer a diachronically accessible outcome to *O*. Here's an example. Suppose you are invited to your friend's cocktail party. You believe that your idol will be in attendance, and so you rent a suit to wear in an effort to impress her. You then learn that she won't be there. It wouldn't be odd to dress up for such an occasion, but had you not already rented the suit, you'd slightly prefer to dress more casually. By wearing the suit you can hide that you've suffered diachronic misfortune so long as it's plausible—as it very well might be—that you all along preferred to wear a suit to the party. Although it will be obvious that you are wearing a suit to a function at which your idol is not present, it needn't be obvious that this is a sub-optimal outcome.

In Gamble Game Show, however, you cannot tell a plausible story according to which you haven't suffered diachronic misfortune. You prefer the dude ranch to the cruise throughout, but are uncertain (at time t_1) about which box contains which prize. You believe the dude ranch vacation to

be in box *B*, so you elect to put the $50 voucher in box *B* during Round 1. You learn at Round 2 that you were mistaken: the dude ranch vacation was in box *A*. What you learn at time t_2 is public: you cannot hide the facts about which prize is in which box. Furthermore, the basis on which you made the decision to put the $50 in box *B* during Round 1 was also public: everyone knew it was likely that box *B* contained the dude ranch vacation, and so you cannot hide which prize you preferred.

4.3.2 Plausibility

What makes a story about your behavior *plausible*? In order for the narrative to be plausible, it's not enough that your diachronic behavior merely meet some formal constraints. The story must also attribute attitudes to you that seem reasonable. What counts as *plausible* will depend on the kinds of things that we around here consider to be relatively natural to care about.

Here's an example. Imagine you are going on a camping trip. The forecast calls for rain, so you rent an expensive raincoat. When you get to the campsite, however, it becomes clear that it is not going to rain. You could still wear the raincoat, but you opt not to do so. You've suffered diachronic misfortune—given how things turned out, it would've been better overall had you not rented the raincoat in the first place—but it's clearly not irrational to not wear the raincoat unnecessarily, and there is no rational pressure whatsoever to do so. You would reveal that you've suffered diachronic misfortune whether or not you wear the raincoat. There is no *plausible* story about you according to which you rent the raincoat, it doesn't rain, you wear it anyway, and you haven't stumbled into a suboptimal outcome. The weather is public, so you cannot disguise your diachronic misfortune in the first of the two ways discussed above. Furthermore, it's not reasonable—given the kinds of things that we around here care about—to take you to *prefer* wearing a rented raincoat unnecessarily to enjoying the sunny day having never rented the raincoat in the first place. People don't wear raincoats on sunny days.

A similar point holds in Gamble Game Show: there is no plausible story about you according to which you put the $50 voucher in box *B*—thus revealing that you prefer the dude ranch to the cruise—and then discover that the cruise prize is in box *B*, and take the contents of box *B*, and yet haven't stumbled into a sub-optimal outcome. Which prize is in each box (in Round 2) is public, and it's not reasonable to take you to prefer $50 plus the cruise vacation to $50 plus the dude ranch vacation.

4.3.3 The Audience

For whom are we constructing these narratives? Our stories are partially directed toward the other members of community, and partially directed toward ourselves. As a heuristic (because it is not always possible to tell who's watching when), we might find it helpful to pretend that there is a semi-omniscient God, whose epistemic access to us is not different in kind or grain from the access afforded to our communities' members, watching us at all times. We've come to internalize the desire to weave our diachronic behavior into a flattering yet plausible self-narrative non-instrumentally. And so, the thought goes, we feel the force of this desire whether we know that our behavior is the subject of public scrutiny or not. The claim is that social evolution has imbued us with a motive to act almost as if we're always being watched. And, insofar as we are often both the authors of and the *audience* to our own behavior, there is a sense in which we *are* always being watched.[24]

4.4 Rational Agency

The view I've sketched shares some important similarities with Velleman's account of agency (see Velleman 2009, for example). Velleman holds that it is constitutive of agency—or rather, constitutive of *action*, of which agents are the authors—to aim for *intelligibility:* that is, to understand what one is doing when one acts, and to act in a way that makes sense (to oneself and others).[25] Our standing desire to make ourselves intelligible is, for Velleman, an *inescapable* one: without it, our actions wouldn't be actions and we wouldn't be agents.

I agree in several respects with the spirit of Velleman's view, but my position diverges in two crucial ways. First, I'm arguing that we have a standing desire to act in ways that can be felicitously woven together into a flattering self-narrative. Spinning such a story involves, in part, acting in a way that is intelligible, but it requires more than this, too. We want our self-narratives to be *flattering* and that goes beyond mere intelligibility. The story of a hapless loser might be intelligible, but it's not flattering.

[24] The idea is that we can, and often do, adopt an outsider's view of our own actions, and evaluate ourselves and our behaviors from that perspective (Smith 1759; Hogan and Briggs 1986).

[25] Velleman (2009) draws an analogy between social interaction and theatrical improvisational acting. Improvisational actors try to do what makes the most sense given the character they are aiming to enact. We are, the thought goes, akin to improvisational actors enacting ourselves. Action is, according to Velleman, a kind of self-enacting performance. (See also the subtle discussion of social behavior in Goffman 1959, which also analyzes social interaction as analogous to theatrical performance. Social interaction is akin to a performance in which "actors" create and manage the impressions they impart to their "audience.")

Second, I don't think this standing desire is a *constitutively* inescapable aspect of agency. Mosquitos are arguably agents, for example, but their actions are not guided by such a motive. Rather, the desire is inescapable for us in a much weaker sense—a sense akin to Velleman's notion of *natural inescapability*. Given the kind of social creatures we are, it's reasonable to think that we've internalized this desire as way of getting along with one another. And, furthermore, because the desire gives rise—in some sense—to *who we are* as people, the desire becomes implicated deeply into our self-identities. So, if we didn't have the desire, it's not that we'd cease to be *agents*. Rather, if we didn't have the desire, we'd cease to be recognizable as anything like the deeply social agents we are.

My view also shares certain similarities with Ruth Chang's account of agency: *Hybrid Voluntarism* (Chang 2009, 2013, 2017). On her view, there are two kinds of reasons: *given* reasons (i.e., considerations that are reasons in virtue of something other than your own agency) and *will-based* reasons (i.e., considerations that are reasons in virtue of some act of will). If your given reasons fail to fully determine what to do, you can create a new will-based reason by "putting your agency behind" some feature of one of your options. In particular, if your given reasons are in equipoise (i.e., you fail to have more, less, or equal reason to choose one thing over another), you can, through an act of will, determine what you have most all-things-considered reason to do. For example, suppose you're choosing between the alpine ski vacation and the beach vacation. And suppose that your given reasons have run out: you don't have more, or less, or equal reason to choose one over the other. By focusing on one of the distinctively valuable features of, say, the beach vacation (e.g., the pleasant way the sand will feel beneath your feet), you can create for yourself a decisive will-based reason to choose it. In so doing, you constrain your future choices by making yourself into a certain kind of person. In Generous Game Show, if your given reasons with respect to the prizes have run out, then you might, by choosing to put the voucher in box *B*, thereby create a new will-based reason to choose box *B* when the time comes.

My view is different from (although consistent with) Chang's. On my view, you have a standing desire to spin a flattering yet plausible autobiographical narrative. What you do during Round 1 affects what you should do during Round 2—not because of some newly created will-based reason—but rather because it constrains how this desire can be satisfied. By putting the voucher in box *B*, you endow the option *take box B* with a property it wouldn't have had otherwise—namely, the property of being integratabtle into your flattering yet plausible narrative. Because you care about your narrative, this gives you a reason to take box *B*. But this isn't a reason directly created by an act of will. Moreover, as mentioned in section 3 you might treat your decision in Round 1 as one between gambles that turn on what you'll feel like doing in Round 2. And so you might put the voucher in box *B* because you predict that you'll feel like taking

it, not because you've willed yourself a new reason to. In fact, your given reasons needn't have run out: if you want future-you to get what future-you wants, and you predict that future-you will want box *B*, your given reasons support putting the voucher in box *B*. On my view, you'll nevertheless have reason to take box *B* (even if your prediction didn't pan out). On Chang's view, however, you wouldn't; you haven't created a new will-based reason to choose box *B*. Finally, on Chang's view, it's unclear to what extent your past choices constrain your future ones. Your commitment to a course of action can be undone, and you can *drift*; you can plump for one course of action over another without creating a new will-based reason. On my view, however, your past choices affect your future ones whether or not you've "put your agency behind" them. On my view, what matters is your concern for your self-narrative, not some private exercise of your will (although that might matter, too).

5 Does the Explanation Show Too Much?

Here's a potential problem. There are examples of diachronic behavior, very similar to those that strike us as irrational, which *do not* strike us as irrational. The worry is that the explanation I've given cannot be correct because it overgenerates; it predicts that certain cases of diachronic misfortune should strike us as irrational that, as a matter of fact, do not.

In this section, we'll look at one such case, and I will argue that my explanation doesn't make such a problematic prediction after all. And, in fact, the explanation I've given in terms of the standing desire to maintain plausible deniability about suffering diachronic misfortune nicely explains the asymmetry between those cases of diachronic misfortune which strike us as irrational and those that don't.[26] Consider the following case of diachronic misfortune (borrowed from Sartre 1946):

[26] Moss (2015) recognizes the possibility of asymmetries of this kind in cases in which agents forego sure gains because of a change of heart. She suggests that these asymmetries can be better accommodated if there are no genuine diachronic requirements of rationality than if there are. And so they ultimately provide a point in favor of Time-Slice Rationality.

She says, "It is not clear that agents in these situations [like the one described in the Sartrean story below] are strictly forbidden from changing their minds. In fact, we are intuitively disposed to forgive some agents who forego sure money, even when their change of heart is not prompted by any change in their evidence" (186). But also acknowledges that our intuitions sometimes go the other way, pointing out that:

> In general, we are most inclined to reject apparent mind changing as irrational when it happens quickly, unreflectively, repeatedly, or for strategic reasons. These intuitions can be comfortably accommodated by a theory according to which changing your mind is not itself impermissible, namely because the salient features of these cases may provide evidence that they do not involve the same sort of genuine changes of mind exhibited by agents in [those cases in which we're disposed to be more forgiving, like the Sartrean story]. By contrast, it is more difficult for blanket injunctions against mind changing to accommodate the intuition that changing your mind can sometimes be okay. (186)

The Sartrean Sequence. You have to choose between fighting the Nazis or tending to your sick mother. There are pros and cons to each. You care about various things, and you haven't a clue as to how to weigh them off against each other. You ask your French philosophy professor for advice, but he's no help. You decide to fight the Nazis. You complete your basic training, but then you reconsider and return to your mother.

The Sartrean Sequence

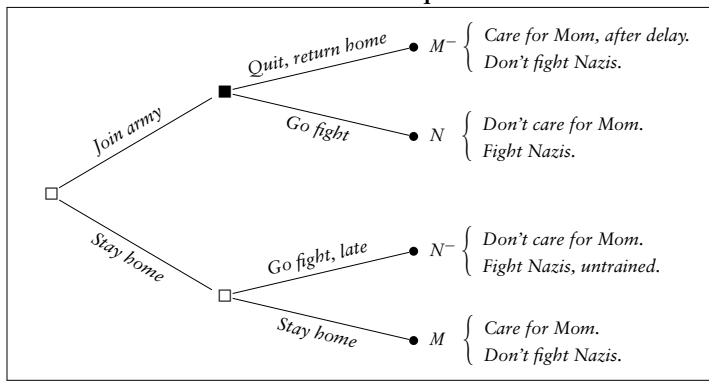

FIGURE 4. The Sartrean Sequence.

You suffer diachronic misfortune by performing the sequence

$$\langle \textit{Join army}, \textit{Quit, return home} \rangle$$

which results in an outcome (M^-) that is clearly worse than the outcome that would've resulted (M) had you performed the sequence

$$\langle \textit{Stay home}, \textit{Stay home} \rangle$$

instead. It's better to stay with your mother from the get-go than it is to stay with your mother only after abandoning her while away at basic training.

Despite the structural similarities between this case and Generous Game Show, we're inclined to judge your behavior more harshly in the latter than in the former. Furthermore, an argument analogous to the one sketched for why your behavior in Generous Game Show strikes us as irrational (i.e., that we have standing desires to spin flattering self-narratives about

Absent some story about how Time-Slice Rationality *can* plausibly accommodate these intuitions, it's not clear to me why we should expect its opponents to have a more difficult time making sense of them, especially in light of the point made in Carr (2013) that one can be *blameless* and nonetheless irrational (50–55). Opponents of Time-Slice Rationality can accommodate the asymmetry by claiming that we're inclined to be more forgiving in some cases than in others, even though a diachronic norm has been violated in both.

our diachronic behavior, and that end is better served by taking the box that contains the $50 voucher) seems to go through equally well in Sartrean Sequence. After completing basic training, when you are deliberating about whether to continue on with the army or to return home to care for your mother, you know that by returning home you give up any hope of telling a plausible story about yourself according to which you've avoided suffering diachronic misfortune. On the other hand, you also know that if you continue on with the army, there *is* a flattering (and plausible) self-narrative that could be told: a story according to which you decided to join the army to fight Nazis, and then went off to do so.

If we have the standing desire to maintain plausible deniability about suffering diachronic misfortune, and the desire is operative in Generous Game Show as well as in **Sartrean Sequence**, why are we inclined to be more forgiving about your behavior in the latter than the former? There are three important differences between these two cases, each of which affects the force of the desire to avoid revealing diachronic misfortune in the latter case.

Difference 1: Stakes. Here's one important difference between the cases. In Generous Game Show (the case in which we're prone to be less forgiving), the stakes are relatively *low:* very little hangs on what you end up doing. You'll be going on a vacation, which you'll find enjoyable no matter what you do. In Sartrean Sequence (the case in which we're prone to be more forgiving), however, the stakes are relatively *high:* what you ultimately end up doing matters a great deal. Your decision about what to do affects other people who matter a great deal to you. What you do matters to your mother, and it matters to your compatriots. This is a decision about which we might think some handwringing is appropriate—compulsory, even.

I contend that this difference in stakes, in part, accounts for our inclination to be more forgiving in the one than the other. It's *not* synchronically irrational to fail at maintaining plausible deniability about having suffered diachronic misfortune when your desire to do so is *outweighed* by other considerations. Your desire to maintain plausible deniability is only one among many, and it is only irrational to fail to do what you most prefer to do *all things considered*. When the stakes are relatively high, the potential satisfaction or frustration of this desire for a flattering self-narrative is just a drop in the deliberative bucket, quite possibly lacking the power to tip the scales.

Moreover, in Sartre Sequence, the stakes are high in a particular way: they're morally weighty. It seems *morally* inappropriate—selfish, or at least viciously self-regarding—for your desire for a flattering self-narrative to outweigh considerations of significant moral importance. Not only might the reason this desire provides fail to tip the scales, it might fail to be a reason of the right kind. Suppose, after completing basic training, you are offered $50 to stay the course. I wouldn't think you irrational if you

turned it down in order to return home to your mother. Similarly, I don't think it's irrational for you to turn down "the offer" of spinning a more consistent self-narrative by staying to fight. Neither consideration—the $50, the consistency of your self-narrative—is of the right kind to make the difference when so much else of moral importance it at stake.

Difference 2: Duration Between Decisions. Another potential factor is the difference in *duration between actions in the sequence*. In Generous Game Show, relatively little time passes between your decision in Round 1 and your decision in Round 2. In Sartrean Sequence, however, months pass between your decision to join the army and your later decision to ultimately return home to your mother.

The more time that elapses between the actions in the sequence, the more forgiving we're disposed to be of agents who fail to maintain plausible deniability about having suffered diachronic misfortune. Here are two reasons this might be the case. First, when the duration between actions in the sequence is very small, it's less plausible—given background assumptions about how humans generally work—that you lacked the ability to *self-bind*—that is, to perform the sequence "all at once" by forming an intention and following through on it. In other words, when the decisions occur one right after the other, we're inclined to interpret the story in such a way that bringing about the outcomes directly are assumed to be feasible options for you; the tree-diagrams, in this case, would misrepresent the decision.[27] If you have the ability to self-bind, though, bringing about a suboptimal outcome is straightforwardly synchronically irrational. And when the decisions happen quickly, it's harder to screen off the possibility that the agent has the ability to self-bind in a way that our intuitions can easily grasp. The second reason is this. The more time that passes between actions in the sequence, the easier it is to fill in the story so that plausible deniability has been maintained. This is because, as years fade, so does one's own and one's "audience's" memories. My life has changed a great deal since I was in kindergarten; many years have passed, and I don't feel beholden to the projects or plans set into motion back then. So much time has passed that I don't risk undermining my self-narrative by effectively ignoring, along with everyone else, the preferences I had in kindergarten.[28]

[27] A similar point (about how to represent self-binding in decision trees) can be found in Hammond 1976 and Meacham 2010a. For more on self-binding and it's implications for decision theory, see Arntzenius et al. 2004; Elster 2000; and Holton 2009.

[28] Hare (1989) makes a similar point when he says, "I wanted, when a small boy, to be an engine-driver when I grew up; when I have graduated as a classical scholar at the age of 18, and am going to take the Ph.D. in Greek literature, somebody unexpectedly offers me a job as an engine driver. In deciding whether to accept it, ought I to give any weight to my long-abandoned boyhood ambition?" (156). See also Bykvist 2003, 2006, on the import of past preferences on current decision making.

As time marches forward, the importance of paying service to one's past preferences in order to maintain a flattering self-narrative decays.

Difference 3: Transformative Power. Another difference between the two stories—in addition to the difference in stakes, and the difference in duration between decisions—is that the decisions made in Sartrean Sequence are, and the decisions made in Generous Game Show are not, in some sense *transformative*. (In fact, this brings us closer to the point Sartre himself seemed to use the example to make. Because of our existential predicament, you (or the student, rather) must choose. And in so doing, one affirms certain values that come to constitute one's moral or practical identity. (See Korsgaard 2009; Chang 2009, 2013, for further discussion.) A choice might be transformative because it will result in a transformative experience (Paul 2014, 2015).[29] Chang (2015) argues that, in some cases, the choice *itself* might be transformative. In either case, we sometimes face decisions that have the power to fundamentally alter the kind of person we will become.

The decision to fight Nazis or stay with Mom is, in some sense, a decision that's also about what kind of person to become. It's potentially transformative. And—I claim—if what you decide will partially constitute your rational identity, or change the kind of person you will become, or involves acting on your deeply-held values (as opposed to mere preferences), we're prone to be more forgiving. This is because transformative decisions usually involve a switch in the "audience" to whom you're interested in projecting a flattering autobiographical narrative. Furthermore, if the kind of transformation brought about by your decision is radical enough—in particular, if it involves a change in your underlying core *values*—you have reason to repudiate your past behavior, and thus lack a compelling reason to integrate into your autobiography going forward.[30] So the fact that one of your available actions can be better integrated into a unifying narrative should matter little to you. (In fact, depending on who you're choosing to become, the fact that one of your possible actions can be better integrated into a unifying narrative might amount to a point *against* performing that

[29] Paul (2014, 2015) argues that a transformative experience can be *epistemically* transformative, *personally* transformative, or both. An experience is *epistemically transformative* if the only way to know what it's like to have it is to have it. Seeing the color red for the first time would be epistemically transformative. An experience is *personally transformative* if having it would significantly and fundamentally change what you care about. Paul's central example—becoming a new parent—is an example of both. The distinction doesn't matter much for our purposes, but I will be focusing more heavily on the personally transformative aspects of these decisions (see also Ullmann-Margalit 2006, 2007).

[30] Another popular example of this phenomenon is Parfit's Russian Nobleman, who intends to give his land to the peasants but anticipates that his socialist ideals will fade with age (Parfit 1984, 327–329). It's natural to think of the Russian Nobleman as essentially two different people. When you face a transformative decision, you have to decide who you are to become, and you might decide to become a different person than who you are now.

action.) You still desire to spin a flattering self-narrative, of course, but it should be a story about your new self; the story of who you *were* has ended.

So here are three important differences between the two cases—a difference in stakes, a difference in duration between decisions, and a difference in the transformative power of the decisions—all of which contribute, in their own ways and in varying degrees, to our inclination to be more forgiving about some cases of diachronic misfortune than others.

6 Does the Explanation Show Too Little?

Here's another potential problem. Imagine an agent who *doesn't* have anything even resembling a desire to maintain plausible deniability about having suffered diachronic misfortune: an asocial alien, for example. Doesn't it also seem irrational for *this* agent to place the $50 voucher in box *B* and then take box *A*? But, given that this agent doesn't care about spinning a flattering social story, it needn't have done anything instrumentally irrational.

That's true, but I think we have reason to be suspicious of our intuitions in these cases, and that the view I've sketched can vindicate this suspicion. Let me briefly recapitulate the dialectic. There are cases like Generous Game Show in which certain sequences of actions strike us as irrational. On the one hand, my proposal aims to vindicate those intuitions: typically, you have done something irrational if you perform such a sequence. On the other hand, my proposal aims to provide an error-theory: despite appearances, you've done nothing *diachronically* irrational. My proposal ascribes to us a standing, non-instrumental desire—the desire to spin flattering yet plausible self-narratives—which, unless outweighed by other considerations, will typically render one of the actions in the sequence *synchronically* irrational. The reason that it might seem like there are fundamental diachronic norms of rationality is that typically when it seems like such a norm has been violated, some particular *synchronic* norm has been. In the case of the asocial alien things are not so simple. Proponents of Time-Slice Rationality should say that the asocial alien, despite our intuitions to the contrary, has not behaved irrationally *in any sense*. The asocial alien, by hypothesis, has no standing desire to spin a flattering yet plausible self-narrative, and so needn't have violated a synchronic norm by performing the offending sequence.

However, I think we have reason to be suspicious of the intuition that the asocial alien has behaved irrationally, diachronically or synchronically. Here's why. Much like the desire to tell flattering stories about ourselves, our intuitions about rationality *themselves* have been forged on the anvil of social interaction. And because this desire is central to our self-identities and central to our ability to get along with each other, it's not implausible to think that we *project* this desire onto others. We can, of course, imagine

asocial agents (and we can stipulate that the protagonists of our thought experiments, e.g., only care about money), but our intuitions about rationality might have trouble being sensitive to these features. In general, the farther away a hypothetical case is from the domain in which our intuitions have been trained, the less we should trust those intuitions.

As it stands, however, this response isn't entirely satisfying. Even if the desire to tell flattering stories about ourselves has been planted deeply inside of us via some evolutionary process, it doesn't follow that we'll have trouble imagining hypothetical agents lacking this desire in a way that will be difficult for our intuitions about rationality to reliably track. Offhand, we seem to encounter no such difficulties with other desires that have a similar psychological status and etiology. Our desire for sweets, for example, might be evolutionarily hard-wired, and yet adopting a strict no-sweets diet doesn't typically strike us as irrational. Nor do we typically judge the celibate monk as behaving irrationally in virtue of being celibate.

I think the desire to spin a flattering self-narrative is importantly different in two respects. First, the underlying reasons that explain why we've internalized such a desire are still operative in many cases. In other words, there are still in many cases *instrumental* reasons to want to present oneself in a flattering light. We continue to be social creatures who must coordinate with each other in order to accomplish our other goals, whatever those other goals happen to be. Whatever your goals are, it will often be in your interest to spin a flattering yet plausible self-narrative. But that's not true of the desire for sweets or sex. Those desires are, generally, no longer instrumental in satisfying your other goals whatever they might be. If, however, they *were* instrumental in satisfying your other goals, it's no longer obvious that it wouldn't be irrational to not desire them. For example, if you're trapped in a candy store (with no hope of escape), sticking to your no-sweets diet would be irrational. We are, as it were, trapped in a social world (with no hope of escape). And so we have instrumental reason to care about our self-narratives.[31]

The second difference is related to the first, but more significant. Our practice of rationally evaluating each other's behavior is *itself* inseparably bound up with our desire to get along with each other. Evaluating someone's behavior as rational or irrational is part of the larger social practice of collectively settling on norms that aid in solving coordination problems.[32]

[31] Thanks to an anonymous referee for suggesting this example.

[32] Gibbard (1990) develops a view very similar to the one I have in mind: the function of rational evaluations is to promote coordination in action and feelings. A similar idea concerning *epistemic* rationality can be found in Dogramaci 2012, 2015. Dogramaci (2012) argues that "the simple use [of epistemically evaluative assertions] promotes the *coordination* of epistemic rule-following across the linguistic community" (522) which can help "extend our common epistemic reach by enabling each person to serve as an 'epistemic surrogate' of any other person" (524). A similar idea concerning *knowledge* can be found in Craig 1990.

The idea is that rational evaluations—using words like 'rational' or 'irrational' to describe someone's behavior—are partly *expressive*. I judge you to be instrumentally irrational insofar as your behavior deviates from the predictions of my folk psychological theory *and* I express disapproval. This is consistent with understanding instrumental rationality to be about taking the best means to one's ends. If I judge you to be instrumentally irrational, I don't think you've taken the best means to your ends. I can only make such a judgment if I have some idea of what your ends are. In real-life cases, my beliefs about your ends come from constructing a folk psychological theory based on (somewhat meager) evidence about your past behavior. My folk psychological theory issues predictions about how you'll behave in the future. If your future behavior defies these predictions, I must either revise my theory or judge that you've behaved irrationally on this occasion.[33] One function of doing the latter, I conjecture, is to exert pressure on you to bring your future behavior into line with our expectations.

Our intuitions about the rationality of someone's diachronic behavior, then, might not merely track whether her behavior is sensible from her perspective, but also the extent to which it would be easy to predict her behavior. When we imagine the asocial alien, perhaps we can't help but imaginatively treat her as a member of our rational community. The asocial alien might not care about her self-narrative. But were she a member of our community, we'd exert pressure on her to construct one. We'd interact with her in ways to make her care. Given that one way to exert such pressure is with the use of words like 'rational' and 'irrational,' it's not unreasonable to expect our intuitions about the rationality of the asocial alien's behavior to track whether we'd be disposed to call it 'rational' or 'irrational' were she a member of our community. If this is so, we have some reason to be cautious about the reliability of our intuitions in these sorts of cases.

7 Conclusion

The argument in this paper goes like this. Because we are social creatures, there is rational pressure on us to signal that we have those qualities that would make one a good teammate. It serves our basic ends to work together, to build communities, and to be part of a team. In order to work together on teams, we need to signal that we would make good teammates. Through a process of social evolution, we have come to internalize the desire to act in ways that can be (easily and plausibly) integrated into an autobiographical narrative that represents us in a flattering light. We want,

[33] This is very different from other cases of theorizing. If my theory of planetary motion, for example, makes a false prediction, I have reason to revise my theory. That might make me sad. But I wouldn't express disapproval toward the behavior of the planets. The difference is that, in this case, my expression of disapproval would (of course) have no influence on the future behavior of the planets. But folk psychology is different. In folk psychology, theorists and subjects interact in ways that mutually influence each other (McGeer 2001, 2007, 2015).

that is, to construct a narrative according to which we look like the kind of person with whom others would want on their team.

The desire to present a flattering self-narrative involves wanting to act in ways that will allow us to maintain plausible deniability about either having made an unwise decision at some point, or having unstable preferences. Why is that? People who make unwise decisions don't make ideal teammates. We want to collaborate with others who assess their evidence sensibly, proportion their beliefs to their evidence, and act sensibly in light of these beliefs. Moreover, people with unstable preferences don't make ideal teammates either. We want to collaborate with people whose behavior is (more or less) easy to predict, people who make it relatively easy to read-off their desires from their choice-behavior, etc. We're particularly weary of revealing that we've made an unwise decision that turned on the stability of our own preferences.

When you suffer diachronic misfortune, either you've made a prediction that turned out to be false or your preferences have changed. So, if you don't want to make salient the possibility that you've made an irrational prediction or that your preferences are unstable, you have reason to maintain plausible deniability about having suffered diachronic misfortune—at least in cases in which suffering misfortune is likely to suggest these things.

I've argued that our deeply rooted standing desire to spin flattering yet plausible self-narratives about our diachronic behavior helps explain why certain sequences of actions in certain circumstances strikes us as irrational. The cases in which we are inclined to think that a diachronic requirement of rationality has been violated are also those cases in which, given that the agent in question has such a desire, she will have fallen afoul of the purely *synchronic* rational requirement to take the means that best promote her ends. Our nature as social agents has endowed us with a need for diachronic stability. In this way, our concern for our autobiographies simulates diachronic norms.

Ryan Doody
The Hebrew University of Jerusalem
E-mail: ryan.d.doody@gmail.com

References:

Arntzenius, Frank, Adam Elga, and John Hawthorne. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind* 113 (450): 251–283. https://doi.org/10.1093/mind/113.450.251.

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80 (4): 1095–1111. https://doi.org/10.2307/1960858.

Baumeister, Roy F. 1982. "A Self-Presentational View of Social Phenomena." *Psychological Bulletin* 91 (1): 3–26. https://doi.org/10.1037/0033-2909.91.1.3.

Baumeister, Roy F. 1999. "The Nature and Structure of the Self: An overview." In *The Self in Social Psychology*, edited by Roy F. Baumeister. Philadelphia, PA: Psychology Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: MIT Press.

Bratman, Michael. 2010. "Agency, Time, and Sociality." *Proceedings and Addresses of the American Philosophical Association* 84 (2): 7–26.
Bratman, Michael. 2012. "Time, Rationality, and Self-Governance." *Philosophical Issues* 22 (1): 73–88. https://doi.org/10.1111/j.1533-6077.2012.00219.x.
Briggs, R. A. 2009. "Distorted Reflection." *Philosophical Review* 118 (1): 59–85. https://doi.org/10.1215/00318108-2008-029.
Broome, John. 1993. "Can a Humean Be Moderate?" In *Value, Welfare, and Morality*, edited by R. G. Frey and Christopher W. Morris, 51–73. Cambridge: Cambridge University Press.
Bykvist, Krister. 2003. "The Moral Relevance of Past Preferences." In *Time and Ethics: Essays at the Intersection*, edited by Heather L. Dyke, Vol. 14 of *Library of Ethics and Applied Philosophy*, 115–136. Boston: Springer, Dordrecht.
Bykvist, Krister. 2006. "Prudence for Changing Selves." *Utilitas* 18 (3): 264–283. https://doi.org/10.1017/S0953820806002032.
Carr, Jennifer. 2013. *Justifying Bayesianism*. PhD diss. Massachusetts Institute of Technology.
Carr, Jennifer. 2015. "Don't Stop Believing." *Canadian Journal of Philosophy* 45 (5): 744–766. https://doi.org/10.1080/00455091.2015.1123454.
Chang, Ruth. 2009. "Voluntarist Reasons and the Sources of Normativity." In *Reasons for Action*, edited by David Sobel and Steven Wall, 243–271. New York: Cambridge University Press.
Chang, Ruth. 2013. "Grounding Practical Normativity: Going Hybrid." *Philosophical Studies* 164 (1): 163–187. https://doi.org/10.1007/s11098-013-0092-z.
Chang, Ruth. 2015. "Transformative Choices." *Res Philosophica* 92 (2): 237–282. https://doi.org/10.11612/resphil.2015.92.2.14.
Chang, Ruth. 2017. "Hard Choices." *Journal of the American Philosophical Association* 3 (1): 1–21. https://doi.org/10.1017/apa.2017.7.
Christensen, David. 1991. "Clever Bookies and Coherent Belief." *The Philosophical Review* 100 (2): 229–247. https://doi.org/10.2307/2185301.
Cialdini, Robert B. 2001. *Influence: Science and Practice*. 4th Edition. Boston, MA: Allyn and Bacon.
Cosmides, Leda. 1989. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason?" *Cognition* 31 (3): 187–276. https://doi.org/10.1016/0010-0277(89)90023-1.
Craig, Edward. 1990. *Knowledge and the State of Nature*. Oxford: Oxford University Press.
Davidson, Donald. 1973. "Radical Interpretation." *Dialectica* 27 (3/4): 313–328. https://doi.org/10.1111/j.1746-8361.1973.tb00623.x.
Davidson, Donald, J. C. C. Mckinsey, and Patrick Suppes. 1955. "Outlines of a Formal Theory of Value." *Philosophy of Science* 22 (2): 140–160. https://doi.org/10.1086/287412.
Dennett, Daniel C. 1992. "The Self as Center of Narrative Gravity." In *Self and Consciousness: Multiple Perspectives*, edited by Frank S. Kessel, Pamela M. Cole, and Dale L. Johnson. Hillsdale, NJ: Erlbaum.
Dogramaci, Sinan. 2012. "Reverse Engineering Epistemic Evaluations." *Philosophy and Phenomenological Research* 84 (3): 513–530. https://doi.org/10.1111/j.1933-1592.2011.00566.x.
Dogramaci, Sinan. 2015. "Communist Conventions for Deductive Reasoning." *Noûs* 49 (4): 776–799. https://doi.org/10.1111/nous.12025.
Doody, Ryan. Unpublished. "The Sunk Cost 'Fallacy' Is Not a Fallacy." http://www.mit.edu/~rdoody/SunkCostFallacyIsNotaFallacy.pdf.
Dreier, James. 1996. "Rational Preference: Decision Theory as a Theory of Practical Rationality." *Theory and Decision* 40 (3): 429–276. https://doi.org/10.1007/BF00134210.
Elster, Jon. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge: Cambridge University Press.
Ferrero, Luca. 2009. "What Good Is a Diachronic Will?" *Philosophical Studies* 144 (3): 403–430. https://doi.org/10.1007/s11098-008-9217-1.

Ferrero, Luca. 2012. "Diachronic Constraints on Practical Rationality." *Philosophical Issues* 22 (1): 144–64. https://doi.org/10.1111/j.1533-6077.2012.00222.x.
van Fraassen, Bas C. 1995. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies* 77 (1): 7–37. https://doi.org/10.1007/BF00996309.
Frank, Robert H. 1987. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *The American Economic Review* 77 (4): 593–604.
Gauthier, David. 1997. "Resolute Choice and Rational Deliberation: A Critique and a Defense." *Noûs* 31 (1): 1–25. https://doi.org/10.1111/0029-4624.00033.
Gazzaniga, Michael S. 1998. *The Mind's Past*. Berkeley and Los Angeles, CA: University of California Press.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.
Gintis, Herbert. 2000. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton, NJ: Princeton University Press.
Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor Books, Doubleday.
Goffman, Erving. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. New York, NY: Anchor Books, Doubleday.
Goldie, Peter. 2012. "The Narrative Sense of Self." *Journal of Evaluation in Clinical Practice* 18 (5): 1064–1069. https://doi.org/10.1111/j.1365-2753.2012.01918.x.
Hammond, Peter. 1976. "Changing Tastes and Coherent Dynamic Choice." *The Review of Economic Studies* 43 (1): 159–173. https://doi.org/10.2307/2296609.
Hammond, Peter. 1988. "Consequentialist Foundations for Expected Utility." *Theory and Decision* 25 (1): 25–78. https://doi.org/10.1007/BF00129168.
Hanson, Robin. 2008. "Showing that You Care: The Evolution of Health Altruism." *Medical Hypotheses* 70 (4): 724–742. https://doi.org/10.1016/j.mehy.2007.08.020.
Hare, Richard M. 1989. "Prudence and Past Preference: Reply to Wlodzimierz Rabinowicz." *Theoria* 55 (3): 152–158. https://doi.org/10.1111/j.1755-2567.1989.tb00728.x.
Hedden, Brian. 2015a. "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research* 90 (2): 423–451. https://doi.org/10.1111/phpr.12048.
Hedden, Brian. 2015b. "Time-Slice Rationality." *Mind* 124 (494): 449–491. https://doi.org/10.1093/mind/fzu181.
Hogan, Robert and Stephen R. Briggs. 1986. "A Socioanalytic Interpretaion of the Public and the Private Selves." In *Public Self and Private Life*, edited by Robert F. Baumeister. New York: Springer-Verlag.
Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Hutto, Daniel D. 2007. *Folk Psychological Narratives: The Socio-Cultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.
Korsgaard, Christine. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.
Kurzban, Robert. 2010. *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind*. Princeton, NJ: Princeton University Press.
Kurzban, Robert and C. Athena Aktipis. 2007. "Modularity and the Social Mind: Are Psychologists Too Self-ish?" *Personality and Social Psychology Review* 11 (2): 131–149. https://doi.org/10.1177/1088868306294906.
Kurzban, Robert and Mark R. Leary. 2001. "Evolutionary Orgins of Stigmatization: The Functions of Social Exclusion." *Psychological Bulletin* 127 (2): 187–208. https://doi.org/10.1037/0033-2909.127.2.187.
Levi, Isaac. 1991. "Consequentialism and Sequential Choice." In *Essays in the Foundations of Decision Theory*, edited by Michael Bacharach and Susan Hurley, 70–101. Oxford: Basil Blackwell.
Levi, Isaac. 2002. "Money Pumps and Diachronic Books." *Philosophy of Science* 69 (S3): 235–247. https://doi.org/10.1086/341849.

Levine, Sheen S. and Robert Kurzban. 2006. "Explaining Clustering in Social Networks: Towards an Evolutionary Theory of Cascading Benefits." *Managerial and Decision Economics* 27 (2/3): 173–187. https://doi.org/10.1002/mde.1291.

Lewis, David. 1974. "Radical Interpretation." *Synthese* 27 (3): 331–344. https://doi.org/10.1007/BF00484599.

Machina, Mark J. 1989. "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty." *Journal of Economic Literature* 27 (4): 1622–1688.

Maher, Patrick. 1992. "Diachronic Rationality." *Philosophy of Science* 59 (1): 120–141. https://doi.org/10.1086/289657.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McGeer, Victoria. 2001. "Psycho-Practice, Psycho-Theory, and the Contrastive Case of Autism." *Journal of Consciousness Studies* 8 (5-7): 109–132.

McGeer, Victoria. 2007. "The Regulative Dimension of Folk Psychology." In *Folk Psychology Re-Assessed*, edited by Daniel Hutto and Matthew Ratcliffe, 137–156. Dordrecht, Netherlands: Springer.

McGeer, Victoria. 2015. "Mind-Making Practices: The Social Infrastructure of Self-Knowing Agency and Responsibility." *Philosophical Explorations* 18 (2): 259–281. https://doi.org/10.1080/13869795.2015.1032331.

Meacham, Christopher. 2010a. "Binding and Its Consequences." *Philosophical Studies* 149 (1): 49–71. https://doi.org/10.1007/s11098-010-9539-7.

Meacham, Christopher. 2010b. "Unravelling the Tangled Web: Continuity, Internalism, Non-Uniqueness and Self-Locating Beliefs." In *Oxford Studies in Epistemology*, edited by Tamar Szabó Gendler and John Hawthorne, Vol. 3, 86–125. Oxford: Oxford University Press.

Moss, Sarah. 2015. "Time-Slice Epistemology and Action Under Indeterminacy." In *Oxford Studies in Epistemology*, edited by Tamar Szabó Gendler and John Hawthorne, 172–194. Oxford: Oxford University Press.

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

Paul, L. A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A. 2015. "What You Can't Expect When You're Expecting." *Res Philosophica* 92 (2): 149–170. https://doi.org/10.11612/resphil.2015.92.2.1.

Pettit, Philip. 1991. "Decision Theory and Folk Psychology." In *Essays in the Foundations of Decision Theory*, edited by Michael Bacharach and Susan Hurley, 147–175. Oxford: Blackwell.

Rabinowicz, Wlodek. 1995. "To Have One's Cake and Eat It, Too: Sequential Choice and Expected-Utility Violations." *The Journal of Philosophy* 92 (11): 586–620. https://doi.org/10.2307/2941089.

Ramsey, Frank. 1926. "Truth and Probability." In *The Foundations of Mathematics and Other Logical Essays*, edited by Richard Bevan Braithwaite, 156–198. London: Kegan, Paul, Trench, Trubner and Co.

Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge, MA: MIT Press.

Sartre, Jean-Paul. 1946. "The Humanism of Existentialism." In *Jean-Paul Sartre: Essays in Existentialism*, edited by Wade Baskin, 31–62. New York: Carol Publishing Group.

Schick, Frederic. 1986. "Dutch Bookies and Money Pumps." *The Journal of Philosophy* 83 (2): 112–119. https://doi.org/10.2307/2026054.

Schlenker, Barry R. 1975. "Self-Presentation: Managing the Impression of Consistency When Reality Interferes with Self-Enhancement." *Journal of Personality and Social Psychology* 32 (6): 1030–1037. https://doi.org/10.1037/0022-3514.32.6.1030.

Schlenker, Barry R. 1980. *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations*. Monterey, CA: Brooks/Cole Publishing.

Seidenfeld, Teddy. 1988. "Decision Theory Without 'Independence' or Without 'Ordering': What is the Difference?" *Economics and Philosophy* 4 (2): 267–90. https://doi.org/10.1017/S0266267100001085.

Seidenfeld, Teddy. 1994. "When Normal and Extensive Form Decisions Differ." In *Logic, Methodology and Philosophy of Science IX*, edited by Dag Prawitz, Brian Skyrms, and Dag Westerståhl, Vol. 134, 451–466. Amsterdam, Netherlands: Elsevier Science B.V.

Sexton, Steven E. and Alison L. Sexton. 2014. "Conspicuous Conservation: The Prius Halo and Willingness to Pay for Environmental Bona Fides." *Journal of Environmental Economics and Management* 67 (3): 303–317. https://doi.org/10.1016/j.jeem.2013.11.004.

Shrauger, Sidney J. and Thomas J. Schoeneman. 1979. "Symbolic Interactionist View of Self-Concept: Through the Looking Glass Darkly." *Psychological Bulletin* 86 (3): 549–573. https://doi.org/10.1037/0033-2909.86.3.549.

Simler, Kevin and Robin Hanson. 2018. *The Elephant in the Brain: Hidden Motives in Everyday Life*. New York: Oxford University Press.

Skyrms, Brian. 1987. "Dynamic Coherence and Probability Kinematics." *Philosophy of Science* 54 (1): 1–20. https://doi.org/10.1086/289350.

Skyrms, Brian. 1993. "A Mistake in Dynamic Coherence Arguments?" *Philosophy of Science* 60 (2): 320–328. https://doi.org/10.1086/289735.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Smith, Adam. 1759. *A Theory of Moral Sentiments*. Edited by David D. Raphael and Alexander L. Macfie. 1976. Oxford: Clarendon Press.

Stalnaker, Robert. 1999. "Extensive and Strategic Forms: Games and Models for Games." *Research in Economics* 53 (3): 293–319. https://doi.org/10.1006/reec.1999.0200.

Steele, Katie. 2010. "What Are the Minimal Requirements of Rational Choice? Arguments from the Sequential-Decision Setting." *Theory and Decision* 68 (4): 463–487. https://doi.org/10.1007/s11238-009-9145-3.

Stone, Jeff, Andrew W. Weigand, Joel Cooper, and Elliot Aaronson. 1997. "When Exemplification Fails: Hypocrisy and the Motive for Self-Integrity." *Journal of Personality and Social Psychology* 72 (1): 54–65. https://doi.org/10.1037/0022-3514.72.1.54.

Swann, William B. 1985. "The Self as Architect of Social Reality." In *The Self and Social Life*, edited by Barry R Schlenker. New York: McGraw-Hill.

Tedeschi, James T., Barry R. Schlenker, and Thomas V. Bonoma. 1971. "Cognitive Dissonance: Private Ratiocination or Public Spectacle?" *American Psychologist* 26 (8): 685–695. https://doi.org/10.1037/h0032110.

Teller, Paul. 1976. "Conditionalization, Observation, and Change of Preference." In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, edited by William Leonard Harper and Clifford Alan Hooker, Vol. 6a of *The University of Western Ontario Series in Philosophy of Science*, 205–259. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-010-1853-1_9.

Tice, Dianne M. and Roy F. Baumeister. 2001. "The Primacy of the Interpersonal Self." In *Individual Self, Relational Self, Collective Self*, edited by Marilynn B. Brewer and Constantine Sedikides. New York: Psychology Press.

Tooby, John and Leda Cosmides. 1996. "Friendship and the Banker's Paradox: Other Pathways to the Evolution of Adaptation for Altruism." In *Proceedings of the Royal British Academy*, edited by Walter Garrison Runciman, John Maynard Smith, and Robin Ian MacDonald Dunbar, Vol. 88, 119–143. New York: Oxford University Press.

Trivers, Robert L. 2000. "The Elements of a Scientific Theory of Self-Deception." *Annals of the New York Academy of Science* 907 (1): 141–131. https://doi.org/10.1111/j.1749-6632.2000.tb06619.x.

Ullmann-Margalit, Edna. 2006. "Big Decisions: Opting, Converting, Drifting." *Royal Institute of Philosophy Supplement* 58: 157–172. https://doi.org/10.1017/S1358246106058085.

Ullmann-Margalit, Edna. 2007. "Difficult Choices: To Agonize or Not to Agonize?" *Social Research: An International Quarterly* 74 (1): 51–78.
Velleman, J. David. 2000. "Deciding How to Decide." In *The Possibility of Practical Reason*. Oxford: Clarendon Press.
Velleman, J. David. 2005. "The Self as Narrator." In *Autonomy and the Challenges to Liberalism: New Essays*, edited by John Christman and Joel Anderson, 56–76. Cambridge: Cambridge University Press.
Velleman, J. David. 2009. *How We Get Along*. Cambridge: Cambridge University Press.
Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.
Zahavi, Amotz. 1975. "Mate Selection—A Selection for a Handicap." *Journal of Theoretical Biology* 53 (1): 205–214. https://doi.org/10.1016/0022-5193(75)90111-3.