

Krister Bykvist¹

Transformative Choices and the ‘As Judged by Themselves’- Criterion²

Abstract: One way to evaluate various interventions in people’s lives, including ‘nudges’, is to ask whether they make people better off, as judged by the people themselves. Paul and Sunstein (2019) show that this criterion does not provide determinate answers when the interventions are transformative choices. Pettigrew (2023) argues that there is a way to understand the criterion that provides both determinate and plausible answers. Pace Pettigrew, I shall argue that it is not easy to find a plausible version of the ‘as judged by themselves’-criterion, especially if it is supposed to be both ethically sound and usable for public decision-makers.

¹ Stockholm University/Institute for Futures Studies, krister.bykvist@iffs.se

² Funding from the Swedish Research Council (1-VR-041) is gratefully acknowledged.

1. Introduction

One way to evaluate interventions in people's lives is to ask whether they make them better off, as judged by the people themselves. Paul & Sunstein (2019) apply this 'as judged by themselves'-criterion (the AJBT-criterion, for short) to nudges and show that it does not provide determinate answers when the nudges are transformative and change the judgments of the targeted people. Pettigrew (2023) argues that there is a way to understand the criterion that provides both determinate and plausible answers in cases of transformative nudges. I shall argue that his account is inadequate and that it is unclear whether an acceptable version can be found, especially if we want it to be both ethically plausible and usable for public decision-makers.

Here is the outline of the paper. In section 2, the notion of nudging is introduced and the first pass at formulating the AJBT-criterion is discussed and found wanting. Section 3 discusses the need for idealized judgments. Pettigrew's version of the criterion is critically discussed in section 4. In section 5, I tentatively propose a different version of the theory that avoids some of the problems with Pettigrew's proposal, but, unfortunately, it is not itself problem-free.

2. The AJBT-criterion at a first pass

In the most general sense, a nudge is a benevolent intervention in people's lives that steers them in the right direction while preserving their freedom. It is supposed to be applied in cases where coercion is not justified. Nudges often make use of features of the choice situation, 'the choice architecture', as it is often called, to influence choices. Examples include default rules designed to increase participation in pension schemes and order effects, like putting the healthy snack by the check-out counter. The important question here is what counts as steering people in the *right* direction. It is not supposed to be the direction the decision-maker happens to see fit. Rather, the direction of the steering is supposed to be benevolent in the sense that the steering is done for the sake of the people. Sunstein and Thaler argue therefore that the relevant test or criterion is whether the target people are made better off, *as judged by themselves*. As Paul & Sunstein (2019) puts it,

The ['as judged by themselves' or] AJBT criterion, as we shall call it, asks whether those who have been nudged *ex ante*—for example, with a warning or a reminder—deem themselves to be better off *ex post* as a result. (Paul & Sunstein 2019, 2)

So, the idea seems to be at least this, providing sufficient conditions for permissible and impermissible nudges respectively:

AJBT Ex Post

- A nudge is *permissible*, if the nudgee would judge ex post themselves to be *better off* as a result.
- A nudge is *impermissible*, if the nudgee would judge ex post themselves to be *worse off* as a result.³

‘Better off as a result’ should be understood as ‘better off in the nudged future than in the alternative future’, assuming there is only one alternative, which I will for simplicity do in the following.

Paul & Sunstein (2019) think this criterion can be applied by people when they are making their own choices, ‘self-nudging’, but also by private and public decision-makers who want to steer other people in the right direction. This criterion would for example condone a default rule for participation in a pension scheme, if people would judge ex post themselves to be better off as a result of this nudge. Again, it is important to note that the domain of application is restricted to cases where coercion is not justified. So, the domain does not include severe cases of self-harming where one can justifiably use coercion to prevent the harm.

One reason to adopt this criterion, stressed by Sunstein (2018), is that a permissible nudge will be something that the nudgee herself would endorse ex post. In this sense, it would be in line with the *means paternalism*: we are allowed to steer people in the direction of what would satisfy their own ends. In contrast, *end paternalism* says that we can be permitted to steer people towards ends that they do not share, and which are defined by some more objective criterion of benevolence. Another reason to adopt this criterion, which Paul & Sunstein mentions, is that people are often the best judges of their own good. Often what people think will make them better off will in fact make them better off. I would like to add a third reason, namely, that when people *wholeheartedly* judge something to be better for them, they also have a *positive non-cognitive attitude* towards it – they favour it, prefer it, or take pleasure in it; and the satisfactions of these attitudes contribute to a person’s wellbeing, according to attitude-sensitive wellbeing theories. So, a wholehearted judgment that your life circumstances are good for you will provide an attitude whose satisfaction makes your life better to some extent.

Paul & Sunstein (2019) show that this criterion runs into problems in cases where

³ In stating this criterion, Pettigrew (2023) uses the terms ‘legitimate’ and ‘illegitimate’ instead of my preferred terms ‘permissible’ and ‘impermissible’. But nothing important hangs on this.

the nudge is *personally transformative*, i.e., changes one's fundamental values and attitudes. (I will put aside epistemically transformative choices in the following, choices which enable the subject to grasp and value things they could not grasp and value before the transformation.) They think the main challenge is that in some transformative nudge cases the criterion is *indeterminate* in that it does not specify a unique solution, because no matter what the decision-maker does, the nudgee would judge ex post herself to be better off as a result. They exemplify this with the following cases, among others.

Diane and the health care plans

Diane is automatically enrolled in a Bronze Health Care Plan – it is less expensive than Silver and Gold, but it is also less comprehensive in its coverage, and it has a higher deductible. Diane prefers Bronze and has no interest in switching. In a parallel world (a lot like ours, but not quite identical), Diane is automatically enrolled in a Gold Health Care Plan – it is more expensive than Silver and Bronze, but it is also more comprehensive in its coverage, and it has a lower deductible. In that parallel world, Diane prefers Gold and has no interest in switching. (Paul & Sunstein 2019, 4)

No matter which health plan Diane is automatically enrolled to, she would judge ex post herself to be better off as a result. So, the ASJT-criterion would say that nudging Diane to choose either plan would be fine. They also list more radical examples where more fundamental values seem to change as the result of nudges.

Rick, the patient

Rick has a serious illness. The question is whether he should have an operation, which carries with it with potential benefits and potential risks. Reading about the operation online, Rick is not sure whether he should go ahead with it. Rick's doctor advises him to have the operation, emphasizing how much he has to lose if he does not. He decides to follow the advice, and a year later, he is glad he did. In a parallel world (a lot like ours, but not quite identical), Rick's doctor advises him not to have the operation, emphasizing how much he has to lose if he does. He decides to follow the advice, and a year later, he is glad he did. (Paul & Sunstein 2019, 4)

In Pettigrew (2023, 6), another case is presented in which fundamental values are at stake.

Happy Parent

Happy Parent I am currently child-free, and I am deciding whether or not to

adopt. At the moment, I value remaining child-free more than I value adopting a child and becoming a parent. When I look forward to the two possible futures ahead of me, one in which I am a parent and one in which I am child-free, I value the latter more. I value the things I will be free to do in that possible future: the time it will allow me to strengthen and deepen the bonds with my friends; the volunteering opportunities it will afford me the time to pursue; the extra money I'll have available that I can use to pursue the projects I love and donate to the causes that matter to me; and so on. However, I've spoken to enough parents to know that, were I to adopt, these preferences are likely to reverse. I will likely form a bond with my adopted child so strong that I will prefer the life in which I care for them to the one in which they are not in my life.

No matter what choice the prospective parent made, he would judge *ex post* herself to be better off as a result. So, the AJBT-criterion nudging her to do either is fine. Again, there is no unique verdict.

One might object here that the lack of a unique verdict is nothing to worry about. After all, we are told that either action is permissible. That is a determinate verdict, even though it does not single out one action as *the* right action. But the following version of the Happy Parent case shows why the permissibility verdict is in fact troublesome.

Figure 1

	Choosing to be parent	Choosing to be child-free
Parent	5	2
Child-free	2	10

The matrix shows how much the agent would value the outcomes of the alternative actions depending on whether she was a parent or childfree. As can be seen from the matrix, the prospective parent would assign a *lower* value to their parent-life, if they became a parent, than they would assign to their child-free life, if they remained childfree (the values marked in red). It seems odd then to say that nudging them to be a parent and nudging them to be child-free must both be *equally permissible*. Surely, something speaks against nudging them to be a parent, namely, that they would then rate their life lower than they would if they remained child-free. In that sense, their own end-satisfaction is lower in the life as a parent than in the child-free life. Indeed, if their *ex ante* judgment agreed with their judgment as a parent, then it seems clear that nudging them to become a parent is *impermissible*.

Figure 2

	Choosing to be a parent	Choosing to be child-free
Ex ante	2	10
Parent	5	2
Child-free	2	10

As Pettigrew puts it, if I was nudged to become a parent by the government, ‘not only has the government nudged me to take means to satisfy ends I don’t not have, but they’ve done so knowing that, by doing this, I will change my ends as well. The means paternalist is horrified.’ (Pettigrew 2023, 6). Furthermore, nudging me to become a parent would not just go against my current judgment but also lead to a lower future end-satisfaction.

Indeed, we can make this case even more convincing by assuming that if they remained child-free, they would think their parent life was *bad* (‘overwhelmed by all the challenges’) but think their child-free life would be even worse (‘can’t see myself without my child’)

Figure 3

	Choosing to be a parent	Choosing to be child-free
Ex ante	2	10
Parent	-5	-7
Child-free	2	10

Here it is even clearer that it is *impermissible* to nudge them to become a parent when we have the option to judge them to stay childless. These examples show that we should not accept the first part of AJBT Ex Post. We should not accept that a nudge is permissible, if the nudgee would judge ex post themselves to be better off as a result.

The second part of *AJBT Ex Post* is also unacceptable as the following case shows. Suppose that you are in Kierkegaard’s situation. If you stay unmarried, you will regret that and judge that you are worse off unmarried; if you marry you will regret that too, and judge that you are worse off married. Here the criterion would say that nudging you to do either action would be *impermissible*. So, if these are the only two options, we will be in a bind as a nudger. No matter which direction we steer the person, it would be wrong to do so. Now, some might be OK with normative dilemmas, or might think that there are no cases of forced nudging (see section 5 for more on this), but the following version of Kierkegaard’s case shows why it would be a mistake to think that both nudge-options must be wrong.

Figure 4

	Choosing to be married	Choosing to be unmarried
Married	10	20
Unmarried	-10	-20

Here it would be absurd to think that there must be a dilemma where each nudge-option is equally wrong, since the nudgee would *positively* value their married life, if they got married, and *negatively* value their unmarried life, if they remained unmarried. Indeed, it seems clear that it would be *permissible* (and not impermissible) to nudge them to be married, if they valued ex ante getting married as much as they would if they got married, and, moreover, judged it better than remaining unmarried.

Figure 5

	Choosing to be married	Choosing to be unmarried
Ex ante	10	5
Married	10	20
Unmarried	-10	-20

Now, one might question the assumption that it is possible to have a forced choice between nudging you to be married and nudging you to stay unmarried (I will say more about this in section 5). Is it not always possible to refrain from nudging? Even if it is possible to refrain from nudging in the case above, the objection stands. Suppose that if no nudging took place, you would choose to be unmarried with the consequence that you will come to hate it (-20). Surely, it is still permissible to nudge you to get married with the consequence that you will love it as much as you love it ex ante (10).

So, this example shows that the second part of AJBT Ex Post is also mistaken. It is not true that a nudge is impermissible, if the nudgee would judge ex post themselves to be worse off as a result. Furthermore, this case also shows that it would be a mistake to think that A nudge is permissible *only if* the nudgee would ex post judge themselves to be *better off* as a result. So, we cannot turn the sufficient condition for permissible nudges in AJBT Ex Post into a sufficient and necessary condition for permissible nudges.

Can we live with AJBT Ex Post being it is *silent* on many cases, because it only provides sufficient conditions for permissible and impermissible nudges? That is doubtful. There are cases where it is clear that an acceptable nudge-criterion should not be silent but where AJBT Ex Post is silent. Here is a schematic case.

Figure 6

	Choosing A	Choosing B
Ex ante	5	5
After A	5	5
After B	5	2

Here the nudgee is indifferent ex ante between A and B. Furthermore, they would not judge themselves to be better off ex post by being nudged to do A; instead, they would be *indifferent* between A and B. So, AJBT Ex Post will have to be silent on whether A is permissible. But surely, this should be deemed permissible, and it would be impermissible to nudge them to do B. If A were chosen, the nudgee would be positively value A and B to the same positive degree, whereas if B were chosen, they would end up with an alternative they would value less than they would have valued A, if A were chosen.

One could fix the silence issue in this case by revising the AJBT-criterion in this way:

- A nudge is permissible, if the nudgee would judge ex post themselves to be *at least as well off* as a result.

Then we can say that A is a permissible nudge-option in the previous case. However, this version does not help with the parent cases, since if one judges oneself to be better off, then one judges oneself to be at least as well off, at least if one is minimally rational and understands that ‘better off’ entails ‘at least as well off’. This means that the revised AJBT-criterion must say that it permissible to nudge the person to be a parent in the second parent case, even though this will have as a result that she will hate her parent-life when she could have been judged to be child-free, which would have resulted in a life she would love.

Note that judging oneself ex post to be at least as well off is not even a *necessary* condition for permissibility. In other words, this principle does not hold:

- A nudge is permissible only if the nudgee would ex post judge themselves to be *at least as well off* as a result.

The second version of the Kierkegaard case shows why. Here it is permissible to nudge the person to get married, even though she would then judge herself to be worse off, and thus not as well off (if she is minimally rational).

3. The need for idealised judgments

It is clear that the AJBT-criterion has to be fixed and that the fixes we have considered so far are insufficient. Pettigrew considers two further possible fixes. They both involve considering the *ex ante perspective* before the nudge as well as the *ex post perspective* after the nudge. Before we consider these fixes, we need to sort out one issue about *ex ante* judgments I have skated over so far. One might wonder how it is even possible to consider *ex ante* judgments about being nudged given that a nudge might not work very well if the nudgee is aware of being nudged. As Pettigrew points out,

If you're told that an option has been placed at the top of the list because the person who made the list wants you to choose it, your contrarian side might kick in and you might be minded to thwart their attempt by choosing something else, or you might simply randomise so as not to feel a dupe. (Pettigrew 2023, 7)

Pettigrew's answer is to understand the *ex ante* judge as a *hypothetical* judgment, a judgment about the actual situation that one would make in a certain *idealized* situation. Pettigrew thinks that in this idealized situation 'the nudgee is given as much time as they need to reflect on the choice; any irrational ways of thinking, such as status quo bias, temptation, etc. are removed; they are provided with all the evidence relevant to the choice; any cognitive limitations, such as limitations in their logical or statistical reasoning, are removed; and they are equipped with unlimited cognitive resources, such as computing time and power, with which to assess the evidence.' (Pettigrew 2023, 4).

However, Pettigrew's answer invokes a rather extreme form of idealization that will take us very far from the actual situation. Indeed, the stipulated changes to the psychological make-up of the actual person might make the ideal person's foundational values rather different from the actual person's. Who knows what I would value if I had no normal psychological limitations? Perhaps I would love the fancy wine I now fail to appreciate. Perhaps I would come to love mathematics, I now found so difficult and prefer it to philosophy. But so what? A means paternalist should be interested in providing the means to some of our actual ends, not the means to some *merely hypothetical* ends. To ignore actual ends and focus on satisfying merely hypothetical ends instead thus smacks of end paternalism. It would also be pretty pointless to give you the means to merely hypothetical ends. Why should we give you advanced maths books that you cannot even understand and have no interest in understanding just because your ideal self would love advanced maths. Furthermore, if the AJBT-criterion is supposed to be used in practice as a test for legitimate policies, it will be

very difficult to apply it correctly if one has to rely on these extreme idealizations, since it is so difficult to know what a person would judge under extreme idealizing conditions (it might not even have a determinate answer).

Pettigrew is right that we need to ask a hypothetical question, but this can be done without making extreme idealizations. It is enough to assume that the nudgee has a current *latent* view about being nudged and that it would survive some exposure to relevant information. We may use their past behaviour and speech as evidence of their view on being nudged in the actual situation.

4. Pettigrew’s proposals

Let us now return to Pettigrew’s proposals. The first fix is to apply the AJBT-criterion to both the ex post and ex ante judgments about being better off.

AJBT Ex Ante/Ex Post

- A nudge is permissible if and only if the nudgee would judge themselves to be better off both ex ante and ex post as a result of the nudge.

This will help with indeterminacy issues. Take our first parent case.

Figure 7

	Choosing to be a parent	Choosing to be child-free
Ex ante	2	10
Parent	5	2
Child-free	2	10

Then, on the revised criterion, it would be permissible to let me be child-free and impermissible to nudge me to become a parent.

However, there are problems with this version. First, it will give the wrong verdict in this parent case.

Figure 8

	Choosing to be a parent	Choosing to be child-free
Ex ante	5	5
Parent	5	5
Child-free	5	2

Since, the nudgee is indifferent both ex ante and ex post as a parent, nudging someone to be a parent cannot be permissible, even though the end-satisfaction in the parent-future is higher than in the child-free future.

Note that it would not help to replace ‘better off’ with ‘at least as well off’ in the criterion.

- A nudge is permissible if and only if the nudgee would judge themselves *at least as well off* both ex ante and ex post as a result of the nudge.

For then we run into this problem.

Figure 9

	Choosing to be a parent	Choosing to be child-free
Ex ante	5	5
Parent	-5	-5
Child-free	5	5

Here the nudgee is indifferent between being a parent and being child-free both ex ante and ex post. However, nudging them to be a parent would be impermissible, since they would as a result value being parent negatively. In contrast, if they were not nudged, they would retain their original values and keep valuing being child-free positively.

Pettigrew suggests a different version of AJBT Ex Post/Ex Ante. He claims that the previous versions of the AJBT-criterion have all been formulated in terms of what he calls *local utilities*, the utilities that encodes the values one has at a certain time. He thinks it is a mistake to focus on only local utilities. Instead, we should focus on *global utilities*, the ones that one should use when one makes decisions. The global utility, at a time, of an outcome is a weighted average of all local utilities of the outcome, past, present, and future. A lot of factors decide which weights to use, but Pettigrew thinks we have a defeasible obligation to assign at least some weight to each local utility, past, present, and future. However, morally abhorrent local utilities can defeat this obligation. We also have a specific obligation to our past selves if they have made sacrifices for us. He also thinks that it is permissible for the current self to give *less* weight to the local utilities of future and past selves than to the local utility of one’s current self. Armed with these notions, we can now formulate Pettigrew’s AJBT-criterion (or at least one plausible interpretation of it):

- A nudge is permissible if and only if the global utility of the nudge would be greater both ex ante and ex post as a result of the nudge.

(To deal with indifference-cases a version of this criterion would replace ‘greater’ with ‘at least as good as’, but I will put them aside for the moment). How does this criterion work? Let us start with this simple case, where ‘Parent’ and ‘Child-free’ are used as names of the outcomes of the corresponding actions. I follow Pettigrew in assuming that it is permissible to give weight $\frac{3}{4}$ to one’s current local utility and $\frac{1}{4}$ to one’s future local utility.

Figure 10

	Choosing to be a parent	Choosing to be child-free
Ex ante	2	5
Parent	5	2
Child-free	2	5

Global utility ex ante of *Parent* and *Child-free* will be calculated in the following way.

Figure 11

	Choosing to be a parent	Choosing to be child-free
Ex ante local utility	2	5
Parent’s local utility	5	2
Child-free’s local utility	2	5
Ex ante global utility	$(2 \times \frac{3}{4}) + (5 \times \frac{1}{4}) = 1.5 + 1.25 = 2.75$	$(5 \times \frac{3}{4}) + (2 \times \frac{1}{4}) = 3.75 + 1.25 = 5$

So, the global utility ex ante of *Child-free* is greater than that of *Parent*.

Parent’s global utility of *Parent* and *Child-free* will be calculated in the following way.

Figure 12

	Choosing to be a parent	Choosing to be child-free
Ex ante local utility	2	5
Parent’s local utility	5	2
Child-free’s local utility	2	5
Parent’s global utility	$(5 \times \frac{3}{4}) + (2 \times \frac{1}{4}) = 3.75 + 0.5 = 4.25$	$(2 \times \frac{3}{4}) + (5 \times \frac{1}{4}) = 1.5 + 1.25 = 2.75$

So, the parent’s global utility of *Parent* is greater than that of *Child-free*.

Hence, the criterion does not allow the nudger to nudge the nudgee in the direction of becoming a parent, which seems to be the right verdict. But the criterion permits the nudgee to remain child-free, since the global utility of Child-free is greater than that of Parent, both ex ante, as shown above, and from the perspective of the nudgee if they remain child-free, as is shown below.

Figure 13

	Choosing to be a parent	Choosing to be child-free
Ex ante local utility	2	5
Parent's local utility	5	2
Child-free's local utility	2	5
Child-free's global utility	$(2 \times \frac{3}{4}) + (2 \times \frac{1}{4}) = 1.5 + 0.5 = 2$	$(5 \times \frac{3}{4}) + (5 \times \frac{1}{4}) = 3.75 + 1.25 = 5$

However, it is important to note that Pettigrew's criterion does not honour the ex ante judgment if as a parent you would assign *sufficiently high* value to being a parent. Here is an example.

Figure 14

	Choosing to be a parent	Choosing to be child-free
Ex ante	2	5
Parent	20	2
Child-free	2	5

Here the ex ante global utility of *Parent* would be greater than that of *Child-free*, as is shown below.

Figure 15

	Choosing to be a parent	Choosing to be child-free
Ex ante local utility	2	5
Parent's local utility	20	2
Child-free's local utility	2	5
Ex ante global utility	$(2 \times \frac{3}{4}) + (20 \times \frac{1}{4}) = 1.5 + 5 = 6.5$	$(5 \times \frac{3}{4}) + (5 \times \frac{1}{4}) = 3.75 + 1.25 = 5$

And the Parent's global utility of *Parent* would also be greater than that of *Child-free*, as shown below.

Figure 16

	Choosing to be a parent	Choosing to be child-free
Ex ante local utility	2	5
Parent's local utility	20	2
Child-free's local utility	2	5
Parent's global utility	$(20 \times \frac{3}{4}) + (2 \times \frac{1}{4}) = 15 + 0.5 = 15.5$	$(2 \times \frac{3}{4}) + (5 \times \frac{1}{4}) = 1.5 + 1 = 2.5$

So, Pettigrew's criterion does in fact allow the government to override your current values by nudging you to satisfy ends you do not have yet, knowing that, by doing this, they will change your ends as well. Shouldn't the means paternalist be horrified then?

Pettigrew would answer that the government satisfies the ends *set by your current global utility*. The problem is that the nudgee might not actually have these ends. True, you *would* have these ends, if you adopted a permissible set of weights and aggregated according to Pettigrew's method by summing the weighted local utilities. But not many people are likely to have done this, since both the criterion for permissible weights, e.g., obligations to give extra weight to past local utility if one's past self has made sacrifices for the present self, and the aggregation method are controversial. The former assumption hinges on Pettigrew's controversial metaphysical idea that persons are made up of distinct multiple selves and that when we make decisions we should think as if we are acting on behalf of all selves, past and future. To just sum the weighted local utilities allows that big sacrifices of current local utility can be compensated for by many small future local utility gains. It thus seems that Pettigrew will have to go for a version of *end* paternalism after all, if by 'end' we mean 'ex ante actual end'.⁴

Another noteworthy feature of Pettigrew's criterion is that it presupposes that it makes sense to *sum* weighted local utilities. Since we are not adding up differences, but levels, we actually need ratio scale measurability. None of the other criteria or their test cases presuppose that. They only require that it makes sense to compare values *ordinally*, and that we can distinguish between positive and negative values. The exacting measurability requirements are especially concerning if we want to use the AJBT-criterion in real policy-contexts. Rather complex calculations involving hard to get utility information are required before we can establish whether we are permitted to nudge someone.

⁴ This is conceded by Pettigrew (2023, 15-17).

5. A better proposal?

It would thus be welcome to have a version of the AJBT-criteria that honours means paternalism, only requires ordinal value comparisons, and avoids controversial metaphysical assumptions about persons. We also need not make sure that it does not focus only on comparative ex post judgments, for then we get into the problems of not paying attention to the degree of value one would assign to the life, if one were to live it. But here is a very simple suggestion that avoids the problems with Pettigrew's proposal.

A nudge is permissible if and only if either

- (a) the nudgee judges ex ante themselves to be at least as well off after the nudge, and
- (b) they would ascribe a greater value ex post to their nudge-life, if they were nudged, than they would ascribe to their alternative life, if they were not nudged

or

- (c) the nudgee judges ex ante themselves to be better off after the nudge, and
- (d) they would ascribe at least as high a value ex post to their nudge-life, if they were nudged, as they would ascribe to their alternative life, if they were not nudged

Conditions (a) and (c) guarantee that one is never allowed to go against the nudgee's ex ante comparative judgment. So, in this sense, one respects their actual ends, and means paternalism is complied with. Conditions (b) and (d) implies that a nudge is permissible only if that the nudge would result in at least as high a degree of future end-satisfaction. This means that my proposal is not an AJBT-criterion that is formulated exclusively in terms of what is *better*, or *at least as good*, as judged by the nudgee's themselves. Such comparative judgments play a role only in the conditions (a) and (c) which are about the ex ante perspective. In contrast, the conditions (b) and (d), which are about the ex post perspectives, are formulated in terms of comparisons between absolute judgments about how well or badly off one is. So, it is still 'as judged by themselves'-criterion, but one that also invokes comparisons *between* absolute value judgments. This criterion will give the right verdicts in the all the cases listed above (I leave it to the reader to work this out.)

Another virtue with this proposal is that it can take care of certain *epistemically* transformative cases. Suppose that you face a choice that would affect your ability to make comparative judgments. Ex ante you judge A to be better than B. But if you choose A, you will judge it to be good but not be able to fully grasp B and therefore

form no attitude towards B, and if you choose B you will judge it to be good but not be able to fully grasp A and therefore form no attitude towards A. Here is an abstract illustration, where the blanks signify that no attitude is formed.

Figure 17

	Choosing to be A	Choosing to be B
Ex ante	5	2
A	5	
B		2

A concrete illustration might be two possible futures for a soon to become dementia patient. As a result of her dementia, she will no longer be able to grasp both alternative future lives, but she will be able to grasp and assess the life she ends up living. Pettigrew's account would not be applicable here, since global utility is not well-defined. In contrast, my proposal will say that it is permissible to nudge the person to choose A, which seems right. After all, you both ex ante judge A as better than B, and have higher end-satisfaction in A than in B.

Finally, this criterion is not as demanding as Pettigrew's as regards the utility information needed and the calculations required. In fact, we only need ordinal information about the level of value assigned by the person and no aggregation.

Now, one might object that it does allow a nudger to change one's ends so that a greater future end-satisfaction is achieved. That is true, but the criterion does not allow that this is done at the expense of the current nudgee. In this sense, it provides a rather strong protection of *present tense* autonomy. Some might think it is too strong. Future end-satisfaction should sometimes be allowed to trump current autonomy. Perhaps, but then we definitely are leaving the means paternalist framework, which was the underlying motivation for the AJBT-criterion and Pettigrew's own solution. Perhaps that in the end we should do this. But that is another and much longer story.

A more important objection is that my account implies normative dilemmas, as the following abstract case illustrates.

Figure 18

	A	B
Ex ante	2	10
A	8	2
B	2	5

A-nudging is impermissible, since it is not in line with ex ante goals, so violates conditions (a) and (c). B-nudging is impermissible, since it minimizes future end-satisfaction, so violates conditions (b) and (d).

What to do? One possible reaction, which I have hinted at in earlier sections, is to deny the possibility of forced nudging; a nudger can always refrain from nudging. More exactly, if the options are A and B, the options for the nudger are (i) A-nudge, no B-nudge, (ii) no A-nudge, B-nudge, (iii) no A-nudge, no B-nudge. In the case above, the verdict from the revised account would now be that options (i) and (ii) are both impermissible, and option (iii) is the only permissible one. However, this move comes with a cost. It would deem all nudges impermissible whenever we have a case with the following structure:

Figure 19

	A	B
Ex ante	A	b
A	C	D
B	E	F

where $b > a$, $c > d$, $f > e$, and $c > f$. This may seem too restrictive, as the following instance suggests.

Figure 20

	A	B
Ex ante	2	10
A	15	10
B	2	10

My account would deem both the A-nudge and B-nudge as impermissible. But one might think that the B-nudge is permissible since B will satisfy the *continuous* ends of the nudgee.

Another option is to retract to sufficient conditions for permissibility and impermissibility.

A nudge is permissible *if*

- (a) the nudgee judges ex ante themselves to be at least as well off after the nudge, and
- (b) they would ascribe a greater value ex post to their nudge-life, if they were nudged, than they would ascribe to their alternative life, if they were not nudged

or

- (c) the nudgee judges ex ante themselves to be better off after the nudge, and
- (d) they would ascribe at least as high a value ex post to their nudge-life, if they were nudged, as they would ascribe to their alternative life, if they were not nudged

A nudge is impermissible *if*

- (a) the nudgee judges ex ante themselves to be at least as well off before the nudge, and
- (b) they would ascribe a lesser value ex post to their nudge-life, if they were nudged, than they would ascribe to their alternative life, if they were not nudged

and

- (c) the nudgee judges ex ante themselves to be worse off after the nudge, and
- (d) they would ascribe at least as low a value ex post to their nudge-life, if they were nudged, as they would ascribe to their alternative life, if they were not nudged

Of course, this would mean that the criterion would be silent in many cases, and that is obviously a cost. Note, however, that the criterion will not be silent in cases illustrated by Figure 6. Here it will say that nudging the person to choose A is permissible, which is the correct verdict.

The third main option would be to nevertheless find a way to capture all our intuitions about cases, use only ordinal information, and respect ex ante autonomy. I am not especially hopeful that this can be done. Indeed, we might have hit an impossibility theorem, which shows that our pet convictions together constitute an inconsistent set. Something will have to give. However, the sufficient conditions for permissibility and impermissibility above seem to be poor candidates for rejection. If so, we at least have a partial account, which is better than nothing.

References

- Paul, L. A., & Sunstein, C. (2019). ‘As judged by themselves’: Transformative experiences and endogenous preferences’, *Harvard Public Law Working Paper*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455421
- Pettigrew, R. (2023). ‘Nudging for changing selves’. *Synthese* 201:22.22
<https://doi.org/10.1007/s11229-022-04020-2>

Sunstein, C. R. (2018). Better off, as judged by themselves: A comment on evaluating nudges. *International Review of Economics*, 65, 1–8.