



NYU | LAW

New York University School of Law



**October 1st, 2026, from 4-7 pm
Lester Pollock Room, FH, 9th Floor**

Colloquium in Legal, Political, and Social Philosophy

**Conducted by
Sophia Moreau and Liam Murphy**

**Speaker: Niko Kolodny, University of California, Berkeley
Paper: Preference Change for Objectivists**



Colloquium Website: <http://www.law.nyu.edu/node/22315>

Preference Change for Objectivists¹
Niko Kolodny (kolodny@berkeley.edu)

Abstract

It matters whether we get what we want. But what if what we want changes? What if what I want today to happen tomorrow is not what I wanted yesterday to happen tomorrow? Which want, if either, do we satisfy? What if manipulation or oppression causes my wants to change? Does satisfying them still matter, in the same way? This paper explores these questions, from an objectivist standpoint, which holds that getting what we want matters, when it does, only because, first, we have reason to want what we do and, second, our wanting it partly constitutes something of further value.

Introduction

Some think—and I’m among them—that it matters whether we get what we want. Indeed, some think—and I’m not among them—that that’s all that matters. Desire-satisfaction theorists of well-being say, roughly, that one enjoys greater well-being just insofar as one’s desires are better satisfied.

Desire-satisfaction theorists of rationality say, roughly, that one is rationally required to choose the option that would best satisfy one’s desires. And desire-satisfaction utilitarians say, roughly, that one is morally required to choose the option that would maximize the satisfaction of everyone’s desires. These theories face many questions. I’ll focus on two. First, does it matter *when* we desire what we do? What if my preferences change, so that what I want today to happen tomorrow is not what I wanted yesterday to happen tomorrow? Which desire, if either, do we satisfy? Second, does it matter *from where* our desires come? What if manipulation or oppression causes my preferences to change? Does satisfying them still matter, in the same way? (Strictly speaking, these questions, or versions of them, arise even if my preferences never change. But these questions stand out most starkly when my preferences do change. Hence the title.)

Here’s how I came to these questions. I was discussing with some AI-safety people what goal we should give AI, if we had it in our power to give AI a goal. They took it as axiomatic that the goal would be to satisfy preferences, whatever they might be, no preference being objectively better than any other.

¹ For feedback on the first half of this paper, I thank an audience at Florida State University. For feedback on the second half of this paper, I thank audiences at the Freie Universität Berlin, the Humboldt Universität zu Berlin, and the Université de Fribourg, a conversation with Johann Frick, and correspondence with Daniele Bruno, Stefan Gosepath, and Thomas Schmidt. For feedback on the introduction, I thank my graduate seminar at Berkeley.

They were at the same time deeply concerned that AI might modify our preferences, precisely to make them easier to satisfy.

This puzzled me. If the goal was to satisfy preferences, why not welcome AI's discovery of a more efficient way to realize that goal? What was the problem? Of course, one might be concerned that AI might modify our preferences to make them for bad or worthless things—say, for doing nothing but drinking Soylent alone in darkened cells—and then satisfy those preferences. But that couldn't be the AI-safety people's concern, if the goal really was satisfying preferences, whatever they might be—if there was no arguing about the independent worth of the objects of preference. And, of course, one might be concerned that AI might modify our preferences in such a way that their satisfaction (or beliefs about their satisfaction) would produce less pleasure—say, again, by giving us preferences for doing nothing but drinking Soylent alone in darkened cells—and then satisfy those preferences. But that couldn't be the concern, if the goal really was the satisfaction—mere fulfillment—of preferences, rather than the production of feelings of satisfaction that might attend the fulfillment of preferences (or beliefs about their fulfillment).

So, what else could the concern be? One answer, which some of my interlocutors voiced, was: “Don't you prefer that AI not modify your preferences?! I certainly do!” To this, I replied: “But if that's the problem, isn't there an easy solution? If what we prefer now is what matters, and if we prefer now that AI not modify our preferences in the future, then just give AI the goal of satisfying the preferences we have now, and we needn't worry that AI will modify our preferences in the future.” The reply was something like: “But surely you don't want AI to satisfy the preferences that we have now for what happens in the future, but instead whatever preferences we might have in the future for what happens in the future. After all, what if in the future we look back at the preferences we have today in the way that we look back at the preferences of the ancient Romans, for slavery and gladiatorial contests? We don't want the preferences that we have today to be locked in! It's a given, then, that we will program AI to satisfy the preferences that we have in the future for what happens in the future. And that brings us back

to the problem: what if AI modifies our preferences so that our future preferences are OK with modified preferences and then satisfies those modified preferences?”

This reply left me again confused. On the one hand, the thought that it was a problem if AI modified our preferences seemed to depend on the idea that AI should satisfy our *present* preferences for the future, preferences that are now opposed to modification. On the other hand, the thought that AI *would* modify our preferences seemed to depend on the idea that we would program AI to satisfy our future desires for the future, because AI *should* satisfy our future preferences for the future. This raised, in a stark way, the question of how we should deal with the conflict between our present preference for the future and our future preference for the future. More generally, what are we to say when what one desires at time t_1 to happen at t_3 is not what one desires at t_2 to happen at t_3 ? Should the desire at t_1 be privileged over the desire at t_2 ? Should the desire at t_1 and the desire at t_2 each be given a vote? Something else? Call this the **Problem of Time**.

So that's one answer to the question of what the worry could be about AI modifying our preferences, if not the worry that AI might modify our preferences for the worse. Another answer is simpler to state. This is that my interlocutors held that the correct goal was not actually the satisfaction of preferences period, but instead the satisfaction of preferences *formed in the right way*. If we arrived at preferences for drinking Soylent alone in darkened cells, then so be it, but only if we arrived at those preferences in the right way. Having our preferences modified by AI would not be the right way. So the satisfaction of our modified preferences would not further the correct goal. Yet what way was the “right” way, I wondered, and why should it matter? Call this the **Problem of Origin**.

At first, I thought that these problems—the Problem of Time and the Problem of Origin—weren't my problems, since I don't think that “the goal” is satisfying preferences whatever they might be. I don't subscribe to the desire-satisfaction theory of well-being, or the desire-satisfaction theory of rationality, or desire-satisfaction utilitarianism. But I began to worry. I am not so much of an “objectivist” that I deny that what we want can be normatively and prudentially significant, making a difference to what we have reason to do and to what contributes to our well-being. Suppose that I am deciding between life as a poet

and life as a philosopher. Surely it is a reason to be philosopher rather than a poet that I want to be a philosopher rather than a poet. And surely the fact that I want to be a philosopher rather than a poet makes it the case that, other things equal, things would be better for me if I were a philosopher rather than a poet. So, the questions arise for me: what if yesterday I wanted to be a philosopher today, whereas tomorrow I will want instead to have been a poet today? Which desire, if either, matters to what I have reason to do or to my well-being? Similarly, we might imagine that my desire to be a philosopher has all sorts of origins, some of them being quite suspect, such as the expectations of overbearing philosopher parents. If my desire did have such a suspect origin, would that affect whether it matters to what I have reason to do or to my well-being?

The Problem of Time

I am going to focus on the Problem of Time for the first half of this paper, turning to the Problem of Origin in the second half. I will begin by considering the Problem of Time as it arises for desire-satisfaction theories of well-being and rationality in all of their generality. (I will assume that desire-satisfaction utilitarianism does not require separate treatment, since it seems derivative from the desire-satisfaction theory of well-being.) Part of my point will be that the Problem of Time is less of a problem for a more restricted, objectivist view, of the sort that I take, which holds that desires have only a limited normative and prudential significance. But that will be easier to see if we begin by considering the Problem of Time as it arises for the more general views.

Before going further, we should say just a bit more about desire. Contrast two frameworks. In one framework, desires are mere theoretical constructs derived from actual or hypothetical choice behavior. On one version of this approach, for illustration, if one's choice behavior (or one's "preference ordering" over actions) meets certain conditions, then we can construct a credence function—which assigns probabilities to states of the world—and a utility function—which assigns degrees of desirability to states of the world—such that one action, *a*, is at least as preferred as another, *b*, iff the expected utility of *a* is at least as high as the expected utility of *b*, where the expected utility of *a* is the sum over all states, *s*, of the probability of *s* conditional on *a* times the utility of *s* given *a*. To say that one desires one state of

the world more than another, on this view, is just to say that one's choice behavior can be represented by a utility function that assigns a higher value to the former than to the latter.

In another framework, desires are not mere theoretical constructs. Instead, they are psychological states with a certain strength or intensity that can be determined, if not by introspection, then at least independently of choice behavior. This is the framework that I will assume we are working within. It is hard to see how otherwise desire could play the roles in which we are interested. In choosing to be a philosopher rather than a poet, I see my desiring to be a philosopher more strongly than my desiring to be a poet as a weightier reason to be a philosopher than a poet. That would make little sense if the desire was simply a construct of choice behavior. (Compare Pettigrew 2020 14–15.) I assume further that strengths of desire can be compared across different time-stages of the same person, as well as across different people, and meaningfully summed, so that we can say that one person desires X twice as much as someone else desires Y, so that the first would derive twice as much desire-satisfaction from X as the second would derive from Y. Again, it is hard to see how otherwise desire could play the roles in which we are interested.

The Tyranny of Other Times

As a point of entry into the Problem of Time, consider two cases introduced by Derek Parfit and Richard Hare. (I follow Bykvist 2003 and Pettigrew 2020 in leading with these examples.) Here's Parfit:

When I was young what I most wanted was to be a poet. This desire was not conditional on its own persistence. I did not want to be a poet only if this was what I later wanted. Now that I am older, I have lost this desire.... Does my past desire give me a reason to try to write poems now, though I have no desire to do so? (1984 157)

Here's Hare:

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 on 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? (1989 156)

I take it as intuitively clear that my past desire does not give me any reason to write poems now or to accept the job as an engine driver. Nor would doing so promote my well-being. Let us call this problem the **Tyranny of Other Times**.

As the label is chosen to suggest, the problem concerns not only past desires for the future, but also future desires for the past. Suppose that one knows that, although one does not want to be a poet now, one knows that later in life one will want to have been a poet now, even against one's will. Suppose that one will not experience any disappointment in the future from the belief that in the past one was not a poet. Perhaps, no matter what, one will be content in the belief, true or false, that, in the past, one was a poet. Intuitively, one's future desire gives one no reason to be a poet against one's will now. Nor would it contribute to one's well-being to be a poet against one's will now.

Time-of-Choice View

So how do we avoid Tyranny? The Time-of-Choice View answers that the desires that count are just those held at the time of choice. Plugged into the desire-satisfaction theory of rationality, the Time-of-Choice View says that one is rationally required to choose the option that would best satisfy one's desires at the time of choice. This delivers the desired result that it is not rational to write poems now. Choosing now to write poems would not best satisfy one's desires at the time of choice, since the time of choice is now, and one doesn't want to write poems now.

However, first, the Time-of-Choice View cannot be plugged into the desire-satisfaction theory of well-being, because, when it comes to well-being, there is no privileged time of choice.

Second, the Time-of-Choice View implies that I would have been rational, when I was younger, to choose to lock myself, if I somehow could have done so, into a life of poetry writing that I did not want to lead later. Similarly, if I don't in fact care now whether my future desires are satisfied, then the Time-of-Choice View implies that it is rational for me to make no provision now for the satisfaction of my future desires. Call this the **Problem of Imprudence**.

One might wonder how judging Tyranny to be Tyranny is compatible with judging Prudence to be Prudence. If it is Tyranny—a bad thing—to let preferences at other times influence the present, understood as the time of choice, how can it be Prudence—a good thing—to let preferences at other times influence the present? The difference is not that in the case of Tyranny, *past* preferences influence the present whereas in the case of Prudence *future* preferences influence the present. Again, Tyranny can be

matter of *future* preferences influencing the present. We need to make a different distinction. It is Tyranny to allow one's future desires for the *present*—another time—to determine what one does in the present. By contrast, it is not Tyranny, but instead Prudence, to allow one's future desires for the *future*—the same time—to determine, at least to some extent, what one does in the present. This is how the two judgments are compatible.

Time-of-Satisfaction View

The Time-of-Satisfaction View offers another way out from Tyranny, this time without risking Imprudence. It says that the desires that count are just one's "simultaneous" desires: desires held at one time for what happens at that same time, or put another way, desires that are satisfied (if they are satisfied) at the time that they are held. By contrast, non-simultaneous or "intertemporal" desires—desires held at one time for what happens at another time—do not count. Plugged into the desire-satisfaction theory of rationality, the Time-of-Satisfaction View says that one is rationally required to choose the option that would best satisfy one's simultaneous desires. Plugged into the desire-satisfaction theory of well-being, it says that one's well-being is promoted by and only by the satisfaction of one's simultaneous desires. And it can be plugged into the desire-satisfaction theory of well-being, since it doesn't require privileging the time of choice.

The Time-of-Satisfaction View avoids Tyranny. It implies that it is not rational to write poems now, when one does not desire to write poems, and that doing so does not contribute to one's well-being. The adolescent desire to write poetry later in life is not a simultaneous desire. It is instead an intertemporal desire: specifically, a past desire for what happens in the present. So it does not count. At the same time, the Time-of-Satisfaction View avoids Imprudence. According to it, it would not be rational to neglect now what one's future self will want for one's future. Those are simultaneous desires—specifically, future desires for what happens in the future—and so they count.

So why not accept the Time-of-Satisfaction View? First, it seems that it is rational now to act to satisfy one's postmortem desires, for what happens after one's death. It also seems that such satisfaction contributes to one's well-being. After all, we often feel that we have very strong reasons to satisfy

postmortem desires, for the sake of the person whose desires they are. But after one dies, one does not desire anything. There are no simultaneous postmortem desires; they are all intertemporal desires for the future. Call this the **Problem of Postmortem Desires**.

A second problem with the Time-of-Satisfaction View is nicely illustrated by Richard Pettigrew's example of Aneri, who is deciding on a career.

If [Aneri] chooses to become a conservationist, her values will stay close to those she currently has—indeed, they will remain the same, but her job will not fully satisfy those values. If, on the other hand, she becomes a police officer, her values will change considerably, but her job will satisfy those values better (2020 23–4).

In other words, becoming a police officer promises greater simultaneous desire satisfaction over Aneri's life. Therefore, the Time-of-Satisfaction View implies that it is rationally required. It mandates what we might call **Disloyalty**. While it might be rationally permissible to become a police officer, it doesn't seem rationally required. The view that it is rationally required, as Pettigrew puts it, "seems to pay too little attention to our current values. What would it mean to value something if you were happy to do anything that will radically change your values in order to get more of what you would then come to value?"

Pettigrew illustrates the same point with someone who changes their desires to ones that will be better satisfied by their environment rather than changing the environment to satisfy the desires they have now.

Choosing that

might seem inauthentic, weak, and revealing of a lack of true commitment. We might naturally ask of such an individual: to what extent were those really your values if you were prepared to jettison them so easily to obtain values that would be more easily satisfied? (2020 210–11)

A third problem with the Time-of-Satisfaction View is illustrated by another example from Pettigrew, of someone who never expects that their desires will be satisfied when they have them.

Chinara... is... deciding whether to embark on writing the novel she's been planning for years. But she's spoken to a few writers and she thinks the following is very likely: throughout the process of writing it, she'll have a very strong preference in favour of having written a book—she'll think of that as a real achievement that is worth celebrating. She won't much prefer the process of actually writing it—she has little time for half-written novels and only endorses the situation of having actually completed one. However, she also knows that, after a writer has been through the process of completing a novel, they no longer have that strong preference for having completed a novel (2020 22–3).

As Pettigrew observes:

Suppose Chinara does decide to write her novel. Then, at each moment from the first time she puts pen to paper until the day she sees copies of the book on the shelves of her local bookstore, she will not, at that moment, get anything that, at that moment, she prefers or endorses” (2020 23).

Nevertheless, it seems that it is rationally permissible for Chinara to write her novel and that writing it contributes to her well-being. Call this the **Problem of the Never Satisfied**.²

Time-Neutral Aggregation

The next two approaches in effect treat each moment of the person’s life as though it were a different person. According to Time-Neutral Aggregation, we take each time-slice in turn. The satisfaction of all of that time-slice’s desires, even its intertemporal desires, count. We then aggregate the desire-satisfaction of each time-slice, according to some distributive principle. What it is rational for the person to choose, at any given time, is the best distribution of desire-satisfaction of all of the time-slices, with each time-slice counted equally. The person’s lifetime well-being is higher to the extent that the distribution of desire-satisfaction is better.

Since Time-Neutral Aggregation, unlike the Time-of-Choice View, counts desires held in the future, it avoids Imprudence. And since it, unlike the Time-of-Satisfaction View, counts intertemporal desires, it might well have the resources to respond to the Problems of Postmortem Desires and the Never Satisfied.

However, Time-Neutral Aggregation leaves the door open for Tyranny. Plausible distributive principles will imply that, so long as past and future desires for being a poet now are sufficiently

² This may be the juncture to address the bearing of Paul’s (2014) influential work on “transformative experience” on our discussion. Paul’s work, as I understand it, highlights two kinds of problem cases. First, there are cases, like Aleri’s, in which the choice now itself affects what the chooser will prefer in the future (by bringing about a “metaphysically” transformative experience). Our discussion of Disloyalty to come, it seems to me, will address what is distinctive about such cases. Second, there are cases in which the chooser lacks information, at least prior to the choice, about what they will prefer in the future (but, as a result of an “epistemically” transformative experience, will gain such information). While Paul suggests that this deprivation of information is often so severe as to stand in the way of “authentic” choice, I am not convinced that there is a special problem here. In general, we are in the dark, to varying degrees, about what we want—now, let alone in the future or past. But, in general, we are in the dark, to varying degrees, about all sorts of factors that bear on what we have reason to choose now. We do our best, relying on what sources of information we have, including, importantly, the experience and testimony of others. It’s a problem, to be sure, but it’s a problem that was always with us. In any event, I will assume, at least for the sake of argument, that we have the relevant information about what do want or will want or have wanted. The Problem of Time is still with us, it seems to me, even if we have that information.

long-held in comparison to the present desire not to be a poet now, one is rationally required to be a poet now. The benefits to all the past and future time-slices outweigh the costs to the present time-slice. Similarly, Time-Neutral Aggregation implies that, in such a situation, being a poet now would make for a better life. And Time-Neutral Aggregation leads to Disloyalty. Since Aneri's future time-slices will enjoy greater desire-satisfaction if she gives up conservation for police work, that is what she is rationally required to do, even if she is presently committed to conservation.

Time-of-Choice-Centered Aggregation

Time-Neutral Aggregation was the first of two approaches that treat each moment of the agent's life as though it were a different person. The second such approach is Pettigrew's favored approach, which we might call Time-of-Choice-Centered Aggregation. Again, we treat each moment of the agent's life as though it were a different person. Again, all of the desires of a given time slice count, even intertemporal desires. However, not all time-slices are treated *equally* in decisions about what to do. Instead, the desires of time-slices are weighted based on their relation to the present, understood as the time of choice. What it is rational for any time-slice to do is to realize the best distribution of desire-satisfaction across all of the time-slices, with each time-slice *weighted by its relation to the present*.

Since Time-of-Choice-Centered Aggregation counts intertemporal desires, it has at least some resources to solve the Problems of Postmortem Desires and the Never Satisfied. If the weights attached to the potentially Tyrannizing past or future are low enough, then Tyranny might be avoided. Similarly, if the weights attached to a future that might lure us into Disloyalty are low enough, then Disloyalty might be avoided. Finally, if the weights attached to the future are high enough, then Imprudence might be avoided. So perhaps Time-of-Choice-Centered Aggregation ticks all of our boxes?

One structural tension is immediately evident. The weights attached to the future need to be low enough to avoid Tyranny and Disloyalty, but high enough to avoid Imprudence. Why think that there are such weights?

A second structural worry is that it is unclear how Time-of-Choice-Centered Aggregation can be made into a theory of well-being, since, where judgments of well-being are concerned, there is no privileged time of choice.

Further worries come to light when we consider how, according to Pettigrew, the weights should be assigned.

In particular, we should assign weight to past, present, and future selves in line with the degree of connectedness between our present self and those selves. In what follows, we will consider three ways in which one self might be connected to another: the first self might have made a sacrifice from which the second self benefits...; the first and second selves might be cognitively connected, so that they share experiences and cognitive states, such as memories, opinions, and beliefs...; and the first and second might be conatively connected, so that they share many values in common (2020 160).

Consider first Pettigrew's suggestion that the desires of one self have greater weight to the extent that they have sacrificed for the present self. (While it is natural to think that the sacrificing self must be in the past, it might in principle be in the future.) Pettigrew motivates this weighting by first considering a case in which Roshni's grandmother sacrificed in the past to benefit Roshni now and desired, before her death, that Roshni study now. As a result, Pettigrew suggests, Roshni has reason to satisfy her grandmother's desires, by studying now. Pettigrew then draws a parallel to a case in which one's past self has made sacrifices to benefit one's present self. Similarly, he suggests, one consequently has reason to satisfy the desires of this past self. However, this opens the door once again to Tyranny. It seems to me, although evidently not to Pettigrew (169), implausible that there is any reason to be a poet now, even if your past self sacrificed to benefit your present self and desired that your present self be a poet.

Similarly, suppose that your friend sacrificed in your past to make possible some benefit that you enjoy now. Perhaps they planted the trees whose shade you now enjoy. At the time they were madly in love with their ex, wishing that their ex would be happy throughout their ex's lifetime. They subsequently went through a bitter break-up and remarried happily. You now are given an opportunity to repay your friend's sacrifice, by helping their new spouse. Alternatively, as it happens, you could help their ex. Pettigrew's view seems to suggest that you ought to help their ex. After all, it was their former self, not

their present self, who made the past sacrifice from which you now benefit, and what their past self desired was that their ex, not their then unknown future second spouse, benefit.

In sum, then, what Pettigrew says about Roshni's case, that a person has reason to fulfil the postmortem wishes of another person who sacrificed for them in the past, seems reasonable enough. We will in due course try to explain it. But the correct lesson to take from the case, I think, is not that, in general, time-slices have reason to fulfill the wishes of other time-slices who have benefitted them.

What about Pettigrew's suggestion that past or future selves have a say to the extent that they are cognitively or conatively connected one's present self? To be sure, this promises to avoid Disloyalty: that is, to explain why we don't abandon present projects for the sake of greater future desire satisfaction from the pursuit of future projects. The future self who would pursue those other projects is conatively distant from one, and so its interests are discounted. This explains why Aneri is not rationally required to become a police officer rather than a conservationist.

Anything less and the leap in the utility that [Aneri] assigns to conformity as a police officer is not sufficient to outweigh her disapproval of that life in the past and present. One upshot of this is that the rise in utility must be quite substantial, and discounting of the future self on the basis of distance in time and distance in values must be quite small in order to push Aneri towards the choice of which her current self would disapprove (2020 214).

But it seems a mistake to avoid Disloyalty by discounting the interests of future selves. To see this, suppose that Aneri first visits the dentist, who informs Aneri that she will suffer a toothache next year, something that she will want very much next year not to suffer, unless she returns for a filling tomorrow. Aneri makes an appointment. Then Aneri visits the fortune teller, who informs Aneri that she will soon change careers, so that by next year she will have police-officer values. Has Aneri learned anything that gives Aneri less reason to go to the dentist tomorrow? I find it implausible that she has, yet it seems to be what Pettigrew's system implies. This may be an instance of the structural tension noted earlier: that if we reduce the weights attached to the future to avoid Disloyalty, then we are led to Imprudence.

The Harmony View

Might there be some way to tweak the Time-of-Satisfaction View, which so nicely avoids Tyranny and Imprudence, so as to address the Problems of Postmortem Desires, Disloyalty, and the Never Satisfied?

Kristen Bykvist proposes a utilitarian theory, the Harmony View, which requires choosing the “action that will best satisfy (i.e., maximise the total sum of satisfaction of) all personal wants that are grounded in sustained priorities” (2003 125). It might be repurposed as a theory of well-being or rationality, and thought of as a modification of the Time-of-Satisfaction View. The key innovation is the notion of sustained priorities. Bykvist explains that a “person’s priorities at a time are a ranking of her wants in order of intensity.” Simplifying a bit, a “person’s priorities at t_1 for t_2 is sustained at... t_2 iff... either (a) she has no priorities at... t_2 , or (b) her priorities at t_1 for t_2 coincide with her priorities at $[t_2]$ for t_2 .” Thus, a “person’s personal preference at t_1 for something to happen at time t_2 counts iff the person’s priorities at t_1 for t_2 are sustained at... t_2 ” (2003 124).

Perhaps the simplest way of presenting the Harmony View is to list the four relevant kinds of preference and ask whether the Harmony View holds that preferences of that kind count. First, simultaneous preferences count (since $t_1=t_2$, guaranteeing that the ranking at the “two” times is the same). This avoids Imprudence, in just the way that the Time-of-Satisfaction View avoids Imprudence. Second, intertemporal preferences when the two times have different rankings don’t count. This avoids Tyranny. My past self ranked being a poet now over being a philosopher now, whereas my present self ranks being a philosopher now over being a poet now, so the past self’s desires for the present don’t count. In this respect too the Harmony View functions like the Time-of-Satisfaction View. Third, intertemporal preferences when the two times have the same ranking do count, unlike the Time-of-Satisfaction View. This might seem to give us the resources to avoid Disloyalty and to solve the Problem of the Never Satisfied. Finally, postmortem preferences count (since there are no priorities at t_2 , after death), also unlike the Time-of-Satisfaction View. This seems to solve the Problem of Postmortem Preferences.

To illustrate the third kind of preference—namely, intertemporal preferences when the two times have the same ranking—suppose that my desires for eating are as follows:

On Monday, I want +5 I eat the cake on Tuesday, +1 I eat a salad on Tuesday, +1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday

On Tuesday, I want +3 I eat the cake on Tuesday, +1 I eat a salad on Tuesday, +1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday
 On Wednesday, I want +1 I eat the cake on Tuesday, +4 I eat a salad on Tuesday, +6 I eat the cake on Wednesday, +1 I eat a salad on Wednesday

Note that my rankings on Monday and Tuesday for what happens on Tuesday agree with each other, but not with my rankings on Wednesday for what happens on Wednesday. This means, according to the Harmony View, that my want on Monday for what happens on Tuesday counts. None of the other intertemporal wants count. So we have:

On Monday, I want +5 I eat the cake on Tuesday, +1 I eat a salad on Tuesday, ~~+1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday~~
 On Tuesday, I want +3 I eat the cake on Tuesday, +1 I eat a salad on Tuesday, ~~+1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday~~
 On Wednesday, ~~I want +1 I eat the cake on Tuesday, +4 I eat a salad on Tuesday~~, +6 I eat the cake on Wednesday, +1 I eat a salad on Wednesday

Suppose I am given the option of eating the cake on Tuesday or on Wednesday. According to the Harmony View, I should eat the cake on Tuesday, since Monday's (+5) and Tuesday's (+3) wants for Tuesday combine to outweigh Wednesday's (+6) want for Wednesday.

The first thing to notice about this case is that it is not obviously intuitive that I should eat the cake on Tuesday, especially given that I want on Wednesday to eat the cake then twice as much as I want on Tuesday to eat the cake then. So this implication doesn't seem to be an asset and may even be a liability.

The second thing to notice about this case is that the result seems highly unstable, as well as dependent on what might seem to be irrelevant factors. Suppose that between Monday and Tuesday, my desire for whether I have a green salad or a beet salad on Tuesday flips. Then we have:

On Monday, ~~I want +5 I eat the cake on Tuesday, +1 I eat a green salad on Tuesday, +0 I eat a beet salad on Tuesday, +1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday~~
 On Tuesday, I want +3 I eat the cake on Tuesday, +0 I eat a green salad on Tuesday, +1 I eat a beet salad on Tuesday, ~~+1 I eat the cake on Wednesday, +4 I eat a salad on Wednesday~~
 On Wednesday, ~~I want +1 I eat the cake on Tuesday, +4 I eat a salad on Tuesday~~, +6 I eat the cake on Wednesday, +1 I eat a salad on Wednesday

Since my rankings on Monday and Tuesday no longer agree, my Monday desire for Tuesday no longer counts. So it follows that I should eat the cake on Wednesday, rather than Tuesday. The seemingly

irrelevant change of preference for salads between Monday and Tuesday has the consequence that I should eat the cake on Wednesday rather than Tuesday. The Harmony View seems to violate something like independence of irrelevant alternatives.

Although it is not an issue that Bykvist discusses, the Harmony View might initially seem to avoid Disloyalty. Recall that if Aneri chooses to be a police officer, Aneri will come to want to be police officer and will enjoy greater satisfaction than Aneri will enjoy if she stays a conservation officer.

Imagine that the situation is as follows:

Past: +5 conservation officer in the future, +3 police officer in the future
Present: +5 conservation officer in the future, +3 police officer in the future
Future, if chooses to continue as a conservation officer: +5 conservation officer in the future, +3 police officer in the future
Future, if chooses to become a police officer: +9 police officer in the future, +0 conservation officer in the future

If Aneri continues as a conservation officer, then, given that her past and present priorities will have been sustained into the future, she will receive +15 from being a conservation officer in the future. By contrast, if Aneri chooses to become a police officer, her past and present priorities will not have been sustained into the future, and she will receive only +9 from being a police officer. So Aneri is not rationally required to become a police officer, and Disloyalty is avoided. (By contrast, on the Time-of-Satisfaction View, the past and present desires for the future do not count, and so Aneri will receive only +5 from being a conservation officer in the future, but +9 from being a police officer in the future. So, as we saw before, the Time-of-Satisfaction View leads to Disloyalty.)

But this neat explanation seems vulnerable to irrelevant alternatives. Suppose that in the future, Aneri will switch from preferring vanilla to chocolate to the reverse.

Past: ~~+5 conservation officer in the future, +3 police officer in the future, +1 vanilla in the future, +0 chocolate in the future~~
Present: ~~+5 conservation officer in the future, +3 police officer in the future, +1 vanilla in the future, +0 chocolate in the future~~
Future, if chooses to continue as a conservation officer: +5 conservation officer in the future, +3 police officer in the future, +1 chocolate in the future, +0 vanilla in the future
Future, if chooses to become a police officer: +9 police officer in the future, +0 conservation officer in the future, +1 chocolate in the future, +0 vanilla in the future

In this case, Aneri's priorities will not be sustained into the future, and so her past and present desires to be a conservation officer in the future will not count. Aneri is required to be Disloyal.

Might the Harmony View have resources to solve the Problem of the Never Satisfied? True, the Harmony View allows some intertemporal preferences to count. But the preference to have written the book need not be embedded in sustained priorities. Suppose:

At t_1 , I want ~~+10 to have written the book at t_2 , +0 to watch a movie at t_2 .~~
At t_2 , I want +0 to have written the book at t_2 , +1 to watch a movie at t_2 .

Since priorities at t_1 are not sustained into t_2 , the desire at t_1 to have written the book at t_2 does not count. Given a choice between watching a movie at t_2 or having written the book at t_2 , I should watch the movie.

Does the Harmony View at least solve the Problem of Postmortem Desires? Again, all desires for what happens after one's death count, because one has no priorities after one's death, and "sustained priorities" are disjunctively defined to say that priorities are sustained at a time if either one has no priorities at that time or... such and such. However, it seems somewhat ad hoc to solve the Problem of Postmortem Desires simply by adding that disjunct.

In any event, it isn't clear that the Harmony View solves the Problem of Postmortem Desires in an appealing way. One difficulty is that the Harmony View implies that all postmortem desires count equally no matter when one has those desires. Intuitively, however, later postmortem desires take priority, or even completely replace, earlier postmortem desires. Suppose that during my first marriage, from age 30–60, I wanted that my first spouse be provided for after my death, even if we should no longer be together, whereas during my second marriage, from age 60–80, when I die, I wanted, with the same intensity, that my second spouse, and my second spouse alone, be provided for after my death. According to the Harmony View, there is more reason, on my behalf, to see to it that my first spouse is provided for, since I held that preference for a longer stretch of time. Bykvist's response to this problem is to say that my later preference is, because it is later, better informed (2003 131). But it is not clear why one's later

preference must always be better informed. Indeed, in this example, it is not clear what there is to be better informed about.

A second difficulty is that because the Harmony View counts all postmortem desires, while discounting many other intertemporal desires, it allows postmortem desires to swamp other intertemporal desires in counterintuitive ways. Suppose that every day for my last 1000 days, I want very slightly +1 that my estranged son attend my funeral and very strongly +900 that he visit me before my death so that we can be reconciled. Suppose that my priorities for my last day otherwise change the night before my last day. In that case, only my simultaneous preference on my last day that my son visit me on my last day counts. By contrast, all of my postmortem preferences that he attend my funeral count. So according to the Harmony View, if someone can buy my son either a ticket to visit me on my last day or a ticket to my funeral, but not both, they should buy him a ticket to my funeral, which satisfies a 1000-day-long desire of intensity +1, rather than a one-day-long desire of intensity +900. This is so even though I have had the latter desire, of +900, for 1000 days.³

³ In fairness to Bykvist, we should note that he suggests that the Harmony View is needed to deliver what he takes to be the intuitive conclusion in the following case.

You have known Anna for fifty years, and for as long as you have known her, Anna has been dying to read a certain rare book. Now, another friend of yours, Eric, has recently formed a desire to read the same book. This new desire of Eric's is of exactly the same intensity as Anna's. Now, today you happen to come across a single copy of the book in a bookstore. If you assume that reading the book will affect your friends' future happiness (in terms of total preference satisfaction) equally, for whom should you buy the book? I think there are reasons to give the book to Anna (2003 128).

But, first, it is unclear to me that, intuitively, there is more reason to give the book to Anna than to Eric. So it is unclear whether the Harmony View's implication that we should give the book to Anna is an asset rather than a liability. Consider fleshing out the example in the following way:

Anna and Eric have wanted to read in retirement, at age 70, having had little time during their busy working lives to do so. Eric wanted from age 20–45 to read *War and Peace*, and nothing else, in retirement, then from age 45–70 to read *Ulysees*, and nothing else, in retirement. Anna wanted from age 20–70 to read *Ulysees*, and nothing else, in retirement.

Why should the fact that Anna's mind settled on *Ulysees* earlier than Eric's mean, as the Harmony View implies, that she has a stronger claim to the sole remaining copy? Or consider the following version of the case:

Anna is impatient. She has desired at age 20 to read *Ulysees* at age 20, without any thought to reading it at other times, then at 21 to read it at age 21, and so forth. Now at 70 she desires to read it at 70, having for the first time a desire to read it at that age. By contrast, Eric has had no interest in reading until age 69, at which age he desired to read *Ulysees* at age 70. He now desires at 70 to read it at 70, and his priorities are sustained from 69 to 70.

According to the Harmony View, we have more reason to give the book to Eric. Again, is this clear? One might just as soon have the intuition that we have more reason to give it to Anna or that it is a tie. I fear that our intuitions about the case, or cases, of Eric and Anna are just a muddle.

In any event, if we have the intuition that we should gift the book to Anna in the original case, it is not clear that the Harmony View supports it in a robust way, for reasons that may now be familiar. Suppose that Anna has

A problem even for the objectivist?

The Problem of Time, then, presents real difficulties for desire-satisfaction theories of rational choice and well-being. We have struggled to find a view that avoids Tyranny, Imprudence, and Disloyalty, while solving the problems of Postmortem Desires and the Never Satisfied. Now, again, my first thought was: “So much the worse for the desire-satisfaction theories of rational choice and well-being!” But on further reflection, I worried. It is not clear how a value-based theory of rational choice or an objective-list theory of well-being escapes the difficulties. Why, for instance, does one have reason to be a philosopher now, rather than a poet? Presumably, because one wants now to be a philosopher now and not a poet. But one wanted in the past to be a poet now and not a philosopher. So why should one’s present want take precedence over one’s past want? How do we avoid being thrown back on the same dialectic?

To be sure, the objectivist might deny that the fact that one cares about being a philosopher now is itself a reason to be a philosopher now. Instead, the reason to be a philosopher now is that, given that one cares about being a philosopher now, being a philosopher now promises more pleasant experiences. But the fact that you want to be a philosopher now seems to be a reason even if being a philosopher does not promise more pleasant experiences. We might imagine, for example, that if you pursued poetry, you will be given a pleasure-inducing drug to make up for the hedonic shortfall. Similarly, the objectivist might deny that the fact that one cares about being a philosopher now is itself a reason to be a philosopher now. Instead, the reason to be a philosopher now is that, given that you care about being a philosopher now, you are more likely to succeed in being a philosopher now. But, again, we can imagine that this is not so. After all, caring more about something can make you less likely to succeed. In any event, the fact that one is less likely to succeed if one *tries* to be poet than if one tries to be a philosopher is only reason not to *try* to be a poet rather than a philosopher. It isn’t reason not to *be* a poet rather a philosopher. Yet one’s caring seems to give one such reason. And if that doesn’t seem sufficient, then consider a slightly revised

preferred vanilla to chocolate for the last fifty years, but yesterday started to prefer chocolate to vanilla. In that case, her priorities have changed, and so the Harmony View ignores her preferences before yesterday for the book. Eric started to want the book the day before yesterday and there has been no change in his priorities. So, according to the Harmony View, we should gift the book to Eric, who has wanted the book for only two days, rather than to Anna, who has wanted it for fifty years.

Aneri. This Aneri* will enjoy even greater experiential benefits and greater success as a police officer than as a conservation officer. Nevertheless, the fact that Aneri* cares, now, about being a conservation officer can make it reasonable for her to remain a conservation officer (compare Raz 1999 316–17).

On the objectivist view, then, the fact that one cares about being a philosopher rather than a poet is a reason to be a philosopher rather than a poet, independent of considerations of experience or success. But why is it a reason? The basic thought is this. When one engages in an independently worthwhile pursuit that one cares about, this complex constitutes something further of value, which we might call a “personal project.” This further value has both normative and prudential significance. Once one cares about a project, until such time as it is replaced, one has “project-based reasons” to “succeed” in the project—that is, to realize what one cares about in caring about the project—and to continue caring about it going forward. And once one cares about a project, until such time as it is replaced, success in the project contributes to one’s well-being.

If it seems contrived to suggest that caring about an independently worthwhile “object” can constitute something of further value, consider friendship. The fact that one has the attitudes of friendship toward someone, caring in a distinctive way about them, partly constitutes something further of value, namely, a friendship. Far from a contrived suggestion, this seems a commonplace. As with projects, once the friendship has begun, until such time as it is replaced, one has friendship-based reasons for the distinctive pattern of behavior and caring in which a friendship consists. And once the friendship has begun, until such time as it is replaced, the flourishing of the friendship (and perhaps of one’s friend) contributes to one’s own well-being. The same might be said of other valuable “attitude-dependent” relationships, such as those of romantic love, which exist only insofar as the people in them have certain attitudes toward each other.

In these cases, of projects and attitude-dependent relationships, the normative and prudential significance of what we care about commences with, and only with, caring and lasts until replacement and no further. To “replace” a project or relationship—to finally define an imperfect term of art already used several times now—is not just to cease to care about it, as might happen through depression, a coma, or

death. Instead, it is to cease to care about it, followed by a sufficiently rich conative life organized around other projects or relationships.

In a certain philosophical mood, one might puzzle about why this should be. Why do you have reasons, say, to spend time with your friends while they are your friends, rather than before they are your friends or after they are your friends? The immediate answer is that so doing would not constitute a friendship. If it is asked, again in a certain philosophical mood, why it couldn't constitute what we might call a "schmendship," I suppose the answer is that it could. If it is then asked why schmendships aren't just as valuable as friendships, it is hard to know how to answer. Why are the things that are valuable for their own sakes valuable, whereas the things that are not valuable for their own sake not valuable?

An attentive reader may have noticed that we have recently shifted from talk of "desiring" or "wanting" or "preferring" to talk of "caring." This is deliberate. Following Scheffler (2010), I believe that insofar as we are thinking of "wanting" as a component of valuable projects and relationships, wanting should primarily be thought of as caring about something (or its mattering to one), where caring about something consists in being "emotionally vulnerable" to it. One is liable to feel, say, grief if one believes that what one cares about has been harmed, frustration if one believes that it has been hindered, anxiety if one believes that it is at risk, elation if one believes that it is thriving, and so on.

In normal cases, one does not simply care about a project, in the sense of simply being emotionally vulnerable to it. Rather, again following Scheffler 2010, one values it. That is to say that, in addition to caring about it, one also has certain normative beliefs. One believes that the underlying pursuit is independently valuable, such that one has reasons to care about it (that one's caring about it is not misplaced) and, perhaps, one has reasons to pursue it even if one does not care about it. Further, one believes that, because one does care about it, it constitutes something of further value (which we are calling, somewhat artificially, a personal project) which gives one reason to succeed in it, gives one reason to continue caring about it going forward, and gives others reason to respect it, e.g., not to interfere with or undermine it. Something similar might be said of attitude-dependent relationships.

While I focus on projects and relationships, I should note that they may be species of a genus with the same structure, in which caring about something of independent value constitutes something of further value, which is then normatively and/or prudentially significant. Other possible species are causes, identities, traditions, ways of life, and memberships in groups.

Avoiding Tyranny and Imprudence

This approach avoids Tyranny straightforwardly. To pursue poetry now, when one has replaced it with philosophy, would not constitute a valuable project, any more than spending time with a stranger would constitute a friendship. So one does not have project-based reasons to pursue poetry. Nor would doing so contribute in the same way to one's well-being.

At the same time, this approach avoids Imprudence. Future projects do not give one *project-based* reasons now, since nothing done now, when one doesn't yet care about the relevant pursuits, would amount to the pursuit of a valuable project. However, future projects do give one *prudential* reasons now, since things done now can help one to pursue those projects in the future, when one will care about them, doing which will contribute to one's well-being. For example, one might have reason not to expend all of one's resources on one's current passion for rock-climbing, when one knows that in old age one will replace that passion with a fondness for gardening. Although one does not now have project-based reasons to garden, one nevertheless has prudential reasons to make provision for it, lest one's well-being in old age suffer.⁴

⁴ "But doesn't something like Tyranny reassert itself?" it might be objected. "Grant that we have prudential reasons, to promote our own well-being. Grant that our well-being depends on the success of our projects, including our past projects. Grant that, at least sometimes, there are things that we can do now to make our past projects successful. Doesn't this mean that we have reason to see to it now, when we can do so, that our past projects were successful, just as we have reason to see to it now, when we can do so, that our future projects will be successful? Suppose that one has shifted from the project of poetry to the project of philosophy, leaving some of one's poems unfinished. One did not fully succeed in one's prior project of poetry. Suppose then that one could now take one day a week away from philosophy to finish one's poems, in order to make one's past project of poetry a success. Does one have reason to do this, and moreover prudential reasons, so that one will have lived a better life? That seems implausible, whereas it seems perfectly plausible that one has reason to take one day a week away from philosophy to provide for what one will do after one's retirement as a philosopher, so that one will live a better life. The worrisome possibility, to be clear, is not that one has reason (conflicting with one's reasons to be a philosopher now), to be a poet now in order to satisfy one's past desire that one be a poet now. Instead, the worrisome possibility is that one has reason (conflicting with one's reasons to be a philosopher now) to be a poet now in order to make one's past poetry a success, which will make it the case that one will have lived a better life."

Solving the Problem of Postmortem Desires

How do we solve the Problem of Postmortem Desires? Why does the fact that I desire that my work be completed, say, after my death give me reason to see to it that it is completed after my death, mean that my well-being is promoted by its being completed after my death, and give others reasons to see to it that it is completed after my death?

First, my project-based reasons are ones that I have now only because I care now about my project. Nevertheless, they are reasons to succeed in my project, or to realize what I care about in caring about the project, not reasons to satisfy my future desires. So there is no difficulty in saying that I have project-based reasons to see to it that my project is completed after my death. The fact that doing so will not satisfy a desire that I have after my death is neither here nor there. Second, success in my project contributes to my well-being until such time as it is replaced. A project that I have when I die is not replaced. So the success of such a project after my death contributes to my well-being. So, third, others have reasons, based in my well-being, to see to it that such projects succeed.

This avoids two of the difficulties that confronted Bykvist's Harmony View. First, on our account, it is only postmortem desires held up until death, not postmortem desires replaced before death, that count. So there is not more reason, on my behalf, to see to it that my ex is provided for after I die, just because I spent more time over the course of my life desiring that. Second, our account does not pose the same problem of swamping by postmortem desires. There is nothing in our approach that implies that there is more reason, on my behalf, to buy my son a ticket to my funeral than, as I vastly prefer, to buy him a ticket to my bedside in time for us to be reconciled.

This approach also explains why Pettigrew's Roshni has reason to do what her grandmother wanted when she died, without implying that one has reason to do what one's own past self wanted, which opened the door to Tyranny. In the former case, her grandmother's concern for Roshni's education

The mistake in this line of thought is to grant that our well-being depends on the success of our past projects—or at least that it depends on the success of *replaced* projects when that success is realized *after* they have been replaced. Again, the realization of what we care about in projects and relationships contributes to our well-being only if it is realized before such time, if any, when we have replaced the project or relationship.

was not replaced, whereas in the latter case, one's past self's concerns were replaced. (Pettigrew anticipates this response, but rejects it on the grounds that the past selves' concerns were not replaced, since those selves, like Roshni's grandmother, went out of existence (169–70).) Similarly, this approach avoids the implication that you have reason to benefit your friend's past, tree-planting self by benefiting their now ex.

Why does well-being work this way? Why does the realization of the objects of our caring still contribute to our well-being when our caring is terminated by depression, coma, or death, but not when it our caring takes different objects? It is a difficult question to answer. But part of the story may have to do with the importance to us of being able to define for ourselves what our good is, what contributes to our well-being. In replacing what we care about, we redefine what our good is. When we merely cease to care through depression, coma, or death, we do not redefine what our good is.

Avoiding Disloyalty

Let us turn from the Problem of Postmortem Desires to Disloyalty. For us objectivists, the question is: Why is it reasonable, at least within limits, to stick with the projects we have, like Aneri*, rather than to replace them with projects that promise greater well-being? On our account, this is explained by the competition of two kinds of reasons. On the one hand, there are *prudential* reasons, concerning one's well-being, which argue for replacing the project: for Aneri*'s becoming a police officer. On the other hand, there are *project-based* reasons to continue with the project one has: for Aneri*'s continuing as a conservation officer. So long as the loss of well-being isn't too great, the project-based reasons hold their own against the prudential ones, and it is at least reasonable to stick with the project that one has.

This explanation differs from Pettigrew's. For Pettigrew, the self that would result from pursuing the project that promises more well-being is conatively further from one's present self, and so its interests are discounted. But, again, this seems to get the wrong result in the case of Aneri's visit to the dentist and then the fortune teller. It implies that having learned from the fortune teller that she will have, in a year's time, the new values of a police officer, rather than her current values as a conservation officer, she has learned of reason to cancel her next appointment with the dentist. On our approach, by contrast, we do

not discount the interests of selves that are more distant from us in time or values when considering prudential reasons. Instead, we recognize that project-based reasons are distinct from, and potentially in competition with, prudential reasons in general, without any discounting of prudential reasons relative to the present. Project-based reasons compete with prudential reasons of the further future just as much as they compete with the prudential reasons of the nearer future just as much as they compete with the prudential reasons of the present.

Solving the Problem of the Never Satisfied

Finally, let us turn to the Problem of the Never Satisfied. What do we say of Pettigrew's novelist, who desires having written the book until such time as they have written it, at which point they will no longer desire having written it? Why do they have reason to write the book? And why does doing so contribute to their well-being?

Again, on our view, one has project-based reasons to pursue a project that one has not replaced. The novelist does not replace the project before it is completed. After all, the novelist completes it! So, the novelist has project-based reasons throughout. And on our view, granted, success in a project that occurs after one has replaced it does not contribute to one's well-being. But success in a project that occurs before one has replaced it does contribute to one's well-being. So writing the novel does contribute to the novelist's well-being.

To be sure, something is missing from their life. Recall that we are understanding desiring as caring, that is, as being emotionally vulnerable. So, as we are imagining the example, upon completing the book, the novelist ceases to be emotionally vulnerable to its completion. They feel no satisfaction in finishing it. It no longer means anything to them, and they are on to the next thing. That's a shame, in that they are denied pleasant experiences, which would make their lives better. But that's compatible with what we have said: that they have reason to complete the book and that completing it contributes to their well-being.

From the Problem of Time to the Problem of Origin

By way of interim conclusion, we return to my discussion with the AI-safety folks. What was their worry about AI's manipulating our preferences, if not that AI might change our preferences for the worse? One possible worry, recall, was that we want in the present that AI not change our preferences in the future. Hence, by changing our preferences in the future, AI would be frustrating our preferences in the present. Can we make sense of this worry, or a worry like it, on our objectivist account?

On our objectivist account, insofar as preferences make a normative or prudential difference, it is because they partly constitute things of further value: projects and relationships. The preference that AI not modify our preferences in general does not partly constitute a project or relationship. So the preference for AI's not modifying our preferences in general does not make a normative or prudential difference.

If our objectivist account registers a worry in the vicinity—having to do with a conflict with our present preferences and distinct from the worry that AI might change our preferences for the worse—it is that AI might somehow modify our preferences for projects and relationships, so that we replace our present projects and relationships with no less valuable projects and relationships. On our account, we have, like Aneri*, project- and relationship-based reasons now to stave off that change. It is an open question, though, whether we would be worse off as a result of the change and so have prudential reasons to stave it off. Perhaps, again like Aneri*, our well-being would not suffer.

Of course, this still leaves the other interpretation of the worry about AI's modifying our preferences: namely, that the goal is just satisfying preferences formed in the right way. And that leads us from the Problem of Time to the Problem of Origin.

Two guiding examples

I begin with two guiding pairs of examples. First pair: Both Julie and Eunice know that having a career and being a homemaker are both perfectly and equally good lives to lead. Living in a perfectly just society, Julie, having desired a career, comes in time to prefer being a homemaker. Call this the “just case.” By contrast, living in an unjust society, Eunice, is unjustly, albeit only negligently, denied the opportunity for a career. As a result, she loses her desire for it. In time, she comes to prefer being a

homemaker instead, exactly like Julie. Call this the “unjust case.” (For ease of expression, I shift back from talk of “caring” to “desiring,” “wanting,” and “preferring.”)

I doubt that Eunice’s positive desire to be a homemaker is problematic. (As we will see, it belongs in the same category as the desire of Enoch’s captive pianist or Pettigrew’s second parent, characters we will meet shortly.) But Eunice’s loss of the desire for a career, simply because she cannot have it, does seem problematic in a way in which Julie’s loss of the desire was not.

What is problematic, it seems to me, is not simply that Eunice has been denied a career, but moreover that this has had the consequence of extinguishing her desire for it. Contrast the case in which she loses the desire for the career with a new case in which she retains that desire, despite its never being fulfilled. To be sure, there is something in one respect better about losing it, which has to do with the “state-given” costs of the desire, which we will discuss. But there also seems to be something worse, which is what we are trying to characterize.

Now the second pair of examples, which is admittedly more contrived. Suppose that, as is your desire, you keep a found-object sculpture garden. It’s not to my taste, and I, your neighbor, wish that you kept a plant garden instead. But I hold my tongue. Over time, though, you find that, more and more, the idea of plant gardening appeals to you. The reasons in favor of plant gardening, reasons of which you were perhaps already aware, call out to you in a new way. Eventually you give up your sculpture garden for a plant garden. This is no less worthwhile an activity, and you are no worse off for the change. I rejoice in my good luck. What accounts for the change? If we were to ask an omniscient, we would find that it was a vast confluence of random, chaotic factors, which were not the result of anyone’s doing, or at least anyone’s intentional, reckless, or negligent doing: the pH of the water you drank at some point, the angle at which the sun hit you at another point, etc. All of these things, a monumental cosmic stew, combined to bring about the neurological change that made you newly susceptible to the reasons in favor of plant gardening. Call this the “organic case.”

Now, for a contrasting case, imagine that I deliberately induce what is otherwise the same change in desires, hating your sculpture garden as I do. Whatever the neurological change was that brought about

your conversion in the first case, I gradually engineer the same change, by brain surgery or drugs. The desire thus induced seems problematic, in a way in which the “organic” desire in the previous case does not. Call this the “induced case.”

The problem with the desires in the unjust and induced cases has to do, it seems, with their origin. The problem cannot be a matter of their objects: what they are desires for. The object of the desire is the same as in the just and organic cases. Nor is the problem with the desires in the unjust and induced cases a matter of the other attitudes that they are or are not accompanied by, such as reflective endorsement. The accompanying attitudes are all the same as in the just and organic cases, respectively. Or might it be said that there is still a difference between the induced and organic cases? In the induced case, if you are fully informed, you will reject the desire, precisely because it was induced. It will conflict with what Stefan Gosepath calls your “Entstehungspräferenzen,” metapreferences for how your preferences arise (1992 374–5). But what if your metapreferences are not opposed to induced preferences, perhaps because they have themselves been induced? True, one might insist that, whether or not you *do* reject the desire, you *ought* to, because it was induced. But that’s just to say that the origin of the desire makes a normative difference, which is just what we are suggesting.

If the problem with the desires in the unjust and induced cases has to do with their origin, what is it about their origin that accounts for the problem, and why is the problem a problem? More generally, it seems that the process by which we arrive at a given conative state—by which I mean not only the presence of desires of various strengths but also their absence—can make that conative state problematic. What sort of processes lead to conative states that are problematic in this way, and how are they problematic?

Amor fati

Consider Aesop’s fox. At the start, the fox wants the grapes. But then it realizes that it cannot reach the grapes. As a result, it ceases to want the grapes. (In the story, the fox convinces itself that the grapes are sour and so not worth wanting. But, to have a cleaner case, let us assume that the fox simply ceases to want the grapes without any rationalizing belief.) It would be one thing if the fox, upon realizing that it

cannot reach the grapes, continued to want the grapes, but simply decided not to try to reach them. But the fox ceases to want the grapes at all. Because the fox realizes that it cannot reach the grapes, it ceases to want the grapes even if it could reach them. Let us call this process “amor fati.” We could call it the formation of “adaptive” preferences, but since people use the label, “adaptive,” to mean quite different things, let’s stick to “amor fati.”

Amor fati should seem familiar. It is the process by which Eunice loses her desire for a career. Might that fact be the problem with her loss of desire? To be sure, that couldn’t be the problem with your induced desire for the plant garden. That does not result from amor fati. But we might look to make progress where we can.

What seems interesting about Aesop’s fox is the following structure.

(A) S knows that D+ is unavailable, whereas D- is available

That is, the fox knows that eating the grapes is unavailable, whereas not eating the grapes is available.

(B) Because of (A), S knows that preferring D+ to D- will lead to “dissatisfaction,” whereas preferring D- to D+ will not.

Because the fox knows that eating the grapes is unavailable, whereas not eating the grapes is available, the fox knows that preferring eating the grapes to not eating the grapes will lead to “dissatisfaction,” whereas preferring not eating the grapes to eating the grapes will not.

(C) Because of (B), S prefers D- to D+.

Because of this knowledge, the fox prefers not eating the grapes to eating them.

The structure (A)–(C) seems an instance of an even more general structure:

(1) S believes that in the present state, W, D+ is less available or less probable or more costly relative to D- than D+ would be in some other state, B.

For example, an aspiring musician believes that, as things are, being a musician is more costly relative to being an academic than it would be if they had more musical talent.

(2) Because of (1), S believes that there are drawbacks in W to desiring D+ more intensely relative to D- than there would be in B.

These drawbacks of desiring D+ more intensely relative to D- might be:

- a. that it makes S more likely to act in ways (such as to attempt to attain D+) that are not worth the cost given the chances of success,

- b. that it makes S more likely to endure unpleasant experiences of dissatisfaction, frustration, disappointment,
- c. that it makes S more likely to have stronger unfulfilled desires (assuming that this is a bad thing in itself).

For example, the aspiring musician believes that there is a drawback, as things are, in desiring being a musician more intensely relative to being an academic: namely, that this makes the aspiring musician more likely to attempt to be a musician, which is not worth it.

- (3) Because of (2), S desires D+ less intensely relative to D- than S would if S believed they were in state B.

For example, the aspiring musician desires being a musician less intensely relative to being an academic.

Let us revise our notion of amor fati to cover all instances of this structure.

This is to suggest that amor fati—or at any rate a phenomenon relevantly like Aesop’s fox—can arise even if there is no belief that the object of desire is strictly unavailable. As the example of the aspiring musician shows, amor fati can also arise from beliefs that some end would be more costly or less likely to be achieved. To take another example, suppose that the fox ceases to desire the grapes at all in response to the information that the grapes, while strictly available, cost \$1 (or are somewhat unlikely to be reached). In other words, the fox, believing that the grapes cost \$1 (or are somewhat unlikely to be reached), ceases to desire the grapes at any cost—even if they were free (or sure to be reached). This change in desires would be relevantly similar, it seems, to the originally described change, which was induced by the belief that the grapes were strictly unavailable. Just as it is amor fati to cease to desire the grapes if they were available upon learning that they are unavailable, likewise it is amor fati to cease to desire the grapes if they were free upon learning that they have some cost (or are somewhat unlikely to be reached).

If this is what amor fati amounts to, in what way is it problematic? First, it might be observed that the drawbacks constitute “state-given” reasons rather than “object-given” reasons against desiring D+. The fact that I cannot attain D+, and that because of this, my desire for D+ will lead to dissatisfaction, does not make D+ less desirable, although it makes *desiring* D+ less desirable.

However, it is not clear that the fact that certain desires are responsive to state-given reasons should make them problematic in the way that we are trying to characterize. Suppose that a career in music simply leaves me cold whereas a career in academia grips me. If this is so for *no* reason at all, there need not be anything amiss. Why, then, is there something amiss if it is so (that music leaves me cold whereas academia grips me) for *a* reason, namely the state-given reason that my lack of talent makes caring more about a career in music too costly?

Compare other optional attitudes, such as a well-supported belief that my last dental visit in 2005 was scheduled for 1pm. If I believe it, I don't do so against reason. But I also no longer have any compelling reason to have an opinion on the matter. Presumably, nothing would be amiss if I ceased believing it for no reason at all. Why then is something amiss if I cease believing it for the state-given reason that, as Gilbert Harman would put it, it is "cluttering" my memory (1986)—if some sub-personal mechanism that is sensitive to the state-given consideration that it is cluttering my memory leads me to cease believing it? If nothing is amiss, then why is something amiss if some sub-personal mechanism that is sensitive to the state-given consideration that my lack of talent makes desiring a music career too costly leads me to cease desiring it? So responsiveness to state-given reasons does not seem to be a problem.

Second, it would be a problem if I were to desire D- in a way that is unreasonable given the object-given reasons against desiring it. Suppose I am a captive, and I have, as I know, the option of endlessly remeasuring the size of my cell and the option of reading great works of literature. It would be unreasonable for me to desire remeasuring over reading great works of literature. Suppose now that my captors take away my books. Due to *amor fati*, I then desire remeasuring to reading great works of literature, even if I had the option of reading great works of literature. This is again unreasonable given the relevant object-given reasons. One problem, then, with *amor fati* is that it may lead to desires that are unreasonable in light of the relevant object-given reasons.

Third, S's desires may be problematic because they are what I will call "improvable": because we can make S better off by not satisfying S's desires, but instead by changing S's desires and then satisfying the changed desires. For example, the fact that the fox prefers not eating the grapes may mislead us into

thinking that the best we can do for the fox is to bring it about that it not eat the grapes, when it is possible that we can make the fox better off by making the grapes available to it and informing it of this. This would be so if, once informed, the fox will once again desire eating the grapes to not eating them, and the fox will be better off eating the grapes with that desire than not eating them with the desire not to eat them.

Generalizing somewhat, suppose that S has the structure of desires in (A)–(C). Suppose that one course of action open to us is to satisfy S’s desire for D⁻. Another course of action open to us is to make D⁺ available to S and to inform S of this, in which case S will avail themselves of D⁺. It may be the case, although nothing in the (A)–(C) structure guarantees it is the case, that S will be better off if we take the second course of action. The problem with S’s desires is that they may mislead us into thinking that the first course of action is the best we can do for S. After all, S currently prefers D⁻ to D⁺, which is to say that S currently prefers D⁻ to D⁺ even if D⁺ were available. The trouble is that that is compatible with its being the case that if S knew that D⁺ were available, then S would prefer D⁺ to D⁻.

Generalizing further, suppose that S has the structure of desires in (1)–(3). Suppose that one course of action open to T (perhaps S themselves) is to endeavor to satisfy S’s desires given that S truly or falsely believes they are in W. Another course of action open to T is to get S to believe truly that they are in B, which would change their desires, and then endeavor to satisfy S’s new desires. Suppose finally S would be better off if T took the second course of action. The problem with S’s desires is again that they may mislead T into thinking that the first course of action is the best that T can do for S.

Finally, consider a musician who has the talent for music and who is pursuing music, preferring it to philosophy. Now suppose that this preference would be reversed as soon as they learned that, were they to pursue philosophy with this reversed preference, they would be more likely to succeed in it, at lower cost—or more generally that they would enjoy greater well-being. One might wonder whether they “really valued” a life in music. This is the phenomenon of Disloyalty that we considered earlier. So a further potential reason why amor fati might be problematic, then, is that it might lead in this way to disloyalty to projects.

So there are three potential problems with amor fati: that it leads one to desires that are unreasonable given the object-given reasons for those desires, that it leads one to improvable desires, and that it leads to disloyalty to projects.

The difficulty is that Eunice's amor fati needn't involve any of these problems. We might imagine that being a homemaker is no less valuable than having a career, so that the preference for being a homemaker is not unreasonable given the object-given reasons for that preference. We might imagine that her desire is not improvable. Granted, it may be the case that, if she came to know that the career was available to her, she would indeed desire it and pursue it. But imagine that that would make her no better off: that she is as happy as a homemaker with her desires to be a homemaker as she would be with a career with her desires to have a career. And we might imagine that there is no disloyalty to projects. Perhaps the desire for the career was extinguished only after a hard struggle to maintain it.

Enoch's account

Here we might look for help from a different literature, on consent. In this literature, it is suggested that the origin of a desire can matter in a further way: namely, it can affect whether consent that the desire motivates is valid. And the problem is not limited to consent. It applies just as well to promises and other transactions which can be valid or invalid. Indeed, David Enoch (2020) makes a suggestion that might at first seem to account for our cases. The suggestion is that a desire does not support consent when it results from unjust treatment of the person with the desire. As he puts it: "an important class of cases of nonautonomous preferences is those that were shaped (in the appropriate way) under the causal influence of *unjust* conditions, conditions that violate the rights or entitlements of the relevant agent" (185).

To be sure, this accounts for our examples, as they stand. In Eunice's case, but not Julie's, the loss of the desire for a career is the result of injustice to Eunice. In the induced case, but not the organic case, the formation of your desire for a plant garden is the result of injustice done to you: that is, my performing surgery on you or drugging you without your consent.

Enoch's proposal also accounts for his leading example. This is the self-starving wife, who has been brought up to malnourish herself in order that the male members of her family, who already have

enough to eat, have even more. The contrast here is with “the woman who chooses to malnourish herself in order to benefit her husband, but does this in a context where such a decision is neither common nor expected, and where there are no gender norms that push her in that direction; perhaps she does this out of the purest love (which she fully identifies with).” Whereas the indoctrinated wife’s desire to sacrifice herself is problematic, the loving wife’s desire to sacrifice herself is not, at least not to same degree. The difference seems to be, Enoch argues, in what causes the desire: in particular, that the first desire is brought about by injustice, while the second desire is not. For Enoch, the problem with the change in desires is that this compromises the desirer’s autonomy, which in turn compromises their consent. Something about this seems to me absolutely right, as we will return to.

As it stands, however, Enoch’s proposal includes too much. First, consider people, such as Frederick Douglass or Harriet Tubman, who come to desire to resist injustice precisely because they learned firsthand what it was to be unjustly treated. Their consent to take risks or make sacrifices to resist injustice was surely valid. Their desires were not problematic, although of course the conditions that gave rise to them were.

Second, Enoch himself imagines someone who falls in love with playing the piano after being unjustly locked in the cell containing only a piano, and he concludes that the desire would not be problematic.

[I]f someone kidnaps me and holds me in a room for many years; if there happens to be a piano in that room, and not much else; if I develop a preference for playing the piano; and if after being released I retain my preference for playing the piano—then it seems like the injustice of the kidnapping, though a part of the causal history of the piano playing preference, does not render it less than fully autonomous (189)

Thus, Enoch appears to grant that the captive pianist’s consent, after their release, to undergo surgery in order to be able to continue to play, or to part with money for a tuning, is valid. This seems to be true even if they were deliberately locked in the cell in order to get them to fall in love with the piano, by someone who knew for a certainty that it would happen. Still, the desire would not be problematic.

Pettigrew (ms) gives a case that I take to illustrate the same point. A parent, Holt, is forced to have a first child by the fear of ostracism. However, once they have the child, they come to love being a

parent and so want a second child. Whereas Holt's consent to, say, a medical procedure to have the first child would not have been valid, Holt's consent to a medical procedure to have the second child is valid. This is so even though Holt was forced to have the experience that resulted in the desire for the second child, just as the piano player was forced to have the experience that resulted in the desire to play the piano.

Enoch's reply, it seems, is that all of these etiologies, in which unjust treatment of the desirer leads to a desire that is nonetheless unproblematic, involve "deviant" causal chains (188–9). Consequently, they do not represent counterexamples to his thesis. But unless more is said to flesh out the sense in which they are deviant, this seems like special pleading.

There is also a worry that Enoch's proposal includes too little. Recall that the self-starving wife, as I told the story, has the desire simply because she was indoctrinated to have it. Mothers, suppose, tell their daughters, before they marry, that they ought to sacrifice themselves, and daughters believe it. Suppose that no one except the wife herself will ever know whether or not she malnourishes herself in this way, and so she will not be shamed, let alone punished, for not sacrificing herself or not having the desire to do so. Still the desire seems problematic.

Does the wife's indoctrination count as unjust treatment? The difficulty is that her indoctrination consists merely in her mother's advice. And giving someone sincere advice is not, unless more is said, treating them unjustly. Now it might be replied that when the content of that advice is for the advisee to make themselves a victim of injustice, giving it is treating the advisee unjustly. But what licenses us in saying that the wife makes herself a victim of injustice? After all, if she did it from love, she would not be a victim of injustice. The reason why she is a victim of injustice—it seems Enoch must say—is that she sacrifices herself from a desire *caused by unjust treatment*. But whether the indoctrination counts as unjust treatment is precisely what is at issue.

We could make the same point by further specifying the induced case. Suppose that I don't perform surgery on you or drug you. Instead, subliminal advertising style, I place a flashing light that reads, "Plants!" in my bedroom window, set to flash too quickly for you to consciously notice, knowing

and intending that this will have the effect of inducing a conversion. The means that I use are otherwise permissible; I am otherwise permitted to place unobtrusive lights in my own bedroom window.

(Alternatively, imagine that I hypnotize you, by means that are otherwise permissible, such as uttering things within your earshot.) It seems to me that this makes your resulting conative state problematic, at least to some extent, and that I wrong you, at least to some degree. But if I wrong you, why do I wrong you? Enoch can't answer, on pain of circularity, that I wrong you simply by causing you to have a problematic conative state, since he holds that the state is problematic only because I wrong you. Enoch must say that I wrong you in some prior and independent way. But what way is this?

Here's a second reason for thinking that Enoch's account counts too little. Take the most drastic intentional intervention to change someone's desires that one likes, such as surgery. Surely, such desires are problematic in the way that we are trying to describe. But can't we nevertheless imagine circumstances in which it would not treat the person unjustly to force them to undergo the relevant intervention? What if the threat of such intervention is the only means to deter people from wrongfully harming others, with people given the utmost opportunity to avoid incurring the intervention?

Pettigrew's account

In an appealingly exploratory spirit, Pettigrew gives a non-exhaustive list of seven mechanisms of preference formation, focusing in the end on four. "Attention" consists in "having extended contact with the object of preference." "Imitation" consists in "imitating the preferences of those around you."

"Sanctions" involves forming or changing preferences "because the price of having them is too high."

And "Beliefs" involves "gaining beliefs about what is objectively preferable." Pettigrew then suggests:

When your preferences are formed via SANCTIONS, and the preference-inducing sanctions are imposed by unjust external forces... the effectiveness of any consent based on them is diminished. And similarly, if they are formed via BELIEFS, and the beliefs are formed in an evidential environment that is manipulated by unjust forces. On the other hand, if they are formed via ATTENTION, even if your extended contact with the object of preference is imposed on you by unjust external forces, the effectiveness of the consent based on them is not diminished (18).

He goes on to conjecture that what goes for attention also goes for imitation.

I do think that Pettigrew is onto something with attention. Something like attention seems to explain, as Pettigrew intends it to explain, our more welcoming reactions to Enoch's captive pianist and Holt, the parent of the second child. However, I find it less intuitive that imitation should be, as a rule, unproblematic. It is not clear that, say, the self-starving wife's desire would be unproblematic simply because she was "imitating the preferences of those around" her. Moreover, Pettigrew's proposal does not account for the unjust or induced case, since in neither case is the change in desire the result of sanctions or belief.

In any event, one might have hoped for a more unified story. The closest Pettigrew comes to a more unified story as to why attention and imitation are consent supporting, whereas sanctions and beliefs and are consent vitiating, is to suggest that there is an "act of free choice" involved in attention and imitation, but not in sanctions or beliefs. Whereas attention and imitation don't "determine" a certain response, sanctions and beliefs do.

This is clearest in ATTENTION, whereby extended contact with the object gives me some freedom to choose how I value it. That extended contact could lead me to value it greatly, or to value it a little, or to disvalue it, and so on. Placing me in extended contact with the object of my preference doesn't determine what will result from that.

And similarly for preferences formed via IMITATION. This mechanism is essentially a shortcut. Given the cognitive load associated with forming preferences concerning the myriad different objects and people and situations and courses of action we encounter, we outsource some of this to others. But, in standard situations, it's our choice to offload it in this way. And so the preferences formed via IMITATION involve acts of free choice in the same way that preferences formed via ATTENTION do, and in the way that preferences forced via SANCTIONS and BELIEF don't (21).

How are we to understand Pettigrew's suggestion? One possibility is that whereas we literally choose to undergo the processes of attention and imitation, we do not similarly literally choose to undergo the processes of sanctions and belief. That is suggested by "it's our choice to offload it this way." But, again, the point of the example of the second parent is precisely that they don't choose to undergo the process of attention, but instead are forced to undergo it. Nevertheless, the resulting desire is unproblematic.

A second possibility is that having undergone the processes of attention and imitation, we then, at the end of the process, literally choose to form the preferences we form, whereas having undergone the process of sanctions or belief, we don't literally choose to form the preferences that we form. That is

suggested by “extended contact could lead me to value it.” Now, it is not clear that we commonly, if ever, literally choose to form preferences, as opposed to choosing to undergo processes that we expect will lead us to form preferences in some unchosen way. But suppose that we can choose to form preferences. First, suppose that while we can choose to form preferences, we don’t in fact. After undergoing the process of attention, we form preferences without choosing them. It doesn’t seem that makes the preferences problematic. After all, this is the most natural way of understanding the examples of the captive pianist and Holt. Second, it isn’t clear what it would be to undergo the process of imitating others’ preferences and then choosing to form them. Wouldn’t the imitation of others’ preferences already involve forming the preferences? Third, suppose that we understand undergoing the process of sanctions as fearing sanctions attached to certain preferences. Why couldn’t one, having undergone that process, literally choose to form alternative preferences, so as to avoid the sanctions? That would not make the resulting preferences unproblematic. Finally, and similarly, suppose that, having undergone the process of belief, that is, having been led by an adverse evidential environment to believe that certain objects of preferences were objectively desirable, one literally chooses those preferences. That does not seem to make the resulting preferences unproblematic.

A final possibility is that Pettigrew is suggesting that it is not determined that, or that we cannot know in advance that, if we undergo the processes of attention and imitation, we will form a certain preference, whereas it is determined that, or we can know in advance that, if we undergo the processes of sanctions and belief, we will form a certain preference. However, on the one hand, it might well be determined that attention and imitation would produce a preference. Indeed, we imagined this in the case of the captive pianist. This does not, at least in the case of attention, make the preference problematic. On the other hand, it seems that it might not be determined that sanctions and beliefs would produce a desire. Perhaps there is some ineliminable randomness in whether the sanctions or belief produce a desire. This does not make the desires, if they are produced, any less problematic.

Might it be enough to appeal to the standard conditions for valid consent?

So what is going on? For a while I thought that Enoch and Pettigrew were simply mistaken in thinking that the etiology of desire affects the validity of the consent that the desire motivates. The reason why the consent of the self-starving wife of Enoch's example is invalid, in any natural telling of the case, is simply that, whatever she may desire, her *consent* is given under the influence of a *false belief*: namely, a belief that her interests count for less, or that this is something that women in general ought to do, or some such. We don't need any fancy new theory of the causal etiology of *desire* to explain why her consent is invalid. It is invalid for the familiar reason that consent given under the influence of a relevant false belief can be, for that reason, invalid. Even if the loving woman has the same false belief, she is not consenting under the influence of it, but instead out of love. Moreover, this explanation accounts for other cases, in which there is no injustice. Imagine that lightning strikes a chatbot, which then indoctrinates the woman, who is given by a naturally occurring depression to underrate her own interests, in the same way. Again, her consent would seem to be similarly invalid.

Much the same could be said of Victor Tadros's (2021) example of a woman "who tends to see herself as an object for the sexual pleasure of men, rather than valuing her own sexual satisfaction, as a result of the perpetuation of sexist norms and practices in her society" (311). Is her consent to sex therefore invalid? While Tadros leaves the matter open, he appears to hold that the fact that others "unjustly shap[e] her self-conception" is a factor tending to make her consent invalid, even if it does not "fully determine" whether it is invalid when other factors, such as her interest in a sex life, are taken into consideration (312). But the fact that her self-conception is unjustly shaped seems to me a red herring. Again, suppose that she acquires that self-conception because lightning strikes a chatbot. The problem is simply with her relevantly false self-conception, however it is caused. Moreover, it seems that we can give a clear answer about whether her consent is invalid. If her consent isn't under the influence of the false belief—that is, if other factors, such as genuine sexual attraction, provide a sufficient explanation of why she consents—then her consent (barring some other invalidating factor) is valid, whereas otherwise it is not.

And something similar goes for Pettigrew's parent when they had their first child. The reason why their consent was invalid, on any natural telling of the case, is simply that, whatever they desired, their *consent* was given *under the influence of fear of a cost* that some agent wrongly imposed on their withholding consent: namely, ostracism for not having a child. When the parent has their second child, it is not under the influence of the same fear. Again we don't need any new theory of the etiology of the *desires* that motivate consent. We already knew that consent given under the influence of fear of a cost wrongly imposed on withholding consent can be for that reason invalid.

I still think that all of this is right, as far as it goes. But, on further thought, I came to the conclusion that Enoch, Tadros, and Pettigrew simply need better examples. The induced case is such an example. When you consent to my (eagerly) carting off your sculpture to the dump to make way for your seedbeds, there need be no false belief in the induced case anymore than in the organic case. Nor need there be any fear of a cost to withholding consent in the induced case that is absent in the organic case. (Of course, you might fear the dissatisfaction that would attend my not carting off your sculpture in the induced case, but you have the same cause for fear in the organic case.) Nevertheless, the validity of the consent seems questionable.

A positive proposal

So, again, what is going on? My proposal, compactly stated, is the following:

Your conative state, S, is problematic in the way that we have been trying to characterize just to the degree that intentional, reckless, or negligent intervention is causally responsible for S, in a way that is out of your control and does not pass through your nonmanipulated judgment about the object-given or natural state-given reasons for S.

Quite a bit, though, needs to be unpacked.

"Intentional intervention" is intervention that aims for you to be in S. I leave it open whether "intentional intervention" requires anything more than representing your being in S as a goal and organizing activity around that goal, in which case perhaps AI's modifying our preferences to make them easier to satisfy could count as an intentional intervention.

“Intervention” is to be contrasted with simply letting processes underway take their course. Contrast hypnotizing you with simply refraining from hypnotizing you, when it was in my power to hypnotize you, with the intention that your present desire continues unaffected. If intervening is a matter of degree, as this example might suggest, then the extent of the intervention affects the extent to which the resulting desire is problematic.

“Reckless or negligent intervention” is relative to a duty of due care not to bring about S. The relevant, general duty of due care, I suggest, is to see to it that others do not suffer changes in their conative state that they do not sufficiently control and that they have sufficient reason to avoid, when one can do so at sufficiently low cost to oneself or others. For example, if I know that your drinking from a certain stream will change your sexual orientation, and it is sufficiently low cost for me to inform you of this, then I owe it to you to inform you of this. As this example illustrates, others might have “sufficient reason to avoid” a change in their conative state not because the change would be for a worse new object, but instead because it would be disruptive or disloyal to a no worse present object. “Sufficiently low cost” depends, among other things, on the epistemic burdens of becoming informed of whether others are at risk of suffering changes in their conative state that they do not sufficiently control and that they have sufficient reason to avoid. Perhaps, in many cases, the duty applies only when one already knows that others are at such risk, so that no epistemic burdens are imposed. “Sufficiently low cost” also depends on whether one is under an independent duty to act in a way such that others do not suffer changes in their conative state that they do not sufficiently control and that they have sufficient reason to avoid. If one is already under a duty to bear certain costs, C, to act in a certain way, X, then one cannot object, on the grounds of costs C, to a further duty that also requires one to X. Note, finally, that this duty of due care to see to it that others not suffer changes in their conative state does not presuppose that the changes would be problematic in the way that we are trying to characterize, which would make our proposal circular. The duty of due care presupposes only that there is sufficient reason to avoid the changes, which might be because the changes would be disruptive or disloyal.

“Natural” state-given reasons for S are reasons that are not created by intentional, reckless, or negligent intervention that you do not adequately control.

Your judgment is “nonmanipulated” if and only if any intentional, reckless, or negligent intervention that influences your judgment takes the form of presenting you with reasons, getting you to attend to them, or removing obstructions to appreciating their force.

To be sure, “judgment about the object-given or natural state-given reasons” may suggest something more intellectualized than I have in mind. It can be understood, more generally, as responsiveness to the features of the objects of the desires that seem to make those objects desirable or undesirable or to the natural features of the desires that seem to make those desires themselves desirable or undesirable.

So that’s the proposal. To digest it, consider, first, how it accounts for the unjust case. Others in Eunice’s society bring it about that she loses her desire for a career, by negligently closing that option off from her. This violates a duty not to close that option off from her. But it plausibly also violates the general duty of due care, to see to it that she not suffer changes in her conative state that she does not sufficiently control and that she has sufficient reason to avoid, when others in her society can do so at sufficiently low cost. It cannot be argued that it is too costly to avoid closing that option off from her, since there is an independent duty not to do so. So the change in her desires results from negligent intervention. And the change in her desires occurs in a way that does not pass through her judgment about the object-given and natural state-given reasons for the change. The change occurs instead in response to amor fati, induced by her unjust circumstances, which is at most a responsiveness to nonnatural state-given reasons.

This accounts also for the induced case. I intentionally lead you to have a desire for a plant garden and to lose a desire for a sculpture garden in a way that does not pass through your judgment about the reasons for the change: namely, by surgery, drugs, hypnosis, or what have you.

What, then, is problematic about changes in desires that are problematic in the way that we have been trying to characterize? That, I suggest, following Enoch, the resulting conative states are not fully

autonomous or your own. They are instead under the influence of an external agency, subject to another will, or something like it. To be sure, the kind of autonomy at issue is “relational” or “interpersonal”—or at least “interagential.” It is a matter of the resulting conative state not being subject *to another agency*, as opposed to its being at least *with respect to other agencies*, autonomous or one’s own.

We might express the proposal in terms of two sufficient conditions for a desire’s being autonomous or your own in the sense at issue. First, if others don’t affect your conative state through intentional, reckless, or negligent intervention, then your desires are not subject to their will (even if your desires may be subject to their blamelessly inadvertent action). This explains why your conative state in the organic conversion case counts as your own, and Julie’s conative state in the perfectly just circumstances counts as her own.

Second, if you enter into a conative state on the basis of your own judgment, free from manipulation, about the object-given reasons for it, then you make the state your own, through an exercise, if not of your capacity to choose, then of your rational agency, your capacity to respond to reasons. You are simply recognizing the values that the world presents you with. That is a matter between you and those values, not a matter between you and any other agency. You thus avoid what might otherwise be subjection to another agency. Similarly, if you enter into a conative state on the basis of the natural state-given reasons for it. Things are different, of course, if you enter into a conative state on the basis of nonnatural state-given reasons, which have been created by others.

This explains why the conative states of Harriet Tubman and Frederick Douglass, on the one hand, and Enoch’s captive pianist and Pettigrew’s parent of the second child, on the other hand, count as their own. While they may have had no choice about whether they were put in circumstances that would engage their judgment about the relevant object-given reasons for desire, nevertheless their conative states result from the (nonmanipulated) exercise of that judgment. Much education, as opposed to indoctrination, works in this way. It induces desire by exposure to the object-given reasons, even if the educated are forced to attend to those object-given reasons, like our captive pianist. This is what seems to me right in Pettigrew’s suggestion that “attention” can make a desire consent supporting. It also explains

why, if I merely inform you of natural state-given reasons against a desire, such as stem from its being, in a way for which no one is responsible, unavailable or prohibitively costly, and you respond to those reasons by deliberately working on yourself to rid yourself of the desire, the resulting conative state is, in contrast with Eunice's amor fati, still autonomous or yours.

What follows?

What follows from a desire's being nonautonomous or not one's own, in this way? First, it is simply a bad thing for one. Eunice's loss of desire for a career, for example, is something regrettable for her sake. It is bad enough that society has unjustly denied her an option. Still worse that it has distorted what she cares about.

Second, a project, I suggest, is fully autonomous only if and because the desires that partly constitute the project are themselves autonomous in the sense that we are characterizing. Thus, if the desires are nonautonomous, then the project is not fully autonomous. Note while your induced project of plant gardening is not autonomous for this reason, the same cannot be said of Eunice's project of being a homemaker. Her positive desire to be a homemaker, we have been assuming, is autonomous, in something like the way that the captive pianist's desire is autonomous. If her project is nonautonomous, it is for the reason that she has no alternative.

But what follows from that—that a project is not fully autonomous? It's admittedly difficult to characterize. Consider the Narrow Life. Here there is no problem with the origin of your desires. You desire and pursue an otherwise worthwhile project. However, you had no worthwhile alternative. It was your only choice. Something seems lacking in your life, in comparison with the Wide Life in which you had worthwhile alternatives. Surely, we would make your life better by giving you the Wide over the Narrow Life. But, surely, in the Narrow Life, you still had reason to pursue the one project open to you, and it contributed to your well-being to do so. Should we say that your project-based reasons were nevertheless weaker, and that the project contributed less to your well-being, than in the Wide Life? Whatever we should say about your project in the Narrow Life, in contrast to your project in the Wide

Life, we should say something similar about your induced project, in contrast to your organic project. Something seems lacking in your life, although it's difficult to say what follows from that.

Finally, if the conative state is not autonomous, then it is harder than it would otherwise be to make the case that consent that it motivates is valid. It is harder in the induced case than in the organic case to make the case, for example, that your consent to my (eagerly) carting off your sculptures to the dump to make way for your seed beds is valid. (To be sure, there is a familiar, difficult tension here. On the one hand, one wants to acknowledge that your consent is compromised. On the other hand, one doesn't want to deny you the ability to pursue what you now desire. For an excellent discussion of the tension, see Tadros 2021.)

Does it also follow that one is less morally accountable for actions that are motivated by desires that are nonautonomous in this way—that reactive attitudes such as resentment are less appropriate? Not obviously. If you, say, steal fertilizer to satisfy your desire for plant gardening, it seems that you are just as accountable in induced case as in the organic case (although if you steal from me, then I might lack standing to blame you).

Is the account over-inclusive?

Isn't this account too inclusive? Aren't we all the time bombarded with intentional interventions that aim to get us to desire things by processes that don't pass through our judgment about reasons or manipulate that judgment? Yes, and to the extent that they attempts succeed, our resulting desires are less than fully autonomous, at least to some extent, in the way that I am describing.

“But,” it will be said, “consider a ‘nudging’ case in which I know that being without a hobby, you are looking for one, and that you will come to desire the first one you are confronted with. So I confront you first with gardening instead of found-object sculpture, intending to get you to desire it. Surely that is different from using surgery, drugs, hypnosis, or subliminally flickering light to implant in you a desire for plant gardening.”

The first point to make in response is that the account may actually register a categorical distinction between the cases. If the captive pianist's desire counts as fully autonomous even though they

are presented *only* with the option of playing piano, why should your desire count as less than fully autonomous because you are presented *first* with the option of gardening? In both cases, your desire may result from your unmanipulated judgment about the object-given reasons for the activity in question. (A similar distinction might be drawn between administering a love potion and administering a go-on-a-date-with-me perfume. You may have no say about whether you go on a date with me, but how you respond to my charms while on the date, in captivity as it were, may still be up to you.)

The second point to make is that, in any event, the extent to which your desire is nonautonomous, when it is nonautonomous, is a matter of degree. It depends, in particular, on the extent to the which influence on your desire is under your control. One thing that distinguishes surgery, drugs, hypnosis, and subliminal advertisement is that their influence is entirely outside of your control; you cannot avoid or check or otherwise defend against their influence.

The third point to make is that to say that your desire is nonautonomous is not to say that you have been wronged in being brought to have it. Granted, if the desire is brought about by recklessness or negligence, then, by definition, you have been wronged. But if the desire is brought about by intentional intervention, then you need not have been wronged. It depends, first, on how extensive the loss of autonomy is. It depends, second, on how important autonomy in the relevant area is. And it depends, finally, on the interests of others in being free to engage in the intentional intervention. The just deterrent represents an extreme case in which, despite an extensive loss of autonomy in an area in which autonomy is important, the interests of others in the intentional intervention are compelling. But there may be less extreme cases, in which it would burden others who are trying to influence our desires to have to police too closely the boundary between legitimate and illegitimate modes of influence.

Is what is problematic the success of a plan that involved wrongdoing?

We might be tempted by a simpler story. “What’s problematic about the induced desire is simply that it represents the success of a plan that involved wrongdoing. This explains straightaway why we don’t think that the same thing is bad about Harriet Tubman’s desire to resist the wrongdoing. That certainly doesn’t represent the *success* of a plan that involved wrongdoing.”

First, this story would struggle to account for the just deterrent and the clever advertisement, where there is no wrongdoing. And it would struggle to account for the flashing light, where there is no *independent* wrongdoing, although, perhaps, in this case, the general duty of due care could be invoked.

Second, this story would struggle to account for cases in which the change in conative state happens through recklessness or negligence. In that case, it does not count as the success of any plan. No one may have had a plan of seeing to it that Eunice came to desire homemaking.

Finally, this story would imply that the captive pianist's desire is just as lacking in autonomy as the induced desire for the plant garden. After all, the captive pianist's desire was the success of a plan that involved far worse wrongdoing: namely, forced confinement. And yet it does not seem that the captive pianist's desire is problematic in the same way as that induced desire.

Is the account under-inclusive: imposing costs?

“What if I intentionally raise the cost of an option prohibitively, from C- to C+ so as to get you to cease desiring it? I thereby change the object-given reasons for desiring the option. Suppose that, in response to this change in object-given reasons, you cease desiring the option. Intuitively, your desire is no longer autonomous, but the present proposal does not account for this. I have not manipulated your judgment about object-given reasons. I have simply created object-given reasons.”

My reply is that the change in the cost of the option does not affect your object-given reasons for *desiring* the option, although it does change your object-given reasons for *choosing* the option. Your object-given reasons to desire the option-at-cost-C- and to desire the option-at-cost-C+ remain unchanged. If you cease desiring the option, it is a response to nonnatural state-given reasons (such as the felt disappointment or risk of mistakenly incurring C+) which our proposal does cover.

Is the account under-inclusive: What about “nonrelational” autonomy?

Suppose that you discover that the change in your desire resulted from the flickering neon sign. One possibility is that it was my doing. Another possibility is that it was a cosmic accident: just the right atmospheric conditions led the light to flicker. “Your account is under-inclusive,” it might be said,

“because your desire is just as problematic in the accidental case.” This I would deny. To my mind, there is something problematic in the intentional case that is not problematic in the nonintentional case.

“Granted, but your desire is still nonautonomous in a further, nonrelational sense, in the accidental case.” Nonautonomous in contrast to what? Can we draw a line between the accidentally flashing light and the cosmic stew in the original organic case? I am skeptical.

It might seem nevertheless puzzling that our concern with the autonomy of our desires is principally relational in this way. To dispel the sense of puzzlement, it may help to spell out the components of the autonomy of a project.

Let us assume, for the moment, that there is no question of the autonomy of desires; all desires are autonomous. A project is then autonomous, it seems to me, just when, nonrelationally, we have an adequate range of alternatives to it and, relationally, we are not, in pursuing the project, subject to the will of another. The relational component cannot be reduced to the nonrelational component. The intentional removal of alternatives and increases in their cost compromise the autonomy of a project in a way in which the nonintentional removal of alternatives and increases in their cost do not. (Likewise, manipulating someone into choosing to pursue a project, say by deception, compromises the autonomy of that project even if it does not affect the adequacy of the range of alternatives that they have.) Nor can the nonrelational component be reduced to the relational component. The autonomy of a project is compromised by an inadequate range of options, even if nature alone is responsible for the inadequacy.

We have been suggesting, in effect, that there is another relational component of an autonomous project, a component that concerns desires. It is not enough that we are not, in *pursuing* the project, subject to the will of another. It must also be the case that we are not, in *desiring* the project, subject to the will of another. That is, insofar as the project partly consists in a conative state, that conative state must itself be autonomous in the sense that we have been characterizing.

One might ask why there is not a further nonrelational component of an autonomous project concerning desires: that the relevant conative state be, in some nonrelational way, autonomous. But what would that component be? That just as autonomy nonrelationally requires an adequate range of

alternative options for projects, it nonrelationally requires an adequate range of options for desire? But what would that amount to? To have an option is for it to be the case that if one chose the option, one would pursue it. The analogue could not be that, for an adequate range of values for X, if one desired X, X would be realized. That would be insanely demanding requirement for autonomy and would anyway turn desire into a kind of choice. Could the analogue be that, for an adequate range of values for X, if one desired X, one would desire X? In that case, logic guarantees nonrelational autonomy. Thus, in the case of desire, there is no analogue to the nonrelational component of an autonomous project. There is only the relational analogue: whether, in what we desire, we are subject to another will.

Some might reply that is another nonrelational component of an autonomous project, beyond having an adequate range of alternatives to it. We must also actively *choose* to pursue the project. We fail in this if we simply “drift” into pursuing a project. This is a nonrelational matter. A solitary castaway might drift into pursuing a project rather than choose it. Thus, there is a nonrelational analogue in the case of our desires: being choosing our desires, rather than simply drifting into them.

What would it mean to “choose” our desires as opposed to “drifting” into them? That we literally choose our desires? That seems a hopeless suggestion. That our desires somehow come “from us” and not “from brute nature”? Again, that distinction is what I doubted could be made out in our earlier discussion of the accidentally flickering light. In any event, it isn’t clear to me that, setting aside desire, projects are any less autonomous if we drift into pursuing them rather than choosing to pursue them. Such a view seems to me to betray a kind of fetish for planning.

Conclusion

So, returning to where we started, does our proposal give us reason to be concerned that AI might modify our preferences? It depends on whether it would count as intentional, reckless, or negligent intervention. As noted earlier, even if no human intends AI to modify our preferences, perhaps AI might itself count as intentionally intervening, by representing a change in our preferences as a goal and organizing its activity around it. And even if that were not the case, if AI posed some independent danger of changing our desires in ways that we did not control and that we had reason to avoid, we humans might be reckless or

negligent with respect to that, and if our recklessness or negligence led AI to modify our preferences, then our situation would be like Eunice's.

What, finally, about the idea that the correct "goal" is the satisfaction of all and only those preferences formed in the right way—understood now as being autonomous in the sense that we have articulated—with there being no distinction between better and worse objects of preference? First, such a view would be dubiously coherent, since the notion of autonomy is defined in terms of judgment about reasons for desire. It's hard to know how else to distinguish the captive pianist from the induced gardener, given that both are intentionally made to desire what they do, other than by appealing to the fact that, in case of the pianist but not the gardener, the causal process passes through unmanipulated judgment about reasons. Second, I do not think that even autonomous preferences for worthless objects, themselves, provide us with reasons. They do not constitute valuable projects. Finally, I think that even nonautonomous preferences for worthwhile objects, themselves, can provide us with reasons. They can constitute valuable projects, even if, there is, in a sense difficult to pin down, something defective about them.

References

- Bykvist, Krister. 2003. "The Moral Relevance of Past Preferences," in H. Dyke (ed.), *Time and Ethics: Essays at the Intersection* (Kluwer Academic Publishers), 115–136.
- Enoch, David. 2020. "False Consciousness for Liberals, Part I: Consent, Autonomy, and Adaptive Preferences," *Philosophical Review* 129: 159–210.
- Gosepath, Stefan. 1992. *Aufgeklärtes Eigeninteresse: Eine Theorie theoretischer und praktischer Rationalität* (Suhrkamp).
- Hare, Richard. 1989. "Prudence and Past Preferences: Reply to Włodzimierz Rabinowicz," *Theoria* 55 (1989): 152–158.
- Harman, Gilbert. 1986. *Change in View* (MIT Press).
- Parfit, Derek. 1984. *Reasons and Persons* (Oxford University Press).
- Paul, L.A. 2014. *Transformative Experience* (Oxford University Press).
- Pettigrew, Richard. 2020. *Choosing for Changing Selves* (Oxford University Press).
- Pettigrew, Richard. ms. "Consent and the Formation of Preferences"
- Raz, Joseph. 1999. *Engaging Reason* (Oxford University Press).
- Scheffler, Samuel. 2010. "Valuing," in *Equality and Tradition* (Oxford University Press), 15–40.
- Tadros, Victor. 2021. "Consent to Sex in an Unjust World," *Ethics* 131: 293–318.