Autonomy for Changing Selves*

Richard Pettigrew

March 24, 2021

Our values change. What we value, want, desire, prefer, and how much; for nearly everyone, these will be different at different times in their life. These changes can be gradual or abrupt; they can be long-lasting or short-lived; and they can be induced by forces outside yourself or they can come from within or they can have no specific catalyst at all. Such preference change raises a number of questions for our theorising about rational choice, and these have been discussed at length. In §2 and §3, I'll outline two of these questions along with some of the putative solutions that have been proposed. But preference change also raises questions for our theorising about autonomy, and these have hardly been considered at all. In §4, I'll outline three problems for personal autonomy; and in §5, I'll outline one problem for political autonomy. In §6, I conclude.

1 Examples

To focus the mind, let's begin with some examples of the different ways in which our preferences might change.[1]

A gradual, long-lasting shift with no specific cause. For instance, the British playwright Alan Bennett talks of the "dreary safari" from the left to the right of the political spectrum that some undertake as they grow older.[2] Others move the other way, becoming more radical as they age. And, for some, the shift may not be political at all. They might value the life of contemplation when they are young but come to prefer the life of action

*To appear in *Routledge Handbook of Autonomy* edited by Ben Colburn.

[1]In what follows, I'll talk of preferences, values, desires, and wants interchangeably. We have conative attitudes, which encode how we'd like the world is be, and these are represented variously by preference, values, and so on. Nothing I say will turn on any distinction between different types of conative attitudes or their representation.

[2]Quoted in 'Alan Bennett launches fierce attack on private education' *The Guardian*, 17th June 2014, https://www.theguardian.com/books/2014/jun/17/alan-bennett-attack-private-education-lecture-wrong

later in their life, or they might favour a close-knit group of friends in their youth, but a larger collection of acquaintances when they are elderly.

An abrupt, long-lasting change caused by something outside the agent. For instance, many who are diagnosed with a serious illness change their values, perhaps because of the jolt administered by the news, perhaps because it allows them to see aspects of society differently as an outsider to the norm of health on which it is founded.[3] And of course there are many other major life experiences that might have a similar effect: losing a loved one; being betrayed by a friend; reading a profound work of philosophy or literature; and so on.

An abrupt, long-lasting change caused by a consequence of the agent's own choice. For instance, a person who values spending time with friends and devoting themselves to volunteering with a charity, and has little time for family, might choose to become a parent, reasonably confident that, when their child is born, their preferences will switch, and they'll value time spent with their family more.[4]

A gradual, long-lasting shift caused by a consequence of the agent's own choice. For instance, the person who values self-direction and disvalues conformity, but joins the police force knowing that people who do so nearly always assimilate by adjusting their values to those prevalent in their workplace, and for the police that involves becoming more conformist.[5] Or the person who moves to a country with a different dominant set of values and who assimilates to those.

An abrupt, short-lived change caused by a consequence of the agent's own choice. The most common examples of this are cases of temptation.[6] It's 1pm. I've been working all morning. I'd like to spend five minutes checking social media. But I know that, if I log on, I'll want to spend thirty minutes on there—that is, I know my preferences will change. But I also know they'll revert to my original preferences whenever I log off—so I know the change will be short-lived. This is a reasonably trivial case, but there are other more significant ones with the same structure: a person who is tempted to be unfaithful to their partner when they're in the presence of a particular person to whom they're attracted; people who enjoy gambling, but who dramatically increase the amount of money they're willing to lose the second they sit down at the poker table; and so on.

[3] See (Carel, 2014; Carel et al., 2016).

[4] See (Ullmann-Margalit, 2006; Bykvist, 2006; Paul, 2014; Pettigrew, 2019a).

[5] See (Bardi et al., 2014) for empirical work on how and when such preference changes happen.

[6] For a book-length treatment of such changes, see (Bermúdez, 2018).

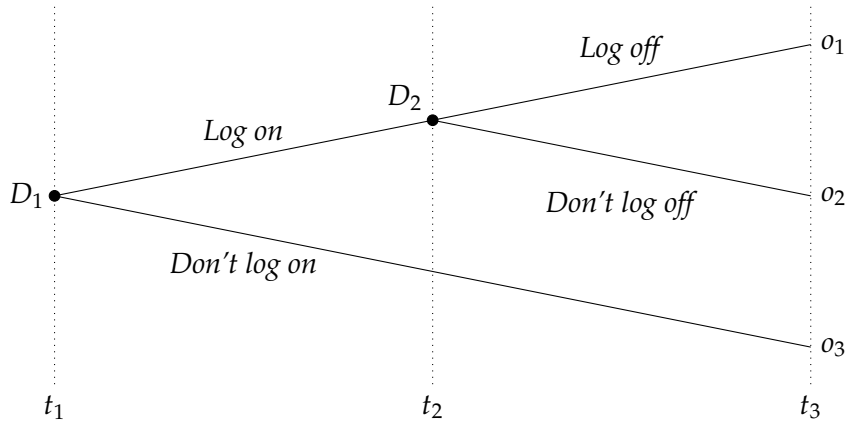

Figure 1: The standard temptation case

2 Temptation

Cases of temptation share a structure.[7] There are three outcomes: o_1, o_2, o_3. In the social media example from above:

- o_1 is the outcome in which I go on social media for only five minutes;
- o_2 is the outcome in which I go on for thirty minutes;
- o_3 is the outcome in which I don't go on at all.

And there are three times:

- t_1: 1pm—the initial time, when I am deciding whether to log on;
- t_2: 1:05pm—if I have logged on, I will be choosing whether to log off at this time;
- t_3: 1:31pm—I'll have logged off either way, whether I stay on for five minutes or thirty.

I face a decision D_1 at time t_1—log on or don't. And, if I log on, I face a decision D_2 at t_1—log off or don't. The table below gives my preferences over the three options at the three times, depending on whether or not I do log on.

	Log on	*Don't log on*
t_1	$o_1 > o_3 > o_2$	$o_1 > o_3 > o_2$
t_2	$o_2 > o_1 > o_3$	$o_1 > o_3 > o_2$
t_3	$o_1 > o_3 > o_2$	$o_1 > o_3 > o_2$

Figure 1 illustrates the decisions I face.

[7] My presentation here owes much to (Thoma, 2018).

As we can see, if I wish to get the outcome that, at 1pm, I most want (i.e., o_1) I must choose to log on. So our standard theory of rational choice seems to tell us that, at 1pm, I'm rationally required to log on. But I know that I will change my preferences once I have logged on, and rationality will require me, at 1:05pm, not to log off and instead stay on for thirty minutes to get the outcome I then value most (i.e., o_2). But then, after a series of two rational choices, I'll end up with an outcome that, at time 1pm and at 1:31pm, I least want. And indeed, there seems to be no rational way to obtain the outcome I value most at 1pm—that is, there seems to be no rational way to resist temptation. This is often thought to pose a problem for the standard theory of rational choice, which says that I should do what I most prefer at each time.

Decision theorists have proposed a host of solutions, or partial solutions. I'll enumerate five here.[8]

First: *the sophisticated choice approach*.[9] The key claim here is that, when you face a decision at one time, you should take into account what you know about how you will choose at future times, and you should choose at the earlier time with that in mind. In decision theory, we are told to choose the action with the best outcome or consequence. For one of the actions available to me at t_1—namely, logging on—its long-term outcome depends on what decision I make at t_2. If I choose not to log on at t_1, then what I do at t_2 doesn't matter; but if I choose to log on at t_1, the outcome of that action is determined by whether I choose at t_2 to log off or to stay on. I predict that I'll choose at t_2 according to the preferences I have at that time. So I predict I'll choose to stay on. So, at t_1, I know that the outcome of logging on is in fact o_2, while the outcome of not logging on is o_3. And since I prefer o_3 to o_2 at t_1, I should choose not to log on. Notice that this is a partial solution to the problem for rational choice theory posed above. It explains how I might rationally choose in a way that will not lead to the outcome that, at t_1, I consider worst. But it doesn't explain how I might rationally choose in a way that leads to the outcome I then consider best.

Second: *the binding approach*. According to this, you add a choice point just before t_1—let's call it t_0, or 12:59pm. At this time, you can choose to add an app to your phone that limits your time on social media to five minutes (decision D_0). If you choose this, you still face the choice to log on or not at t_1, but at t_2, if you have logged on, you no longer face the choice whether to log off or not—the app decides that for you, and it logs you off. Having introduced this new decision, you can now appeal to the sophisticated choice approach. If you choose to install the app, then you

[8] For a much fuller treatment, see (Bermúdez, 2018); for an overview of that treatment, see (Pettigrew, 2019b).

[9] See (McClennen, 1990; Peterson & Vallentyne, 2018).

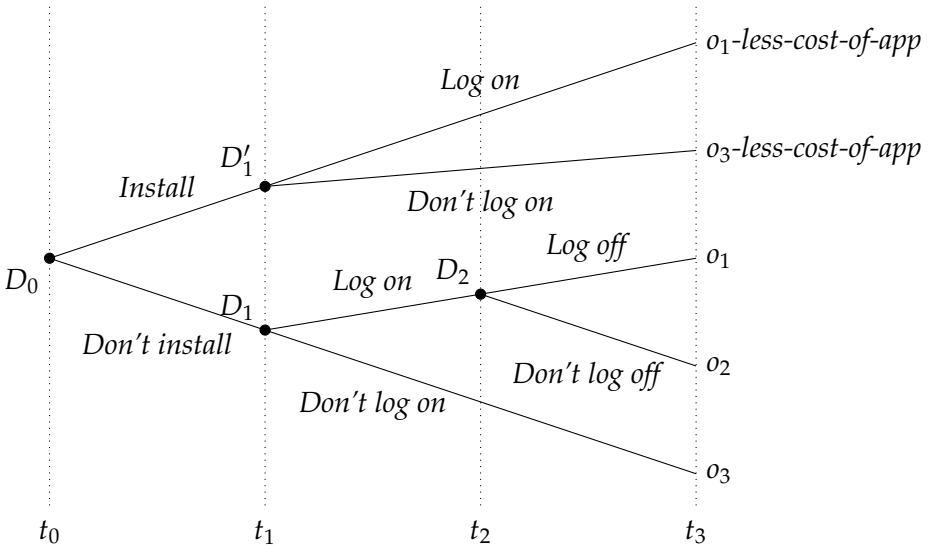

Figure 2: The binding solution to the temptation case

know you'll choose to log on at t_1, since that leads you to the best outcome that is available to you at that point, namely, o_1-*less-cost-of-app*, and this is the outcome you'll secure. If you choose not to install it, you'll choose not to log on at t_1, since you know that if you do log on, you won't log off at t_2, and you'll end up with outcome o_2, which, at t_1, you want less than o_3, which you'll secure if you don't log on. So, if you choose not to install the app, you'll end up with outcome o_3. So, providing the time-limiting app does not cost too much, and you prefer o_1-*less-cost-of-app* to o_3, then applying sophisticated choice at t_0 leads you to log on and enjoy the five minutes on social media that the app allows. This is illustrated in Figure 2 below. Again, notice this is only a partial solution to the problem posed by temptation, since it explains how it is rational to choose actions that result in o_1-*less-cost-of-app*, rather than o_1 itself, which is the preferred outcome.

Third: *the commitment approach*.[10] In standard decision theory, we represent our agent as choosing between a set of available actions, and we conceive of these as performed at a single specific time—the action of clicking the button to log in to your social media account; the action of setting off to the party where temptation awaits; and so on. On the two-tier approach, there is still a set of available actions at each time. But sometimes there is also a set of available plans or intentions or strategies or resolutions or commitments between which you might choose. We conceive of these not as performed at a specific time, but as laying out a schedule for what actions

[10]See (McClennen, 1990; Holton, 2009; Bratman, 2012).

to choose at different times when faced with specific decision problems. Thus, I might choose a plan that commits me to choosing to log on at t_1, and then log off at t_2. If I choose this plan at t_1 and then stick with it at t_2, I will obtain the outcome o_1 that, at t_1, I most want.

Of course, having enriched our set of options with these plans, we must say which plan rationality requires us to choose at a given time; and then we must explain why, having chosen one, rationality requires us to adhere to it at a later time.

On one approach, rationality says that you must choose whichever plan you will consider best at all times, if such exists; and then, at each time the plan covers, you should do what it says at that time.[11] The problem, of course, is that, in cases of temptation such as we consider here, no such plan exists: a plan to choose to log on at t_1 and to log off at t_2 is best by the lights of your preferences at t_1, but not by the lights of your preferences at t_2; a plan to log on and then not to log off at t_2 is best at t_2; but not t_1; and a plan not to log on at t_1 isn't best at either time.

Another approach, which is known as *resolute choice*, tells you to choose at t_1 whatever plan will get you what, at t_1, you most prefer were you to follow it to the letter at all times it covers; and then, at each time it covers, you should follow it to the letter.[12] The problem with this is that it seems to rule out as irrational ever forming new preferences and acting upon them. After all, as soon as a plan to cover your whole life is available, you will choose the one that best satisfies your preferences at that time; and henceforth, you will be bound to follow that plan. Relatedly, it isn't clear why the dead hand of the past should have as tight a grip on your decisions as this account demands.

Fourth: *the group rationality approach*. According to this, we should treat the person who faces the decision to log on or not at t_1 and the person who might face the decision to log off or not at t_2 as separate selves.[13] They are both part of the same person, namely, me; but they are distinct selves with distinct preferences. Once we conceive of them like that, we can see their decisions not as two decisions faced by one person, but as a single decision made by one self and a different single decision made at a later time by another, different self. Together, these selves form a group whose members are making decisions individually, but in the service of the group as a whole—akin to a group of artisans each working on a different component of an elaborate clock that they are building together. And there are a number of theories of group rationality that we might bring to bear on

[11] See (Gauthier, 1994).

[12] See (McClennen, 1990).

[13] See (Peleg & Yaari, 1973; Gold, 2018; Pettigrew, 2019a).

the question of how individuals in such a group should choose. We might, for instance, appeal to game theory, which tells us how two individuals with different preferences should choose when they each face separate decisions whose outcomes depend on the other individual's choice. Or we might appeal to voting theory, or other methods of judgment aggregation, which tell us how to pool the preferences of the individuals in a group to give the group's preferences, and then we might choose on the basis of those. Whichever approach we take, the idea is that, because the two selves identify as part of a group or team devoted to choosing together, they can rationally compromise and cooperate.

Fifth: *the true self approach*. According to this, we should distinguish between those preferences that issue from your 'true self' and those that issue from some normatively less compelling source.[14] The idea is that, in cases of temptation, what you feel you want to do when under the influence of the temptation does not reflect the preferences of your true self. Rather, it issues from somewhere else. In the social media case above, my true self wants to spend only five minutes online, but some more base urge drives me to spend longer on there once I have logged on. According to this approach, my true preferences don't change between t_1 and t_2. What seem to be new preferences at t_2 do not issue from my true self; because of this, they have no normative force. Instead, rationality requires me to choose in line with my true preferences. So, I should choose to log on at t_1 and log off at t_2. If I don't, I'm irrational.

3 Transformative experiences

Perhaps it's because of the sort of examples that motivate our treatment of temptation; perhaps it's because in cases of temptation our preferences change and then change back again after quite a short period of time and only last so long as we're in a particular situation, such as logged on to social media or sitting at the poker table; but we are more enthusiastic to support solutions that privilege our preferences at time t_1 than those that privilege t_2 or even treat the two equally. But there are cases of changing preferences in which that is not the case. These are cases in which the change is long-lasting, and in which the catalyst is not simply being in a particular environment, but rather a profound change in your life. Some have argued that this sort of preference change also challenges our theory of rational choice.[15]

Again, let's begin with an example. I am choosing whether or not to adopt a child and become a parent. At the moment, my life is quite settled. I

[14] See (Mele, 2018; Thoma, 2018; Pettigrew, 2019a).

[15] See, for instance, (Ullmann-Margalit, 2006; Bykvist, 2006; Paul, 2014).

have a very close group of friends, whom I see regularly, and I spend much of my free time helping to run an evening club for cognitively disabled adults. At the moment, I prefer that life to the life I would lead as a parent, where my time for these friends and this evening club would be much reduced. But I know that, were I to adopt and become a parent, I'd value doing that more. I'd come to love the child I adopt and I'd come to want them to flourish so much that I'd prefer spending the majority of my time ensuring their happiness to spending it with my friends or on my work for the club. The experience of becoming a parent, like the experience of moving to a new country, falling in love, or learning you have a terminal illness, are what L. A. Paul (2014) calls *transformative experiences*. By having them, they change core parts of who you are, including certain core values. How, then, should I choose when faced with the adoption process or any other decision on which one or more of the available actions might lead to a transformative experience? Such decisions Edna Ullmann-Margalit (2006) calls *big decisions*. Decision theory typically tells me to choose whichever option I most prefer. But, relative to my child-free preferences, I prefer not to become a parent, while relative to the preferences I would have if I were to adopt, I prefer to be a parent. So when I choose the option I most prefer, which set of preferences should I use? The ones I have at the point I make the decision? The ones I have when the effects of the decision are being felt? Some other preferences entirely? Or perhaps some combination of these? On this question, you might think, standard decision theory is silent. The challenge that such cases of preference change pose is how to say something principled about them.

As in the case of temptation, a number of putative solutions have been proposed, and indeed some resemble the solutions in that case.

First: *the unchanging preferences solution*.[16] In some ways this resembles the true self approach from above, for it also denies that there has been any real change in the true preferences. But this time the reason is not that my child-free or my potential parent preferences don't issue from the true self, it's just that both are reflections of a single underlying preference that remains the same throughout. To motivate the idea, think of someone who, at twenty years old, prefers going on a rollercoaster to going on a vintage steam train, while at ninety prefers the train. In this case, we might say that their preferences haven't changed: in both cases, they prefer doing whatever gives them most physical pleasure; what has changed is what gives them that. When they are young and the rollercoaster gives only the rush of adrenaline and no physical discomfort, it is the rollercoaster; when they are older and it makes them feel sick and faint and their chest feels tight, it is the train. On the unchanging preferences solution, all appar-

[16] See (Stigler & Becker, 1977; Becker, 1998; Nagel, 1978).

ent preference change is like this: the underlying preferences remain the same; what changes are the features of the world that determine what specific outcomes will satisfy those preferences. In the case of the rollercoaster and steam train, for instance, it is the physiology of the individual's body. One problem with this is that, in the case of adopting a child, it isn't clear what plays the role that physical pleasure plays in the rollercoaster example: what is it exactly that I will steadfastly prefer throughout the change wrought by becoming a parent that will lead me to want not to be a parent before that change occurs but will lead me to want to be a parent afterwards?

Second: *the higher-order preference approach*.[17] According to this, when our first-order preferences change, we should appeal to our higher-order preferences to adjudicate. When I choose whether or not to adopt, and I see that my current preferences are different from what they'll be if I do adopt, I should ask myself: which of these preferences would I prefer to have? Do I prefer being someone who values my friendships and my work at the evening club over raising a child? Or do I prefer being someone whose first-order preferences run the other way? If the former, I should not adopt; if the latter, I should. One problem with this suggestion is that my second-order preferences often change in tandem with my first-order preferences. At the moment, I prefer my current, child-free preferences; if I adopt, not only will my first-order preferences change, but so will my second-order preferences, and they'll change so that they prefer the first-order preferences that I'll then have. So the higher-order approach merely pushes the problem back a step: to which second-order preferences should I appeal when I make my decision? Another problem is similar to one that arises in discussions of Frankfurt's and Dworkin's approaches to autonomy, which also appeal to second-order preferences. Why should second-order preferences take normative priority? Why should we bring them into alignment by changing our first-order preferences? Why not change our second-order preferences?[18]

Third: *the group rationality approach*.[19] This is very similar to the approach of the same name from above. Again, it asks you to view a person as composed of a set of successive selves, some of whom will have different preferences from the others. When any one of those selves is called upon to make a decision, they might be required to take into consideration the preferences of the others, and to aggregate those preferences with theirs in some way and use the aggregate preference to make their choice. Let me illustrate with my own favoured version of this approach. Standard decision theory

[17] See (Ullmann-Margalit, 2006).

[18] See (Friedman, 1986).

[19] See (Pettigrew, 2019a).

represents an individual's preferences over outcomes by their utility function, which assigns to each outcome a number that measures how much the individual values that outcome. Suppose we represent in this way each of my selves at different times. Then we might say that, when their turn comes to have decision-making power—that is, at the time when that self exists—each should make decisions using an aggregate utility function that incorporates their own utility function but also the utility functions of my other selves, albeit to different extents. For instance, we might say that the aggregate utility function should be a weighted average of the utility functions of all the selves that make up me. So, in the adoption case, my current self should decide using a utility function that is obtained by averaging their utility function and the utility functions of my past and future selves. Thus, the aggregate utility for the outcome in which I adopt is a weighted average of my current utility for that outcome and the utilities assigned to it by my past selves and the utilities assigned to it by those of my future selves that would come into being were I to adopt. This approach raises the question: what weight should be given to the utilities of the various different selves when we produce the aggregate utility function? Does the current, decision-making self receive more weight than others? Must all selves receive some possible weight? Even the past selves? Are some selves worthy of more weight than others, and why?

4 Personal autonomy for changing selves

These, then, are some of the problems that the phenomenon of preference change raises for our theorising about rational choice, together with a sketch of some proposed solutions. In this section, I'd like to describe some problems it raises for our theorising about autonomy.

The Problem of the Fractured Self. Autonomy is self-authorship. In Kant, this is understood in an extreme form. To be autonomous, your preferences must arise directly from the will; they should not be caused by anything outside that will.[20] Later authors do not require so much. Your preferences might be the result of external influences, such as the society or family in which you're raised or your genetic make-up; but autonomy requires that you would, upon reflection, endorse them. For some, this means that your first-order preferences are endorsed by your second-order preferences.[21] Others, who question why second-order preferences should occupy such a privileged normative role, understand endorsement differently.[22] But, however we understand it, the phenomenon of changing preferences raises

[20] See Section 3 of (Kant, 1785 [1997]).

[21] See (Frankfurt, 1971; Dworkin, 1976, 1988).

[22] See, for instance, (Raz, 1988; Colburn, 2010).

a question for such accounts. For such reflective endorsement is surely determined, at least in part, by some aspect of the individual's preferences. And since those change over time, so might the individual's disposition to endorse. And so arises the question: at what time must you endorse your preferences to count as autonomous? At the time at which you have them? But then the addict will count as autonomous, which some wish to deny. Or at some other time? But then many people who undergo changes wrought by transformative experience will fail to count as autonomous, for they will not, at the earlier time, endorse the preferences that they come to have at the later time—for instance, I don't currently endorse the preferences I would come to have were I to adopt a child.

In general, the phenomenon of changing preferences bolsters what is sometimes called the postmodernist critique of certain accounts of autonomy.[23] This critique is directed against those accounts that require a unified, constant self that charts its course through life and that is able to reflect rationally on the preferences it has at a given time and give a verdict on them. In a life that contains many changes in preferences, it looks difficult to find this unified, constant self. The phenomenon of preference change drives us more towards a view of persons as corporate entities made up of different selves at different times, often with different preferences. And from that point of view, some accounts of autonomy founder.

The Problem of the Unit of Autonomy. This last observation leads to a different, but related problem: what is the unit of autonomy? Is it the person as a whole? Or is it the individual selves that constitute that person? Often, in theorising about autonomy, philosophers draw the distinction between local and global accounts of autonomy. Local autonomy is a property of a person at a particular time in their life, and often more specifically the choices they make at that time; global autonomy is a property of their whole life. Perhaps we can apply the same distinction in response to our question here: selves can be locally autonomous; persons can be globally autonomous. One problem with this suggestion is that selves might endure for a long time. They need not necessarily be momentary entities, but might instead last from one change of preferences to another, which might be a period of years. Another problem is that, even if we say that different concepts of autonomy apply to selves and to the persons they compose, we must say which of the two properties corresponding to those concepts we wish to promote, and if we wish to promote both, which takes precedence when they clash? After all, we study autonomy not merely as an exercise in conceptual analysis, but because we think that the property of being autonomous is one that has normative significance in ethics and politics.

[23] See (Mackenzie & Stoljar, 2000, 10-11) and (Benhabib, 1992, Chapter 5).

The Binding Problem. We see a specific instance of this problem of precedence when we consider the binding solution to the problem of temptation. According to that, when I foresee that my preferences will change temporarily in the future as a result of a particular experience, such as logging on to social media or sitting down at a poker table, and I foresee that I must put myself in that position if I am to have a chance of getting what I now most want, then I am rationally required or permitted to bind my future, tempted self so that they cannot choose the outcome that they will most want. As I spelled it out above, I effect this binding by removing their favoured option for them—in the social media case, by downloading an app; in the gambling case, perhaps by telling the gambling house or website not to let me play beyond a certain stage. That is, I remove an action from the set of those that will be available to my future self. But many theorists of autonomy hold that you deplete an individual's autonomy when you remove their options—particularly if they are options that they would have chosen over the others had they been available.[24] So, if the unit of autonomy is the self, then it seems that binding my future selves reduces their autonomy. If, on the other hand, it is the whole person that is the unit, then we might still say that they are autonomous, for they have, as a single entity, charted their course through life in a way that they will largely endorse.

5 Political autonomy for changing selves

The Problem of Paternalism. Many liberals prize very highly a society in which individuals are autonomous; a society whose members are able to chart their own course through life, pursuing what will satisfy their preferences, at least insofar as they do not thereby frustrate the preferences of others, and reflecting on those preferences and either endorsing them or seeking to alter them. Some think that it is compatible with such a goal that either the government or private companies or citizens seek to alter the decisions of others, sometimes without explicitly telling them that they are doing this. Such attempts are sometimes called nudges, and they became popular in some countries over the past twenty years, building on research in social psychology that showed how effectively to do this.[25] For instance, we know that people are more likely to choose an option when it is placed first on a list than they are if it is placed lower down. So, a government might nudge individuals towards choosing a particular option—agreeing to donate their organs after death, or agreeing to receive the annual 'flu shot—by placing that at the top of a list of related alternatives.

[24] See (Raz, 1988, Chapter 14).

[25] The *locus classicus* is (Thaler & Sunstein, 2008).

When are such nudges compatible with the autonomy of the individuals who are being nudged? Some will say that they are always compatible: after all, no option has been removed; all are still available, and just as easily; all that has changed is how the decision is presented. But others will worry that there are less benign cases. For instance, some might think that subliminal messaging is not compatible with autonomy.[26]

One response is to apply what Thaler and Sunstein call the 'as judged by themselves' test. This says that a particular nudge is compatible with autonomy if the person nudged is happy that they were influenced this way and happy with the decision so made. Thus, while they might be indifferent to the annual 'flu shot, or perhaps even a little bit against organ donation, it is permissible to nudge them towards both if, after having been successfully nudged, and after the mechanism by which it was done is laid out before them, they are happy that they were.

However, as Paul & Sunstein (ms) note in recent work, the phenomenon of preference change through transformative experience raises problems for this criterion. Suppose I am dead set against adopting a child. My current life, with my close group of friends and my work with the disabled club, is much more valuable to me now. However, my government very heavily nudges me towards adopting. And indeed they do so through a variety of rather sinister and covert methods: subliminal messaging, emotional blackmail, and so on. Finally, I agree to adopt. Over the coming months, I come to love my adopted child so much that, when the government's psychological techniques are revealed to me, I declare myself happy that they did this—after all, raising this child is the most important thing in my life and I wouldn't have done it otherwise. The government's nudges pass the 'as judged by themselves' test, but they seem incompatible with autonomy.

So, the question arises: if the 'as judged by themselves' test fails in these cases, is there a better test available?

6 Conclusion

The phenomenon of changing preferences is varied and widespread. It has raised a number of challenges for theories of rational choice, and these have occupied economists, psychologists, and philosophers for a number of decades. But the challenges that they raise for theories of autonomy have been less commonly discussed. As we saw, there are a number: the problem of the fractured self, the problem of the units of autonomy, the binding problem, and the problem of paternalism. No doubt there are many more.

[26] See (Dworkin, 1988, 3-20).

I have sketched the problems here, but I have not offered anything in the way of solutions. I hope those who concern themselves with the theory of autonomy will find it worth their time to address them.

References

Bardi, A., Buchanan, K. E., Goodwin, R., Slabu, L., & Robinson, M. (2014). Value stability and change during self-chosen life transitions: Self-selection versus socialization effects. *Journal of Personality and Social Psychology*, *106*(1), 131–146.

Becker, G. S. (1998). *Accounting for Tastes*. Cambridge, Mass.: Harvard University Press.

Benhabib, S. (1992). *Situating the Self: Gender, Community, and Postmodernism in Contemporary Ethics*. London: Routledge.

Bermúdez, J. L. (Ed.) (2018). *Self-Control, Decision Theory, and Rationality: New Essays*. Cambridge, UK: Cambridge University Press.

Bratman, M. (2012). Time, Rationality, and Self-Governance. *Philosophical Issues (Supp. Noûs)*, *22*(1), 73–88.

Bykvist, K. (2006). Prudence for changing selves. *Utilitas*, *18*(3), 264–283.

Carel, H. (2014). The Philosophical Role of Illness. *Metaphilosophy*, *45*(1), 20–40.

Carel, H., Kidd, I. J., & Pettigrew, R. (2016). Illness as Transformative Experience. *The Lancet*, *388*(10050), 1152–53.

Colburn, B. (2010). *Autonomy and Liberalism*. London: Routledge.

Dworkin, G. (1976). Autonomy and Behaviour Control. *The Hastings Center Report*, *6*, 23–28.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge, UK: Cambridge University Press.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*, 5–20.

Friedman, M. A. (1986). Autonomy and the Split-Level Self. *The Southern Journal of Philosophy*, *24*(1), 19–35.

Gauthier, D. (1994). Assure and Threaten. *Ethics*, *104*(4), 690–721.

Gold, N. (2018). Putting Willpower into Decision Theory: The Person as a Team over Time and Intrapersonal Team Reasoning. In J. L. Bermúdez (Ed.) *Self-Control, Decision Theory, and Rationality*, (pp. 218–239). Cambridge, UK: Cambridge University Press.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Kant, I. (1785 [1997]). *The Groundwork of the Metaphysics of Morals (trans. M. Gregor)*. Cambridge, UK: Cambridge University Press.

Mackenzie, C., & Stoljar, N. (2000). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford: Oxford University Press.

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

Mele, A. R. (2018). Exercising Self-Control: An Apparent Problem Resolved. In J. L. Bermúdez (Ed.) *Self-Control, Decision Theory, and Rationality*. Cambridge, UK: Cambridge University Press.

Nagel, T. (1978). *The Possibility of Altruism*. Princeton University Press.

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

Paul, L. A., & Sunstein, C. (ms). "As Judged By Themselves": Transformative Experiences and Endogenous Preferences. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455421.

Peleg, B., & Yaari, M. (1973). On the existence of a consistent course of actions when tastes are changing. *Review of Economic Studies*, *40*(3), 391–401.

Peterson, M., & Vallentyne, P. (2018). Self-Prediction and Self-Control. In J. L. Bermúdez (Ed.) *Self-Control, Decision Theory, and Rationality*. Cambridge, UK: Cambridge University Press.

Pettigrew, R. (2019a). *Choosing for Changing Selves*. Oxford, UK: Oxford University Press.

Pettigrew, R. (2019b). Review of J. L. Bermúdez (ed.) *Self-Control, Decision Theory, and Rationality: New Essays*. *Notre Dame Philosophical Reviews*.

Raz, J. (1988). *The Morality of Freedom*. Oxford: Oxford University Press.

Stigler, G. J., & Becker, G. S. (1977). De Gustibus Non Est Disputandum. *The American Economic Review*, *67*(2), 76–90.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Thoma, J. (2018). Temptation and Prefence-Based Instrumental Rationality. In J. L. Bermúdez (Ed.) *Self-Control, Decision Theory, and Rationality*. Cambridge, UK: Cambridge University Press.

Ullmann-Margalit, E. (2006). Big Decisions: Opting, Converting, Drifting. *Royal Institute of Philosophy Supplement*, *81*(58), 157–172.