The Many Ways to Achieve Diachronic Unity

By Kenny Easwaran and Reuben Stern

Abstract:
Bratman (1987) has long argued that future-directed intentions play an essential role in promoting the diachronic unity of agency by allowing for self-governance over time. More recently, Bratman (2012) has argued that future-directed intentions can play this governing role in particular circumstances only if there are diachronic norms of coherence that rule out certain combinations of intentions across time. Here, we argue that the norm that Bratman defends is false, and that Bratman is wrong to think that there must be diachronic norms of coherence in order for agents to exhibit self-governance over time. Our strategy is to identify several ways that agents can achieve diachronic unity without any recourse to future-directed intentions or diachronic norms of coherence.

I. Introduction

When Zeb's owner, Martin, placed his favorite toy in front of him, Zeb did something that surprised Martin. Rather than immediately rushing to put the toy in his mouth, as Martin expected, Zeb spent about 15 seconds concentrating his efforts on *not* putting the toy in his mouth, before appearing to allow himself the opportunity to feverishly pounce on the toy.

Martin was very impressed. As he saw things, Zeb's behavior seemed to exhibit some distinctly human property -- i.e., some property that we typically associate with human agency, but not animal agency. What exactly this property was, or whether it was definitely present, Martin could not say. But he was sure that it had something to do with Zeb's apparent ability to execute a plan over time.

Though Martin may not have been able to articulate why he found Zeb's behavior impressive, Bratman (1987) has argued that Zeb's seeming capacity to self-govern over time is suggestive of a capacity to form binding future-directed intentions. As Bratman sees things, if there are "time-slices" of Zeb that would prefer to pounce on the toy, but that nevertheless abstain, then they must do so because they are constrained or governed by some prior future-directed intention of Zeb's. In order for an agent's actions to be bound in this way, there must be rational pressure to follow through with one's prior plans. But it is no small task to determine exactly when, or in what way, agents are bound by their prior intentions since there are clearly contexts in which it is rationally permissible to abandon some prior future-directed intention -- e.g., when Zeb initially intends to abstain for 15 seconds, but subsequently learns that Martin will take the toy away after 10 seconds.

Bratman (2012) argues that the bindingness of future-directed intentions is rooted in a diachronic norm of coherence that rules out certain combinations of intentions across time. Roughly, Bratman argues that agents must stick with their prior intentions or plans unless they acquire reason to abandon them. So, according to Bratman, unless Zeb acquires reason to

abandon his prior plan to abstain from pouncing for 15 seconds, then he must stick with his prior intention on pain of diachronic incoherence.[1]

In this paper, we first argue that the diachronic norm of coherence that Bratman defends is false, and then identify circumstances in which agents can successfully bind their subsequent time-slices without any recourse to future-directed intentions or diachronic norms of coherence. Thus we argue that there are means through which to achieve diachronic unity or self-governance that do not rely on the cognitive resources that Bratman attributes to planning agents.[2] This does not mean, however, that we think there is no role for binding future-directed intentions in diachronic agency. Rather, we show that in some of the circumstances where agents can achieve diachronic unity through alternative means, it is still true that agents would be more unified and better off (in expectation) were their later actions to be governed by their prior intentions. Thus we argue that future-directed intentions play a valuable, but not essential, role in unifying diachronic agents across time.

II. Bratmanian Self-Governance

According to Bratman (2012, p. 79), planning agents have reason to treat their prior intentions as defaults when doing so follows from adherence to the following diachronic norm, D.

> (D) The following is locally irrational: Intending at t_1 to x at t_2; throughout $t_1 – t_2$ confidently taking one's relevant grounds adequately to support this very intention; and yet at t_2 newly abandoning this intention to x at t_2.[3]

Bratman's defense of D is twofold. First, he argues that D is initially plausible insofar as it yields intuitive verdicts when applied to cases. Second, Bratman argues that self-governance over time is possible only given conformity to D. In this section, we argue that Bratman is wrong on both counts.

As Bratman realizes, a reasonable first reaction to D is that, even if true, D never compels an agent to act in accordance with her previous intention because the circumstances in which D applies are circumstances in which she already takes her grounds to support acting in accordance with her intention.[4] That is, since D bars the agent from abandoning her prior

---

[1] Bratman would depict Zeb as acquiring reason to abandon his prior intention upon learning that Martin will take the toy after ten seconds.

[2] When we speak in terms of "achieving diachronic unity," we do not mean to take a stance on the difficult question of what precisely constitutes diachronic unity. If some readers are unhappy with our choice of vocabulary because they think that diachronic unity cannot be understood behaviorally, they should feel free to substitute every mention of the term with its closest behavioral relative.

[3] Bratman clarifies what he means by "locally irrational" in the following sentence: "D states a *local* demand that concerns a specific sub-cluster of attitudes within an overall cross-temporal psychic economy." [emphasis added] (2012, p. 79)

intention only when she already "takes her relevant grounds to adequately support" her prior intention, and since these are circumstances in which it is already settled what the agent should rationally do, it may seem that the class of cases where D accounts for the *governance* of future-directed intentions is trivially small (or empty).[5]

If an agent's grounds could only ever adequately support one of her options, then D would, in fact, collapse into triviality (since abandoning the original adequately supported intention for some alternative would necessarily consist in adopting a new intention that is, by the agent's lights, less optimal than the original). But since, as Bratman (2012, pp. 79-80) points out, an agent's grounds can adequately support *multiple* incompatible alternatives -- e.g, when an agent's grounds equally support two incompatible options, or when the grounds that support two options are incommensurable -- D does sometimes compel agents to act in accordance with their prior intentions. Consider D's application to the following case.

When Carmen woke up this morning, she settled on having an apple for her mid-afternoon snack. Then, when she arrived at her favorite fruit stand at 3 o'clock, she reconsidered and opted for an orange. Was Carmen rational?

There are clearly further specifications of this case where D does not compel Carmen. For example, if Carmen discovers that oranges are on sale, or that today's selection of apples are rotten, then Carmen plausibly no longer takes the relevant grounds to adequately support having an apple, and D therefore says nothing about whether Carmen has done something irrational. But there are also clearly further specifications of this case where D *does* compel Carmen. If Carmen initially takes the apple and orange to be equivalently good, or if she judges her choice to be, well, between "apples and oranges," and if she acquires no reason to prefer the orange to the apple before 3 o'clock (perhaps because she learns nothing new about the fruits), then D says that Carmen must stick with her original choice of the apple in order to be rational.

So far, so good. It is clear that Carmen should *not* be bound by her previous choice of the apple when she learns that the apples are rotten, and it is *prima facie* plausible that Carmen should stick with her prior intention if she acquires no new information that suggests she should do otherwise. So it may seem that Bratman is right to champion D, and that future-directed intentions govern by establishing defaults that determine what agents should do in the absence of new reasons.

Attractive as D may seem, we argue that it suffers from two main problems. First, though we agree with Bratman that an agent's grounds can adequately support multiple incompatible

---

[4] One might reasonably point out that D never *compels* agent because D is not a command or imperative, but rather an evaluative standard. When we speak in terms of D "compelling" an agent to do something, we just mean that D says that the agent must do the thing in question in order to qualify as rational.

[5] Or as Bratman might put things, if the agent intends to x at t_1, and then at t_2 still takes her grounds to support x-ing, it may seem that there is no work for the prior intention to do in constraining what is rational, since the agent's assessment of the relevant grounds at t_2 are in agreement with what she intended at t_1.

alternatives, we argue that the class of cases where D compels agents to act in accordance with their prior intentions is too small to account for the many cases where our current selves are rationally governed by our prior selves. Second, we argue that D is false -- i.e., that adherence to D sometimes results in irrational behavior in the class of cases where D endows intentions with binding force.

Though Bratman is right that the class of cases where D construes intentions as governing is non-empty, it is relatively small. Suppose, for example, that Carmen acquires the slightest piece of information that speaks in favor of oranges over apples -- e.g., that oranges (but not apples) are within arm's reach upon arriving at the fruit stand. In this case, D ceases to say anything about what Carmen should do because she no longer takes her relevant grounds to support opting for the apple. Alternatively, consider a variant of Carmen's case that more closely resembles the classic tale of Buridan's ass, in which Carmen finds herself positioned exactly between the apples and oranges in the grocery store. Even if Carmen initially takes her grounds to adequately support both kinds of fruit, she plausibly ceases to take her grounds to support opting for the orange once she takes even one step towards the apple (because the apple is then closer than the orange). Thus, here, just as in cases where the agent's grounds initially support only one option, D does not depict the agent's intention as governing since there is no tie to be broken.[6]

Since Carmen's prior choice seems to play a role in determining what is rational even in cases like these -- i.e., when Carmen acquires negligible reason to break from her plans, or when Carmen acquires reason to follow through with her plans -- it is hard to see why Bratman thinks that the truth of D can explain what he takes to be the essential role that future-directed intentions (or plans) play in self-governance over time.

Perhaps Bratman's thought is that in the many circumstances where an agent *has* acquired new information that changes what she takes her relevant grounds to support, her prior intention plays a governing role insofar as it is still true that *were* she not to have acquired this new information, then she *would* have been bound by her initial intention. If this is right, it is still hard to see why Bratman thinks that such a thin modal property can drive self-governance over time. When an agent self-governs over time by settling on some course of action, it seems that her governance consists in *actually* constraining the set of actions that the agent's subsequent time-slices can rationally perform, *not* in constraining the set of actions that the agent's subsequent counterpart-time-slices can rationally perform in distant possible worlds. So even if Bratman were right that there are not counterexamples to D, it is hard to see how D could play the central role in rational self-governance that Bratman attributes to it.

---

[6] In this section, we intend for "tie" to refer not only to cases where options are regarded as equally good, but also where options cannot be ranked. If Chang (2002) is right that there are cases of "parity," in which neither of two options is better than the other, but also not equally good, then these may be cases in which a small improvement to one of the two options does not make it better than the other. If true, the set of cases in which D applies is not trivially small. But it is still too small, we contend, for D to account for many cases in which our current selves are rationally governed by our prior selves.

But alas, even though D does not stick its neck out much, it sticks its neck out too much. That is, even within the range of cases in which D depicts intentions as governing, there are counterexamples.

Suppose that Carmen arrives at the fruit stand five minutes early and has time to kill before her friend meets her there at 3 PM. Since there is little to do while standing at a fruit stand, Carmen reopens the question of which fruit to have in order to see what fruit she currently prefers. (It helps pass the time.) Upon doing so, she takes her grounds to adequately support both the apple and the orange, just as she did when she woke up this morning. But this time, she opts for the orange instead of the apple.

According to D, Carmen's behavior is irrational because she (i) intends at t_1 to have the apple at t_2, (ii) takes her grounds to adequately support having the apple throughout $t_1 – t_2$, and (iii) abandons the intention to have the apple at t_2. But this seems wrong. Surely, there is nothing wrong with introspecting to see which fruit she currently wants, given, first, that it is sometimes rationally permissible to reopen settled practical questions, and, second, that Carmen is better off for having done so in this case (because it helps her pass the time). But once Carmen checks what she currently wants, she does not have any reason to stick with her prior choice. Indeed, it seems that in order for Carmen to derive any benefit from reopening the question, she must regard her choice as unconstrained by her previous choice. This means that it seems rationally permissible for Carmen to plump for the orange at 2:55 for the very same reasons that it seemed rationally permissible to plump for the apple in the morning.[7]

Of course, the circumstances of Carmen's case are somewhat rare. By Carmen's lights, she is better off deliberating once more, but we usually think that there are costs (rather than benefits) to deliberating because doing so takes time and effort. This may *prima facie* suggest that cases like Carmen should be set aside, but this is wrong. By focusing on these cases, it becomes apparent that whether an agent should reopen a question depends just on whether the agent is better off (in expectation) if she does so. Because we happen to live in a world in which there are usually costs (rather than benefits) to reopening questions, our prior choices actually do play an integral role in unifying agents across time through self-governance. But if we lived in a world in which we were always bored, and in which we had cognitive resources to go around, our prior choices would not play this particular organizing role.

This suggests that when agents are bound by their previous choices depends just on considerations of expected utility -- i.e., Carmen should stick to her prior plan when doing so makes her better off (in expectation), but otherwise should not. In section V, we address Carmen's case in some detail and propose a model according to which the (ir)rationality of switching plans depends on the costs and benefits of introspecting. But before developing this

---

[7] One might contend that Carmen is irrational because she *reconsiders* her prior intention, not because of what she elects to do upon reconsidering her prior intention. By our lights, this does not square with Bratman's own treatment of D since it is clear that D does not bar Carmen from reconsidering and sticking with the apple upon doing so. Moreover, it seems rationally permissible for Carmen to reconsider since doing so alleviates her boredom.

model, it is worth demonstrating that agents can achieve diachronic unity (and diachronic self-governance) even when there are not costs to introspecting, and absent diachronic norms of coherence. Sections III and IV describe two kinds of cases leading to diachronic unity absent any costs to introspection, and then section V incorporates these costs.

All of the cases that we describe involve nearly maximally disunified selves composed of time-slices that each have their own interests. We demonstrate three different features of preference structure that can yield some sort of diachronic unity even in these conditions, and show that when multiple features are present, an even greater degree of unity occurs. While none of these models is very realistic, we think that they exemplify effects that are present to some degree in actual people in many situations. There are many types of preference structure that we have not investigated yet, such as cases where time-slices care distinctively about their own behavior rather than that of the group as a whole (as in an intrapersonal Prisoner's Dilemma). But given the way these three effects (and probably others) can support each other in the contexts we do investigate, we think it is a mistake to try to explain all diachronic unity in one way.

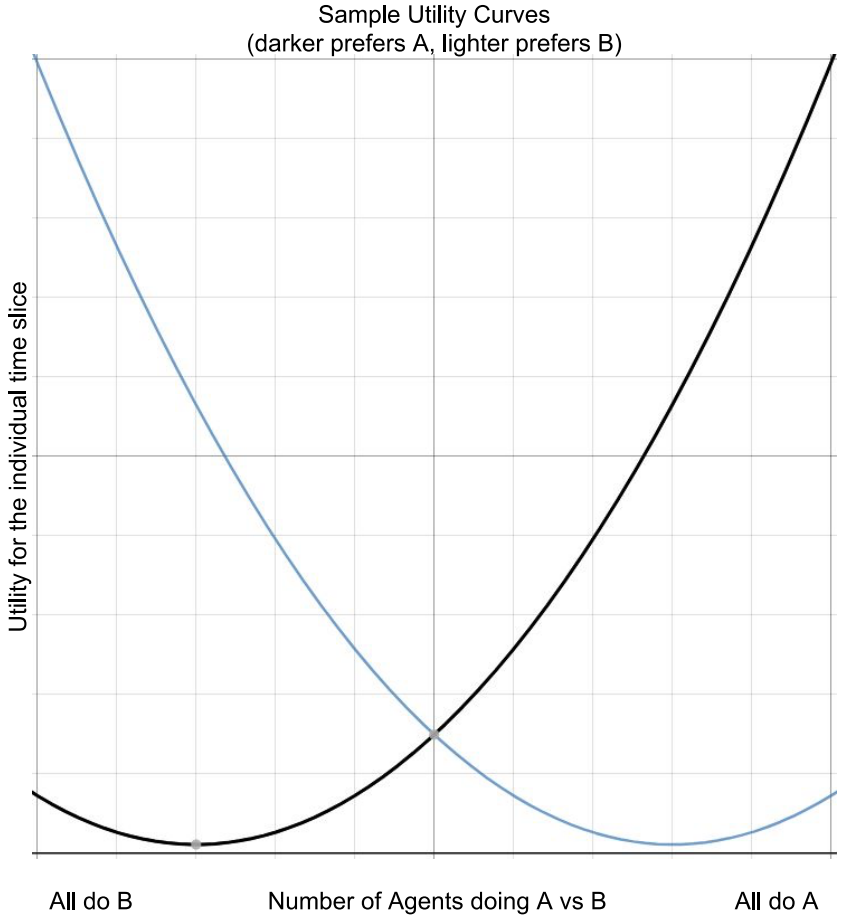
III. Unity in the Absence of Binding Intentions

Aidan is about to have an afternoon off at home. He would like to finally reorganize his closet after some recent shopping trips, which would take all afternoon to complete. But there is also a football game playing all afternoon that he knows some of his friends will want to talk about tomorrow. He is not terribly enthusiastic about football, but he does sometimes enjoy watching the game. He knows that there will be some moments of the afternoon when he is more interested in watching football, and some moments when he is more interested in organizing his closet. Every minute he spends doing one or the other of these activities will be a minute well-spent doing something he cares about to some degree, but he may change his mind about which is more valuable. However, at any moment, he will think it at least as good to have spent all afternoon doing just one of the two (rather than splitting his time) -- i.e., even at moments when he would prefer an afternoon watching football, he will think that a fully organized closet is at least as good as a half organized closet and half a football game; and even at moments when he would prefer an entire afternoon of cleaning, he will think that watching the whole game is at least as good as watching half and organizing half of the closet. But he is likely to change his mind at various points about whether watching the full game (or nearly the full game) is better than organizing the full closet (or nearly the full closet).

What should Aidan do? Is he diachronically unified? Would he benefit from forming a binding future-directed intention? We think that the following is a useful model for considering Aidan's case.

Treat the succession of time-slices of Aidan as individual agents, each confronted with the decision to spend the next minute engaged in action-type A (organizing) or action-type B (watching football).[8] Because each time-slice's preferences depend on what the other time-

slices do, each “agent” has a utility function that is sensitive to the total number of agents that choose A and the total number that choose B. Also, because each time-slice is equally ignorant about what the others prefer (except insofar as the preferences of earlier time-slices are revealed by their behavior), the individual “agents” should be modeled as symmetrically uncertain about what the other time-slices prefer.



When Aidan’s time-slices are modeled in this way, we prove the following results. (See Appendix.) For each concave-up utility function (like those specified above),[9] there is an *i* such

[8] Because we use committees of individual agents (or time-slices) to model diachronic agency, it is worth considering whether our model can be extended to the case of *collective* agency, where the individual agents are individual humans, rather than individual time-slices. (See Weirich 2009 for an interesting discussion of the relationship between diachronic agency and collective agency.) We conjecture that our discussion of when and to what extent time-slices qualify as unified diachronic agents is sufficiently general to apply to the questions of when and to what extent collections of individuals qualify as unified collective agents, but space limitations require us to leave this for later.

[9] A function is said to be “concave-up” if for any two inputs, the average of the corresponding outputs is greater than the output for the average of those two inputs. For a utility function with this feature, there is

that if the first *i* time-slices have all already done the same act-type, an agent with that utility function has higher expected utility for continuing the streak rather than switching, regardless of which act-type he prefers in general. However, if the actions of the time-slices up to the present are either equally split (including for the first time-slice, when there are no actions up to the present) or only one more has done the act the agent disprefers, then he will have higher expected value for doing the one that he prefers generally, even if that involves switching. Finally, prior to knowledge of the action of any time-slice, every utility function in this family has higher expected utility for allowing one time-slice to choose the act of all of them dictatorially rather than allowing each to choose its own act in light of the acts that have come before.[10]

What do these results suggest about Aidan's diachronic unity?

Even in the absence of binding intentions, it appears that Aidan's time-slices can be bound by his prior time-slices, given the particular shape of their utility functions. For example, if Aidan spends enough time organizing the closet (or watching the game), then the rest of his time-slices will be in agreement with his former time-slices that it is rational to continue organizing the closet (or watching the game) despite the fact that they may well have a preference for the alternative activity, simply because they are choosing in the environment of their former time-slices. Thus Aidan's earlier actions can force his later time-slices to be in agreement about the rational course of action even in the absence of binding future-directed intentions. Put differently, once Aidan has established himself sufficiently on one action, he will achieve a simulacrum of self-governance just from his momentary preferences.

Does this mean that binding intentions would have no use to Aidan? No. In the beginning of the afternoon, if Aidan oscillates between preferring football and organizing (as he might, depending on the distribution of the two utility functions across the early time-slices), then he will switch back and forth a few times, and, if he's unlucky enough, he might spend all afternoon switching. The Preference for Unanimity result (described in the Appendix) implies that every time-slice of Aidan would expect to be happier were Aidan able to form a binding intention at the outset to spend all afternoon doing a single one of the activities. So it seems that it would be beneficial to Aidan to be able to form a binding intention, even though it is possible for Aidan to be unified in the absence of such intentions.

Of course, our momentary utilities are not always (or even usually) concave-up like Aidan's, so we can ask whether there are cases where agents can be unified in other ways.[11] The next

---

some value to being more extreme rather than middling. We describe the functions in detail in the Appendix.

[10] One might assimilate these binding intentions to McClennen's (1990) conception of resolute choice.

[11] For example, we can use concave-down utilities to model an agent who achieves diachronic self-governance by creating an environment in which her later time-slices will rationally prefer to adhere to what Bratman has called a "sampling plan." In these cases, every time-slice prefers heterogeneity across time (rather than homogeneity) and the later time-slices will sometimes opt to do whatever the earlier time-slices haven't done, even when doing so involves settling on doing something that they otherwise would not prefer.

example teaches us that there are, in fact, other ways that diachronic unity can be achieved solely through one's momentary preferences.

IV. The Cost of Switching

Brooklyn is playing at the playground. On one end of the playground is a sandbox. On the other end of the playground is a swingset. It takes her a minute to cross the playground to get from one to the other. She has 60 minutes to play at the playground. At some points in time, she values time spent at the swingset more than time spent at the sandbox. At other moments in time, it is the reverse. However, at every moment in time, she prefers time spent at the swingset or the sandbox to time spent crossing the playground. Unlike Aidan, Brooklyn doesn't care about completeness of either activity -- no matter what moment's preferences we consider, each minute spent doing one activity adds just as much utility as each other minute spent doing that activity.

Again, we consider a formal model where each time-slice is considered a separate agent, each has her own utility function, and each is symmetrically uncertain about the utilities of each other time-slice. Again, we show similar results. For each utility function of this form, there is some *i* such that if there are fewer than *i* time-slices left in the sequence, she will keep doing whatever action her predecessor did rather than switching. Earlier in the process, the individual preferences will matter, but only if they are sufficiently strong compared to the cost of switching, and the relative preferences of the other time-slices. (See Appendix.) Finally, prior to knowledge of the action of any time-slice, every utility function in this family has higher expected utility for allowing one time-slice to choose the act of all of them dictatorially rather than allowing each to choose its own act in light of the acts that have come before.

The conditions under which one of Brooklyn's time-slices will act as if bound by her predecessor are different from the conditions under which one of Aidan's time-slices will act as if bound by his predecessor. But both of them will sometimes act as if bound, and every time-slice of each of them would antecedently prefer that the first time-slice had the capability to form a binding intention. Furthermore, increasing the cost to switching in Brooklyn's case will result in more time-slices acting as if bound, and we conjecture that the same is true for increasing the concavity of Aidan's preferences.

Brooklyn, like Aidan, can be unified to some extent in the absence of binding intentions since the rational choice of a given time-slice will sometimes be forced to agree with prior time-slices.[12] That is, even in the absence of binding intentions, Brooklyn's choice to play in the

---

[12] If we consider a case that has features of both Brooklyn's and Aidan's, where the utility function for each time-slice is concave-up in the total number of time-slices that do a given act, and there is *also* a cost based on the number of switches, then we can sometimes get even greater degrees of unity. For instance, we noted that in a case where the utility function was concave-up and the utility of the 50/50 split was at least as great as the utility of the opposite extreme, an agent will have perfect unity if by chance, some initial sequence of time-slices all have the same preference, and the agent will have total disunity if by chance, consecutive time-slices all have opposite preferences. If a cost to switching actions is introduced, then towards the end of the sequence, the cost of switching will outweigh the possible

sandbox at *t* can have the effect of making it rational to play in the sandbox at *t*+1 even if Brooklyn would prefer to swing on the swingset at *t*+1 were she to start afresh. This may make it seem as though binding intentions have no use to Brooklyn (since her prior time-slices succeed at persuading her subsequent time-slices sans any costs of reconsideration), but this is not right. The Preference for Unanimity result shows that all of Brooklyn's time-slices would, prior to any action, prefer to have their behavior governed by a dictator with a master-plan about when Brooklyn should do what. So it seems that Brooklyn's case, like Aidan's, helps us to understand why there are particular circumstances in which it is good to be diachronically unified by binding intentions.

V. The Costs of Introspecting

Aidan and Brooklyn teach us that it is possible for agents' earlier time-slice to govern their later time-slices even in the absence of binding intentions. By initially settling on some particular course of action, Aidan and Brooklyn's early time-slices effectively create an environment in which it is optimal for Aidan and Brooklyn's later time-slices to settle on the same course of action, even though Aidan and Brooklyn's later time-slices would otherwise prefer to do something else.

Though these cases demonstrate that there are multiple ways to achieve diachronic unity, they do not show that an agent's earlier time-slices can govern her later time-slices without changing what the later time-slices want to do.[13] This means that these cases do not shed light on the kind of rational self-governance that exists when Carmen is rationally required to stick with her plan to have the apple because she takes there to be costs to determining what she wants, but remains neutral between the apple and the orange.

In order to model cases like Carmen's, we must introduce the costs and benefits of introspecting. Upon doing so, it is clear that the rationality of *x*-ing *upon introspecting* must be distinguished from the rationality of *x*-ing *absent introspection* because the rationality of these actions can come apart. For example, when Carmen believes that there are perks to thinking about what fruit she wants before her friend arrives, it seems that Carmen prefers introspecting and having an apple (or an orange) to not introspecting and having an apple, and that if Carmen is rational, she will either introspect and have the apple, or introspect and have the orange.

---

benefit of having more of one's preferred action, so even the maximally disunified sequence will exhibit some unity in its behavior. However, for the sequence that is already maximally unified, this cost doesn't add any more unity. A detailed analysis of when the two factors together produce more unity than either does separately would require further investigation of specific utility functions. But even with what we have now, we can see that maximal unity is possible either with a high cost to switching, or with sufficiently concave utilities and a sufficient streak of agreement in the early stages, or with some combination.

[13] One may worry that we have not described cases of genuine self-*governance* because each time-slice is just does what it most wants to do. By our lights, this reply requires too narrow of a conception of governance. For example, an employer clearly *governs* her employees inasmuch as she gives them incentive to prefer taking some actions to others. This is basically what Aidan and Brooklyn do, since they promote an environment in which their later selves will be motivated to serve their earlier interests.

There may be multiple ways to account for the costs and benefits of introspecting. Here, we propose modeling the cost of introspecting as a price that the agent pays in order to infer what she prefers in the moment, or put differently, in order to remove any ignorance about her utility function.[14] Thus we model the agent who does not go to the trouble to reopen a settled question as ignorant about what course of action she regards as optimal in the moment, but as possibly paying a price to infer whether she finds sticking with her prior plan optimal. (See Appendix.)

This cost of introspecting provably can render the agent bound by her prior plans. For example, if Carmen treats her prior plan to have the apple as evidence that she currently wants the apple, then it is rational for her to stick with her prior plan if the costs of introspecting are sufficiently high.[15] Alternatively, if Carmen thinks (absent introspection) that she probably prefers the apple to the orange because there is usually a cost to deviating from standing plans, then even if Carmen, as a matter of fact, is neutral between the two options, it can be rational to stick with the prior plan in order to avoid the costs of introspecting. On the other hand, if Carmen were to think that there was no cost (or even a benefit) to discovering her preferences, then Carmen would have no reason (or even negative reason) to abstain from determining what she wants, and would therefore be rationally required to opt for some course of action that she momentarily (actually) regards as among the optimal actions. Thus when Carmen is actually neutral between the apple and the orange, and thinks she stands to benefit from introspecting (because it helps to alleviate the boredom), she can rationally opt for either the apple or the orange.

As we discuss in the Appendix, if we revise the cases of Aidan and Brooklyn to introduce costs of introspection, then Aidan and Brooklyn will be even more diachronically unified. There is also a formal similarity between the way that a switching cost and a cost to introspection induce unity, though the cost to introspection turns out to be more effective.

VI. The Costs of Reassessing

At this juncture, we have demonstrated that there are multiple ways to achieve diachronic unity in particular circumstances without any recourse to diachronic norms of coherence. But we have not modeled *every* way that agents can achieve diachronic unity. For example, the following case evades our current grasp.

At the beginning of an offensive basketball possession, Diana settles on driving left towards the basket. As she gets closer to the basket, it may be best (in expectation) for Diana to stick to her guns and not reconsider her prior intention (perhaps because slowing down to consider what seems best often leads to turnovers), but, in the event that she does reconsider, better for Diana to go right than left (perhaps because she now judges the path to the right to be more open).

---

[14] In the formal model, we model Carmen as having one known utility function, but as being uncertain about her own psychology.

[15] This is true even though Carmen would discover that she is neutral between the apple and the orange, were she to introspect.

Here, it seems that Diana should not reconsider her prior choice, but not because of any price associated with introspecting. Indeed, Diana may know perfectly well when driving left is optimal and when driving right is optimal; it just takes Diana time and effort (that is better spent driving towards the basket) to determine whether her current circumstances call for driving left or driving right. It is natural to model Diana's earlier time-slice as an expert that Diana's later time-slice can defer to when she does not take the time to investigate the circumstances herself. Were Diana to take the time to survey her circumstances once more, she'd discover that it's better to change courses. But because doing so takes time and effort, it can still be rational for Diana to defer to the prior time-slice's assessment by continuing left.[16]

Though the costs of introspecting can be modeled as a price that one can pay to eliminate uncertainty about what she desires, this exact treatment is not available here because Diana is not uncertain about what she values. However, if we slightly modify the existing model so that Diana can pay a price to reassess the state of play (rather than deferring to her past time-slice), then we can capably represent Diana's choice.[17] The unity here may not be of the same kind discussed in previous cases because it resists changing external circumstances, not internal circumstances. But it does seem that both binding intentions and models like ours can capture the kind of self-governance at play here. Indeed, this model may be interpreted as a way to represent Bratmanian binding intentions.

VII. Lessons Learned

Aidan, Brooklyn, and Carmen teach us that there are ways for agents to be diachronically unified (and self-governed) even in the absence of binding intentions, but also reveal that there are contexts in which diachronic agents can be better off (in expectation) for having binding intentions.

On the one hand, this means that there really is something remarkable about having the capacity to form binding intentions (or to execute plans over time) since collections of time-slices that have this capacity can outperform those that do not. On the other hand, this suggests that there are multiple ways to achieve diachronic unity even in the absence of binding intentions, and that we cannot be sure that an agent has exhibited self-governance (by forming a binding intention) just from observing her diachronic unity since her unity may have resulted from her momentary preferences being aligned in the right way.[18]

---

[16] One might argue that it is wrong to include the option of not reassessing the circumstances because Diana does not actively decide whether to reassess as she drives towards the basket. Though we agree that agents usually are not confronted with conscious choices about whether to reassess (or to introspect, for that matter) we believe that we can nevertheless use an agent's probabilities and utilities to *evaluate* whether she was rational to reassess or introspect.

[17] We discuss this in greater detail in the Appendix.

[18] This suggests that Martin might have been right to be impressed with Zeb's seeming capacity to form binding intentions, but also might have been right not to be certain that Zeb really had this capacity.

Either way, whether an agent should follow through with her prior plans seems to depend on expected utility considerations in a great many circumstances (perhaps excepting cases like Diana's). So whether an agent should follow through with her plans depends on how costly (or beneficial) it is to do otherwise, and on how much she stands to gain from taking the steps necessary to steer herself in a new direction.

References

Bratman, M. (1987). *Intentions, Plans, and Practical Reasons*. Cambridge, MA: Harvard University Press.

Bratman, M. (2012). Time, Rationality, and Self-Governance. *Philosophical Issues*, *22*, 73-88.

Chang, R. (2002). The Possibility of Parity. *Ethics, 112*, 659-688.

McClennen, E. (1990). *Rationality and Dynamic Choice*. Cambridge, UK: Cambridge University Press.

Weirich, P. (2009). *Collective Rationality*. Oxford, UK: Oxford University Press.

Appendix

A. Aidan: Concave-up utilities

Aidan has an afternoon to spend either organizing his closet or watching a football game. He might change his mind about which activity he prefers over the course of the afternoon, but our formal model shows that he may nevertheless make a decision and stick with it. We consider him as a sequence of time-slices, and treat each time-slice as an independent agent with preferences regarding the collective behavior of all time-slices, each given by its own utility function. Importantly, although the time-slices may disagree about which activity is better, they all agree that completion is preferred over splitting time.

Formal assumptions:

There is a fixed finite number n of agents who are to make decisions, and each agent is certain about how many agents there are.

Every agent has two choices, A and B. Each agent has a utility function that depends only on the number a of agents that choose A and the number $b=n-a$ of agents that choose B. For each agent, the utility function is concave up -- for any i and j, the utility when $a=(i+j)/2$ is less than the average of the utility when $a=i$ and the utility when $a=j$. For each agent, the utility of an even split between A and B is at least as high as the utility of one type of unanimity, and is less than the other type of unanimity. One such utility function is given by a^2+3b^2, but many others are possible.

Every agent is certain of the above facts for the utility function of each other agent, but none of them knows each other's utility functions. At the time of action, each agent is certain of the behavior of each earlier agent, but has no information about the behavior of any later one. Prior to observing any behavior, each agent has independent, identically distributed credences about the utility function of each other agent, and these credences are symmetrical with respect to interchanges of A and B.

This could be because there are just two utility functions that are mirror images of each other and the agent has an independent .5 credence that each other agent is of one of these two types, or because there are many utility functions with greater or lesser degrees of favoring of one unanimous type over the other, or greater or lesser degrees of concavity.

Results:

As a result of concavity, and because the 50/50 split is at least as good as unanimity on the disfavored action, each agent also prefers a 50/50 split over any non-unanimous majority for the disfavored action. Furthermore, by concavity, any non-extreme symmetric probability distribution over the behavior of the group has a higher expected utility than a guaranteed 50/50 split, and a lower expected utility than a .5 credence in each extreme.

Lemma 1: If, conditional on a particular initial sequence S of choices, the expected utility of doing act A is at least as high as that of doing act B, then conditional on an initial sequence SA, the expected utility of doing act A is still at least as high as that of doing act B. That is, regardless of whether the agent prefers A to B generally, if a given initial sequence makes act A preferable, then an additional A by one extra time-slice beforehand won't change this fact.

Proof:

We will prove the contrapositive - if for a given utility function, conditional on initial sequence SA, the expected value of doing B is greater than that of doing A, then conditional on initial sequence S, the expected value of doing B is greater than that of doing A.

As part of the proof, we will use this corollary: for any probability distribution over utility functions, the probability of an agent preferring B conditional on initial sequence S is at least as high as the probability of the agent preferring B conditional on initial sequence SA.

Proof, by induction on the number of remaining agents after the current one:

Base case: The length of S is n-2, so that the act after SA is the last one. Assume that the utility function is such that the sequence SAB has at least as high utility as SAA. Each utility function under consideration is a function of b, the total number of B's in the sequence. Since the utility function is concave up, there is a number i such that the utility of a total sequence is monotonically increasing in b when $b>i$ and monotonically decreasing in b when $b<i$. Since the utility of SAB is greater than that of SAA, the number of B's in S must be at least i. Thus, $U(SAA)\leqslant U(SAB)=U(SBA)<U(SBB)$. The expected utility of SA is a weighted average of U(SAA) and U(SAB) and the expected utility of SB is a weighted average of U(SBA) and U(SBB), where the weights are the probabilities of the last agent having a utility function that leads it to do A or B. Since U(SBA) and U(SBB) are at least as high as U(SAA) and U(SAB), and U(SBB) is strictly greater than U(SAA), this means that, given initial sequence S, the expected value of doing B is strictly greater than that of doing A.

For the induction step, assume that we have shown for all utility functions in the relevant family, if EU(S'AB)>EU(S'AA) when there are j unknown agents after the ones mentioned here, then EU(S'B)>EU(S'A). Now we must show that if EU(SAB)>EU(SAA) when there are j+1 unknown agents after the ones mentioned, then EU(SB)>EU(SA).

Note that EU(SB) is a weighted average pEU(SBA)+(1-p)EU(SBB), and EU(SA) is a weighted average qEU(SAA)+(1-q)EU(SAB), where p and q are the probabilities of an unknown agent preferring act B conditional on initial sequences SB and SA respectively. Note that EU(SBA)=EU(SAB), and by the assumption of the theorem we have EU(SAB)>EU(SAA). Thus, if EU(SBB)≥EU(SBA) then we are done.

Lemma 2: An agent's expected utility given an initial sequence S of choices, and given that all later agents have her same utility function, is at least as great as her expected utility given an initial sequence of choices, with no information about the utility functions of later agents.

Proof: The expected utility of a given sequence of choices is a mixture of the expected utility of the next agent doing A and that of the next agent doing B. Conditioning on the agent at that time having the same utility function as the agent under consideration removes all weight from whichever of these two expected utilities is lower for the agent under consideration. Thus, conditioning on every agent having the same utility function as the agent under consideration just increases the expected utility at each step.

Proposition: Conditional on an initial sequence of choices containing i choices of the agent's favored act, and i+1 choices of the agent's disfavored act, the agent weakly prefers to do her favored act, with equality only if i=0 and she is the last agent to act.

Proof: By lemma 1, if a single-act majority for the disfavored act were sufficient to prefer doing the disfavored act, then so would any greater majority. Thus, if all future agents were of the same type as this agent, the actual outcome would be unanimity if i=0, and a split majority for the disfavored act if i>0. Thus, by lemma 2, the actual expected utility of doing the disfavored act with no information about the type of later agents must be no greater than this, which is at most equal to the utility of a 50/50 split if i=0, and strictly less if i>0.

However, if she performs her preferred act, then the information available for the next agent is symmetrical. Thus, with no information about future agent types, her credences over complete outcomes must be symmetrical, and non-extreme if she is not the last agent to act. Thus, the expected utility conditional on her performing her favored act is strictly greater than that of a 50/50 split if she is not the last agent to act, and the expected utility of her performing her disfavored act is strictly less than that of a 50/50 split if n>0. Thus, if either n>0 or she is not the last agent to act, she prefers in this circumstance to do her preferred act.

If n=0 and she is the last agent to act, then she might be indifferent between the two acts, because one produces total unanimity and the other produces a 50/50 split.

Combining this proposition with Lemma 1, we see that conditional on a prior 50/50 split, an agent prefers to perform her favored act.

Thus, we can see that if there are more than two agents in the sequence, and the first two favor different actions (no matter how much or how little they favor the different extremes), the group will not perform a unanimous action.

Preference for Unanimity result:

Thus, if there are more than two agents in the sequence, then the prior credence for each agent will be a non-extreme symmetric distribution over action sequences, which has strictly lower expected utility than a .5 credence in each unanimous action sequence. Since this .5 credence in each unanimous action sequence is the expected result of giving the first agent in the sequence the ability to bind all future agents to act the same way as the first, each agent will in the abstract prefer the ability for the first agent to make such binding decisions, even though she would prefer to act differently from the first if she happened to go second after an agent of the opposite type.

B. Brooklyn: Linear utility with a switching cost

Brooklyn has an hour to spend at the playground, splitting her time between playing in the sandbox, playing at the swingset, and running between the two. She might change her mind about which activity she prefers at any moment, but our model shows that she will sometimes stick with the dispreferred activity, and may even stick with it for the entire hour, despite changing her mind. Again, the formal model considers each time-slice as her own agent with her own utility function, and uncertainty about the utility functions of other time-slices.

Assumptions:

There is a fixed finite number n of agents who are to make decisions, and each agent is certain about the value of n.

There are two types of actions - A and B. For each agent, the utility function is determined by three positive constants, x, y, and s, and takes the form $ax+by-ds$, where a is the number of agents that chose A, b is the number of agents that chose B, and d is the number of agents that chose a different act from her predecessor. For each agent, the switching cost s is greater than the difference of x and y. (The assumption that the total utility of a time-slice is given by her own values for the experience of each other time-slice's behavior is unrealistic if these utilities are taken to be hedonic utilities - in that case it would be more plausible for the total utility of a time-slice to depend on how many time-slices do their *own* preferred behaviors rather than hers. But this assumption may be more plausible if the value of the play involves spinning out an elaborate fiction based on each activity, that depends on the behavior of other time-slices involved in the same activity.)

Each agent is certain of these facts for all other agents, but none of them knows the specific constants of any other agent. Each has i.i.d. credences about the utility function of each other agent, and these credences are symmetrical with respect to interchanges of A and B.

Some agents will have preferences and positions in the sequence such that they will do the act that has higher utility for them regardless of what their predecessor did, while others will have preferences and positions in the sequence such that they will do what their predecessor did, regardless of which act has higher utility for them. If there is a consecutive sequence of agents

of the latter type, we say that they are all “controlled” by the last agent to choose on the basis of her own preferences before that sequence.

Consider an agent about to decide whether to be controlled by her predecessor, or to choose on the basis of her own preference. Let c_i be the number of time-slices (counting herself) that are expected to be controlled by the choice of the ith to last agent if she chooses for herself. $c_i|b\text{-}a|$ is then the expected advantage of choosing based on her own preference rather than choosing the opposite of her own preference, while s is the expected advantage of choosing the act chosen by her predecessor rather than the opposite. Thus, the ith to last agent will be controlled by her predecessor iff $c_i|b\text{-}a|<s$.

Let $p(c)$ be the probability that a randomly selected agent has $c|b\text{-}a|<s$. Then $c_{i+1}=1+c_ip(c_i)$, because the agent in the position i+1 from the end would expect to control herself, and have probability $p(c_i)$ of controlling the c_i that would be controlled by her successor. Since $p(c)$ is the probability that a randomly selected agent has utility function with $s/|b\text{-}a|>c$, we see that $p(0)=1$ (since s is always positive), and p monotonically decreases to 0 as c goes to infinity.

Let $t=s/|b\text{-}a|$. This is the threshold for c at which a given agent would be willing to choose according to her own preference rather than according to the act the previous agent chose. If everyone has the same value of t, then the tth to last agent is the last one that is willing to choose according to her preference, and all later agents will be controlled by her. Since the tth to last agent controls herself, the previous agent will act as if she is the last, and by parallel reasoning, every block of t agents, counting from the end, will be controlled by its first member. Thus, in this situation, there is a very real sense in which we could treat the sequences of t time-slices as the units, rather than the time-slices themselves.

If agents don’t have the same value of t but instead it is distributed continuously, then $p(c)$ is a continuous function, so $1/(1\text{-}p(C))$ is a continuous, monotonically decreasing function going to infinity as C goes to 0 and going to 0 as C goes to infinity. Thus, there is one unique value for which $C=1/(1\text{-}p(C))$, or equivalently $1+Cp(C)=C$. As i gets large, c_i will converge to this value. Thus, as long as it is far from the end of the sequence, everyone whose t is in the most extreme $1/C$ of the population will choose according to her own preferences, while everyone whose t is less extreme will be controlled by her predecessor. However, in the last few segments of the sequence, even agents whose preferences are this extreme will be controlled by their predecessor if there are not enough successors left to make up for her t.

Note that as s gets large while the distribution of $|b\text{-}a|$ stays fixed, C gets larger, and a larger fraction of agents will be controlled by their predecessor. Thus, higher switching costs translate to greater behavioral unity.

Preference for Unanimity result:

The expected value of being bound by the first agent is equal to the average of the two unanimous sequences. The expected value of not being bound is the average of some

symmetric distribution over the sequences. The average of the value of any sequence and its dual is equal to the average of the two unanimous sequences minus *s* times the number of switches that each sequence has. Thus, again every agent will prefer that the collective be bound by the first agent, even though some agents might prefer to switch.

Note that if there is an upper bound to the possible value of |*b*-*a*|, then if *s* gets high enough, there will be no switches, and this can mimic a binding intention. So increases in the cost of switching with no compensating increase in the payoff for any act can paradoxically increase the expected utility of time-slices. However, this increase isn't always monotonic.

C. Carmen: Costs to introspection

Carmen is choosing whether to take an apple or an orange, and is uncertain which she prefers. She can spend some effort introspecting to get certainty about her preferences, or just act according to her uncertain preferences.

For the formal version of this, we will start by considering a one-shot version of the decision, rather than a sequence of time-slices as in the examples of Aidan and Brooklyn. We will provide some discussion of how this effect of a cost to introspection changes things if added to sequential choice cases, but we won't provide a full analysis. This is because the analysis is fairly uninteresting if every time-slice has the same symmetric uncertainty, but if they don't, then the analysis depends on uncertainties not just about each other's utility functions but about each other uncertainties, and the complexity quickly gets out of hand.

For a formal analysis, we can treat Carmen as having a choice between two actions, A and B, and uncertainty about her own utility for each action. Under function U_1, A has greater utility than B, but under function U_2, B has greater utility than A, and Carmen is uncertain which utility function is actually hers. We can formally model this as though Carmen has one *known* utility function, and an *unknown* bit of information about her own psychological state, so that $U_1(A)$ is treated as $U(x_1,A)$, $U_2(A)$ is treated as $U(x_2,A)$, and so on. If p is the probability that she has function U_1, then we can see that the choice to do act A has expected utility $pU(x_1,A)+(1-p)U(x_2,A)$ and the choice to do act B has expected utility $pU(x_1,B)+(1-p)U(x_2,B)$.

If Carmen has the option to introspect, for introspection cost i, then she can use this introspection to do whichever act in fact has higher utility for her. The expected value of this plan is $pU(x_1,A)+(1-p)U(x_2,B)-i$. She will rationally choose to pay this cost iff $i<p(U(x_1,A)-U(x_1,B))$ and $i<(1-p)(U(x_2,B)-U(x_2,A))$. If i is higher than either threshold though, she will just choose to do whichever act appears to have higher expected utility in light of her uncertainty. If U_1 and U_2 are symmetrical with respect to A and B, then this means that she will rationally pay the cost of introspection iff the probability of the less likely utility function multiplied by the difference in utility of the two outcomes is greater than the introspection cost, and otherwise she will rationally just do whichever act has the higher probability of being better. (When Carmen is bored, the introspection "cost" is negative, because she derives entertainment from introspection, so she will definitely pay the cost.) Thus, if an earlier plan is taken as evidence that one's utility function

is more likely to still favor that act than the other, Carmen will rationally stick with this plan unless the possible difference in values between the two outcomes is especially large.

If we compare this to a one-shot version of Brooklyn's case, where she can pay a switching cost *s* to do B, or do A for free, we see that Brooklyn will switch iff $s<(U(x_2,B)-U(x_2,A))$, while Carmen will switch iff $i<(1-p)(U(x_2,B)-U(x_2,A))$. The difference between the two is a multiplicative factor of the probability of the less likely preference, so a relatively low introspection cost may well be more effective than a somewhat higher switching cost at producing behavior like what we saw in the analysis of Brooklyn's case above. (A fuller analysis of an iterated version of this case would require some theory of how the probabilities of the two utility functions evolve over time, and an analysis of how many future time-slices the combination of switching cost and introspection cost allows one time-slice to control.)

Note that a formally similar analysis is available for Diana's case, if we reinterpret x_1 and x_2 not as internal psychological states that determine the value of the two acts, but as external states of the game that determine the value of driving left or right, and interpret *i* not as the cost of introspecting which psychological state one is in, but instead as the cost of analyzing the state of play to determine which direction would be better to drive in. If an earlier decision to drive left is evidence that the game is probably still in a state where driving left is better, then Diana will continue to drive left unless the uncertainty about which direction is better is high enough and the value of the better direction great enough to overcome the cost of analyzing the game.

Combining this notion of an introspection cost with Aidan's feature of concave-up utilities is even more complex, but it can be worked out in a simple case, to see that even a relatively small introspection cost can change things. Consider a case with just three time slices, each having a utility function of either a^2+3b^2 or $3a^2+b^2$. If the first two time slices have done the same act and the third is debating which to do, then his choice is either between 9 or 7 units of utility, or between 27 or 13, depending on which utility function he has. In any case, he would prefer doing the same as the previous two. If the first two time slices have done different acts, then the third is choosing between 13 and 7 units of utility. If the introspection cost *i* is less than 3 units of utility, he will pay it, but if $i>3$, then he will choose arbitrarily, if he thinks either preference is equally likely. If we consider the choice facing the second agent, he can either do the same act as the first agent, or a different one. If he does the same one, then the third agent will go along with the choice, so the overall payoff will be either 27 or 9. If he does the other one, then the third agent will either choose arbitrarily (if $i>3$) or introspect and then choose in line with that (which is just as likely to agree with his current preferences or go against them). Doing the opposite action is only going to have higher expected utility than going along with the first agent if $i<½$ and the second agent introspects and recognizes that his preference goes against the act that the first agent did.

In a sense, the fact that a cost to introspection leads to unified behavior should be the least surprising result of these formal models, because the agent who hasn't yet introspected is making a decision on behalf of a mixture of her own utility function and the utility functions of other time-slices, rather than merely on her own behalf.