"As Judged By Themselves":
Transformative Experience and Testimony

Abstract

One way to evaluate various interventions in people's lives is to ask whether they make choosers better off, "as judged by themselves." This criterion can be understood to borrow from the liberal political tradition insofar as it makes the judgments of choosers authoritative. Giving ultimate authority to choosers might be taken to respect their judgments and to promote their welfare (insofar as people are uniquely situated to know whether choices make them better off). But for certain decisions, the "as judged by themselves" criterion is indeterminate. In these situations, involving life-changing, transformative experiences, the criterion does not offer a unique solution. It is possible that welfarist criteria will resolve the indeterminacy, despite serious questions about incommensurability. Considerations of process are relevant to judgments about choice-influencing interventions that promote transformative experiences.

Some choices are hard. Some of the hardest are those where each option takes you down a path that will change your life, and as you walk that path, it changes you into someone whose preferences are unrecognizable to you now. Our interest is in these kinds of choices, where each life-changing way you might choose is legally, socially, morally, and practically acceptable. How should we make these kinds of choices? It is plausible to think that what you choose should be influenced by factors standardly emphasized by models of rational choice, such as how happy each option could make you, or will generate the most life satisfaction, or, more precisely, which act will maximize your expected value, given everything that is relevant to a life well-lived. These things matter. But, as we'll see, it's not always easy to determine just how they matter.

Consider the choice of whether to have a child. Ella, faced with such a choice, is

a single woman in her thirties, with a thriving career, ambitious and proud of her accomplishments. She's unsure whether she wants to have a baby. She has the means for artificial insemination and to support a child, but agonizes over whether this irreversible life choice is the right one for her, since, right now, she cares much more about career success than parenting, and parenthood is likely to impede her career. After talking with her parents and friends, who strongly encourage her to try for a child, she goes ahead with it. It is an understatement to say that, after becoming a parent, she is happy. Her child becomes the most important thing in her life; to her surprise, she identifies herself, first and foremost, as a mother.

Or consider an example of emigration: Will was born and raised in the United States. He has long identified as American and worked for political change, but in recent years, he has started to question the direction of his country. To take a break from his worries, he decides to spend time in Norway. As he settles in, he forms close friendships; his loyalties and values begin to shift, and he finds himself full of admiration for Norwegians. After repeated urging by his friends, he finally moves to Oslo. Now he is very glad he's made the move: though he feels close to the nation in which he was born and raised, he can no longer imagine thinking of himself as American, or make sense of the way that the electorate decides on its president.

If Ella knew, before she chose to become a parent, that she would value parenting over her career, in what sense would it make sense for her to decide to have a baby? If Will knew that his choice to move to Norway would make him lose touch with the political pulse of the USA, why would he choose to leave his native home, the place to which he was so committed? More generally, on what grounds should one choose a path that changes, in deep and fundamental ways, what one cares about? On what grounds should a person choose to avoid such a path? Our goal here is to demonstrate that, for a certain class of such cases, a central criterion for post-choice evaluation, the "as judged by themselves" (AJBT) criterion, fails. For this class of cases, the fact that choosers deem themselves to be better off as the result of the choice they made—even if they are in fact better off as a result of the choice—does not imply that their choice was better than the alternative. We will have a few things to say about how to evaluate choices in such circumstances.

**

In recent decades, social scientists have learned a great deal about human behavior and in particular, about the role of "choice architecture" in affecting people's decisions.

Choice architecture refers to the background conditions against which people make choices. As an analogy, think of how the way a house is built affects the life of the people living there. An old house filled with nooks and corners, small rooms, and winding staircases encourages one style of living. A modern, open plan encourages another. Or consider the winding paths of a garden with landscaping designed to lead you towards a particularly propitious view.

Similarly, choice contexts can be designed to lead people towards particular options. Is the choice opt-in or is it opt-out? Choices might be greatly affected by the answer. Is a website designed to emphasize Option A, by making it visible and salient, and to downplay Option B? If so, few people might choose Option B. Is one option presented before another? If an item is placed first on a menu, consumers will be more likely to select it, merely because it is first. Designs can also exploit framing effects. For example, people tend to be especially averse to losses, disliking them far more than they like corresponding gains.[1] Whether a change counts as a loss or a gain may depend on how it is framed. More broadly, if people are informed of an existing social norm, they might well move in its direction, simply because it is a norm.

"Nudges" are understood as interventions that preserve freedom of choice while steering decisions in beneficial directions.[2] Consider a GPS device, which allows people to choose their own destination, and also to reject its directions, with the ultimate goal of helping people to get to where they want to go. The guiding idea is that background conditions can be constructed in ways that preserve meaningful freedom of choice while promoting good choices over bad ones. On this approach, policies might include provision of information or default rules, potentially with large effects on outcomes: for example, the GPS may steer more people towards the use of expressways, or an opt-out design may significantly increase participation rates for retirement savings plans..[3] Default rules, use of order effects, and particular ways of describing and framing

[1] Eyal Zamir, *Law, Psychology, and Morality: The Role of Loss Aversion* (Oxford: Oxford University Press, 2014); Daniel Kahneman and Amon Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47 (1979): 263-291; [redacted for anonymous review]).
[2] *See* Richard H. Thaler and Cass R. Sunstein, *Nudge* (New Haven: Yale University Press, 2008).
[3] Jon Jachimowicz et al., "When and why defaults influence decisions: a meta-analysis of default effects," *Behavioural Public Policy* 3(2) (2019): 159-186. doi:10.1017/bpp.2018.43

outcomes are all ways to nudge.[4] Just as you might purchase an open plan house to encourage your teenagers to hang out in a common family space, nudges frequently involve the use of choice architecture to encourage agents to make good decisions.

But how do we decide which decisions are good ones? Which way ought people to be nudged? A pervasive answer points to maximizing the well-being of choosers. But how to we measure that? Along with Richard Thaler, Cass Sunstein has argued that the principal criterion is whether people are made better off, *as judged by themselves* (AJBT).[5] The AJBT criterion, as we shall call it, asks whether those who have been nudged *ex ante* (before the choice) deem themselves to be better off *ex post* (after the choice) as a result. For example, Omar, newly hired, needs to make a decision about his retirement plan. The default option in bis plan has him set aside eight percent of his income each month unless he explicitly opts-out. Omar does not opt out. If, at every point over the succeeding years, Omar judges that he has been made better off as a result, then the nudge (the need for an explicit choice to opt-out), as evaluated using his *ex post* testimony, is a good one.

**

The AJBT criterion can be applied in a context of individual decision making, as when an agent is making a choice, and also from a more general policy perspective, as from the standpoint of private and public choice architects. Choosers might well ask whether enrollment in some plan, or a change of some kind, will make them better off, AJBT, in light of everything that matters to them, and policymakers might ask the same question for relevant populations.

In many cases, choosers can answer the relevant questions about their own well-being far better than policymakers. Even so, the ex post judgments of choosers cannot always be treated as authoritative. Choosers might believe that they are better off when they are actually worse off. To address that issue, we might want to know if some kind of bias or motivated reasoning has contributed to their belief; we might also have to answer hard questions about the contested idea of welfare. If we are concerned with welfare, as behavioral economists tend to be, the question is

[4] Gideon Keren, ed., *Perspectives on Framing* (London: Psychology Press, 2010).
[5] Thaler and Sunstein, *Nudge*.

whether choosers are in fact better off, not merely what they believe. For this reason, giving decisive authority to subjective judgments about well-being is too strong.[6]

Nonetheless, the AJBT criterion has significant pragmatic appeal, perhaps especially when developing policy on which direction to nudge, but also for the *ex ante* assessments of options. It provides prima facie evidence, via testimony, about the value of the nudge for the chooser. If people believe and say that they are better off, we seem to have testimonial evidence in favor of the thesis that the choice has, in fact, improved their welfare. For followers of John Stuart Mill, that reason might be very good indeed.[7] At the very least, then, the AJBT criterion is a useful heuristic, and when choosers are objectively correct about being better off, it would seem to provide guidance for policymaking and evaluation. In addition, the AJBT criterion can claim to draw on strands of a liberal tradition that emphasizes the importance of individual agency. If people believe that they are better off, we might think that there is a sense in which any nudge is consistent with a kind of freedom in choice that liberal policymakers wish to respect. (We will qualify this point below.) Finally, use of the AJBT criterion has significant pragmatic value in how it sharply constrains outsiders – in government and in the private sector -- by directing them to attend not to their own concerns, but to those of choosers.

Our principal goal here is to identify a problem for those who believe (as we do) that the AJBT criterion can be useful and important for the reasons we have specified. The problem is that in an important class of cases the criterion can fail to provide determinate or unique answers. The class of cases concerns ones in which our choices are *transformative*, in the sense that they alter a chooser's core values, and by extension, their preferences.[8] They change what the chooser cares about, and this change is significant enough to bring about a change in the chooser's understanding of themselves. Such changes bring about an endogenous change in the chooser's preferences, and this has implications for our understanding of their post-hoc judgments about the change.

---

[6] On various conceptions of welfare, see Matthew Adler, *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis* (Oxford: Oxford University Press, 2011).
[7] For a sustained objection, see Sarah Conly, *Against Autonomy* (Cambridge: Cambridge University Press, 2013).
[8] Another distinctive feature of these cases is that, ex ante, the chooser may not be able to forecast this value change or imagine themselves "into the shoes" of their changed future self. See L.A. Paul, *Transformative Experience* (Oxford: Oxford University Press, 2014); Edna Ullmann-Margalit, *Normal Rationality* (Oxford: Oxford University Press, 2017).

**

Transformative experiences involve large-scale changes in your values, leading to changes in who you are. Small changes in value correspond to relatively mundane matters – food preferences, films, clothing, laptops – and do not involve self-transformation. We focus, instead, on large-scale shifts in the nature of one's values, which lead to changes in one's core preferences, as in the cases of Ella and Will. Such large-scale changes are *epistemically transformative*, in the sense that they change what a person knows, and by extension, their values. But they are more than this. They are also *personally transformative*: the epistemic change is profound enough that the shift in the person's values is seismic, amounting to a replacement of some of one's "core" preferences. On an approach where one's core preferences define oneself, the shift can be thought of as a process where one's old self, defined by one's old preferences (and values), is replaced by one's new self, defined by one's new preferences (and values).[9]

In personal transformation, in an important sense, the person after the choice is no longer the same kind of person they were before the choice. To make this claim more precise, we can distinguish between selves and persons, and take a persisting person to be constructed from a series of temporally and causally successive selves, in sequence from birth to death. A self at a time, is defined by a person's first-person perspective, a psychological state grounded on their conscious beliefs, values, and fundamental preferences at that time. When a person's beliefs, values and fundamental preferences change enough, we can say their old self is replaced with a new self. Both selves are parts of the sequence that compose the person over time: the old self realizes the person at an earlier time and the new self realizes the person at a later time. We can think of a persisting person as being realized, over time, by a series of causally and psychologically connected selves. When someone's old self is replaced by an appropriately connected, yet very different new self, we may understand this as a change in the kind of person they are.

Now we have the structure in place needed to explain the way a choice can make a transformative change in who a person is. When a person makes a transformative choice, the choice alters their self-identity, at least in important respects, by altering some of their fundamental values. As a result of this change in values, the person's *ex ante* self (the self

[9] One's values ordinarily determine one's preferences: if I value A over B, and I am rational, I prefer A to B.

making the choice) is replaced by their *ex post* self (the self resulting from the choice), changing the kind of person they are. So a transformative change in a person is a radical change in the self that realizes that person, changing who they are across the temporal expanse of the choice process, through a radical change in their values. (A change in values creates a corresponding change in preferences.) As such, transformative choices are choices to change oneself in an especially far-reaching way.

Choices that bring about transformative change have several interesting features. One important element involves the epistemic change involved: what happens when a person is unable to assess, *ex ante*, the kind of value change they are signing themselves up for? The person who is making the choice may need to undergo the transformation before they can appreciate the value—or disvalue—of who they might become. However, our main focus here is on another property of transformative choices: the endogeneity of the self-transformation involved. That is, not only are these choices transformative, but they are transformative in a particular way: when these choices change who a person is, *they change the person's preferences in a way that depends on that very choice*.[10]

Recall: the transformative choice amounts to a choice by an agent to replace their *ex ante* self with an *ex post* self. When the *ex post* state of the chooser (their self, or core self-defining preferences) depends on the act that the chooser performs, the state of the agent depends on the act of that agent. This violates the independence of the *act* from the *state*. Act-state independence is often presupposed in standard treatments of rational action. And the trouble is, if this presupposition is violated, norms with regard to the interpretation of testimonial evidence concerning the satisfaction of one's preferences are also called into question.

**

To get clearer on how transformative endogeneity creates a problem for the AJBT criterion, we'll start by looking at a simple example. Consider a choice involving an *ex post* evaluative judgment that is endogenous to a nudge: Raul has a serious illness. The question is whether he

[10] For additional discussion of how this interacts with experimental social-scientific research and methodological questions involving causal mechanisms, counterfactual dependence, and the fundamental identification problem, see L.A. Paul and Kieran Healy, "Transformative Treatments," *Noûs* 52 (2016): 320–335.

should have an operation, which carries with it with potential benefits and potential risks. Reading about the operation online, Raul is not sure whether he should go ahead with it.

When consulted, Raul's doctor frames the options in such a way as to persuade him to have the operation, emphasizing how much he has to lose if he does not. Raul decides to follow the advice, and a year later, he is glad he did. In a different possible world (a parallel world just like ours right up until Raul consults his doctor for advice, but relevantly different thereafter), Raul's doctor frames the options in such a way as to persuade him not to have the operation, emphasizing how much he has to lose if he does. In this world, Raul also decides to follow his doctor's advice, and a year later, he is glad he did. [11] In each case he starts out being unsure, that is, he does not have a clear preference about whether or not he should undergo the operation, but after he chooses, his *ex post* preferences (to have had the operation, or to have skipped the operation) are satisfied.

In this kind of case, the AJBT criterion can be satisfied by nudges in (at least) two different directions. The endogeneity here is that Raul's preferences are determined by the choice he ends up making. In each version of the case, Raul starts out with indeterminate preferences, makes a choice, and then forms preferences as a result of his choice. And in each case, Raul is happy that he followed the doctor's advice. Crucially, it also seems to be the case that post-hoc, in each case, Raul has in fact satisfied his preferences.

Endogeneity, formally, in this context, simply means that there is a causal connection between the participants' choices and actions, after being nudged, and their post-hoc preference satisfaction. A phenomenon of evident interest is the process by which the preference satisfaction was created. In endogenous cases like these, values and preferences are an artifact of the nudge, and thus satisfaction of these values and preferences is also an artifact of the nudge.

Endogeneity like this raises a question: is nudging the right thing to do? If so, in which direction? There are two features here that, together, raise our question. First, we can fairly say that in these cases, a person might be keenly interested in knowing whether they will believe themselves to be better off, *ex post*; but they will consider themselves to be better off *no matter*

[11] For related discussion, see Harman, Elizabeth (2009). "'I'll Be Glad I Did It' Reasoning and the Significance of Future Desires." *Philosophical Perspectives* 23: 177–199.

what, i.e., no matter which choice they make. But second, they consider themselves "better off no matter what" at least partly because of the endogeneity of the process.

The example brings out how, in such cases, it is not the particular sequence of events that determines preference satisfaction: rather, it is the way the person's preferences evolved in response to the choice they made. Such cases abound, and not just with cases with a forced change. There are countless variations on the story of Raul, where he chose or was nudged to stick with the status quo and would be glad to have chosen or to have been nudged that way, and where he would also be glad to have chosen to to have been nudged to depart from the status quo. In these kinds of cases, where there is no clearly best option, the AJBT criterion is indeterminate: we cannot use it to decide which nudge is better.

**

The problem for the AJBT criterion, at root, stems from the way a person's preferences can evolve in response to the choice they make. Although nudges are small, they can have big effects, and in some contexts, these effects can be transformative. This section develops the metaphysical structure of transformative preference evolution and its implications.

Recall that when a person transforms themselves, their preferences undergo a seismic change. As we put it above, a transformative choice amounts to a choice by an agent to replace their *ex ante* self with a new self: an *ex post* self.

Return to our example of Ella. Imagine Ella, like Raul, could be nudged in two different directions, with very different results. For example, perhaps Ella, now pregnant, is still unsure about what to do, and considers whether to have an abortion. The first way she could be nudged is in the direction of becoming a parent. If she becomes a parent, when interviewed a year later, she will say that she is very happy with her choice. Thus, as judged by herself, we should conclude that, *ex post*, she would be better off. However, she could also be nudged in the opposite direction, where, partly as a result of her choice to have an abortion, she forms a strong preference to live a child-free life. When interviewed a year later, she says that while she was sad to have had to make such a choice, she is very happy with what she decided. Thus, as judged by herself, we should also conclude that, *ex post*, she is better off.

In such a case, Ella could have decided or been nudged each way, and again, for each way, at the time of (post hoc) assessment, she would be glad that she'd chosen that way.

Moreover, at the time of (post hoc) assessment, she would find the alternative outcome to be truly abhorrent. Finally, the case demonstrates the same kind of endogeneity as Raul's. Given that each outcome of this endogenous choice results in post hoc satisfaction, how are we to compare the possible outcomes in order to discover which was better?

The endogeneity arises because transformative choices "satisfy a preference" only by replacing a person's self (and thus replacing their preferences) in such a way as to create the satisfied individual that they end up becoming. The problem, at root, stems from the way that transformative change violates the assumption of act/state independence, as the choice entails that a person's ex post self replaces their ex ante self. The assumption of independence is important in several contexts, and one of those contexts involves utility comparisons.

Normally, when making utility comparisons over changes of state, the agent is kept fixed, in order to meaningfully assess their utility in the new state. If we want to use the AJBT criterion to assess and compare A in state *a* to A in state *b*, it must be the case that the testimony from "A" in each state comes from the same agent. We use A's testimony to assess A in state *a* at time t1, make a change, evolve the world forward to state b at time t2, and then use A's testimony to assess A in state b at time t2, and compare our assessments.

For example, recall the way we would use the AJBT in an ordinary case of nudging, such as with Omar, who chose to save for his retirement. To assess the value of the nudge using the AJBT in a study of the effects of the nudge on retirees, we would assess Omar's satisfaction, based on his testimony, with his state (*a*) before he made his retirement savings choice, and then assess his satisfaction with his new state (*b*) based on his testimony after that choice. To do this in a meaningful way, that is, for the truth of "Omar makes his choice at t1 and he is satisfied with his choice at t2", to have the intended meaning, "Omar" must pick out the same person before and after the choice, that is, at t1 and at t2.

However, in a case where we replace A with a different agent (B) as the world evolves forward into state b, we cannot not simply assume that any change in testimony we observe is meaningful in the intended sense.[12] If, when we say "Omar makes his choice at t1 and he is satisfied with his choice at t2" it is actually the case that our use of "he" picks out the testimony

[12] For related methodological discussion, see L.A. Paul and Kieran Healy (2018), "Transformative Treatments," *Noûs* 52, 320-335 and L.A. Paul and John Quiggin, "Transformative Education," *Educational Theory* 70 (2020): 561-579.

of *somebody else*, e.g., some other Omar, then the truth of the statement "Omar makes his choice at t1 and he is satisfied with his choice at t2" means something very different. (Imagine, when following up with study participants thirty years later, that due to a records mixup, we interview the wrong person--someone else, from a different study, also named "Omar".)

In transformative contexts, we face a variant of this problem. Again, recall that a transformative decision changes the nature of the person transformed, replacing the ex ante self (the self that chooses at t1) with the ex post self (the self at t2 that results from the choice). This reflects the fact that there is a relationship between the change in the state of the agent (which is what we want to evaluate) and the way that change in state entails a change in who the agent is.

Returning to Ella, if she chooses to have the child, the self that she becomes is different from the self that made the choice. In our story, if Ella decides to have a child, when interviewed a year later, she tells us that she is very happy with her choice. That is "Ella makes her choice at t1 and she is satisfied with her choice at t2" is true. The trouble is that we cannot simply interpret this in the simple way we interpreted Omar's case: we cannot simply assume that Ella's testimony as a parent is comparable to Ella's testimony when she made the choice, because the self that we refer to as "Ella" is different at each time.

Similarly, if Ella decides to have an abortion, and when interviewed a year later, she has become a vociferous anti-natalist, she may tell us that she is very happy with her choice. That is "Ella makes her choice at t1 and she is satisfied with her choice at t2" is also true in this world. Again, however, we cannot simply interpret this in the way we interpreted Omar's case: we cannot simply assume that Ella's testimony as an anti-natalist is comparable to Ella's testimony when she made the choice. Why not? Because the self that we refer to as "Ella" at t2 is not the same self as the self we refer to as "Ella" at t1.

A way to see the complexity involved here may be to reflect on your own life. If you have children, reflect on who you were before you became a parent, and imagine (if you can!) what you would have been like now, if you'd stayed childfree. Are you better off as the result of the choice you made? To decide this, you must compare yourself as you are now with who were before you chose. We may think of this in terms of comparing the testimony from your different possible selves. If you, as you are now, are to compare yourself in a meaningful way with your past self, what you care about most, how you spend your time, who you spend your time with, and so on, has almost certainly changed dramatically from your pre-parent life, and would be

even more radically divergent with further childfree years. Who you are has changed quite dramatically, and what it means to have your preferences satisfied has changed at least as much. That is, what it means for you to testify to the fact that your preferences have been satisfied depends on which self you have become.

Our point is that in transformative cases, such an approach is more like comparing the testimony of different people than measuring a change in testimony across the same person's change of state. In these kinds of cases, if there is no clearly best option to choose (because all outcomes lead to more or less equally satisfied individuals), it is unclear how to apply the AJBT criterion in order to assess the counterfactuals in question. In such cases, if we are considering nudging one way or another, we cannot use the AJBT criterion to decide which nudge is better.

The problem is especially salient given that transformative choices involve replacement of one's preferences in a context where, before the choice is made, the nature of the transformative experience is, in some relevant sense, opaque to the agent making the choice. If the choice people can make determines their post hoc preferences about that choice in a way that is disconnected from their own *ex ante* preferences and disconnected from their previous experiences, how should they make their choice? In particular, before they choose, how should they assess and weigh post-hoc testimony from others in order to apply it to their own possibility for self-transformation, if such testimony is endogenously generated from the transformative experience in question?[13]

Here, then, is the root problem that transformative choices raise for the AJBT criterion. Such choices, leading to large-scale alterations in people's lives, can result in endogenous preference change, where this change is tied to a new self that is generated by the process of the change. This new self may have preferences that are very different from the self that made the choice, and may also have preferences that are very different from any alternative self that could have resulted from a different choice. This raises questions about the AJBT criterion. If our assessment of the value of the change involved is merely that, as judged by themselves, people's ex post selves will be happy, and glad to have ended up as they did, this is not sufficient to distinguish between alternative outcomes for the chooser.[14]

---

[13] Vladimir Chituc, L.A. Paul, and M.J. Crockett, "Evaluating Transformative Decisions," *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43 (2021), and [redacted for anonymous review].
[14] Harman 2009.

We conclude that we need a further criterion. As a general rule, the AJBT criterion is a useful test for evaluation of a nudge, but even when that criterion has been satisfied, we will not know in which direction a person should be nudged. For that reason, it is insufficient.

**

We have argued that the AJBT is not sufficient for welfare determination, and thus not sufficient, on its own, for approval of a nudge. In this section, we will explore the implications of this conclusion.

As we pointed out above, cases of transformative choice involve endogeneity of a particular kind. In these cases, the chooser's *very self* is determined by the choice that was made. In a certain sense, when you make a transformative choice, your choice is making you just as much as you are making your choice. A transformative choice isn't just a choice about happiness, or about preference satisfaction. It's a choice about what kind of life you find most appealing, and by extension, about what kind of person you want to be.

Return to the case of Ella, as she considers parenthood. Should she become a parent? It is not clear that the solution is not to have others tell her what to do, or for her to choose merely based on what will make her happiest. In fact, choosing to become a parent may well involve much more suffering than choosing to remain child free. She might be not focus only on what will make her happiest in some narrow sense.

> "A life that seems to be aimed at something of genuine value and importance can at times generate deep satisfaction, but it also can and typically does present frustrations and obstacles that call forth great exertions; it can require great personal sacrifice; it can and often does produce great regrets; and, in many cases, it includes great suffering. The sense of value and importance of a life does not typically make those experiences pleasant or satisfying; it makes their being unpleasant or unsatisfying seem less significant." [15]

[15] William Talbott (2016), "Critical Notice: Transformative Experience", Analysis Reviews, p. 6.

In other words, the choice to become a parent might well not be about what would make you happy. It's about what kind of future you want to have. In the end, if Ella chooses parenthood, she might choose a kind of suffering, but suffering that has a particularly meaningful sense.

Now, this is consistent with the AJBT. As our earlier examples suggested, in many cases, we can imagine someone like Ella testifying that she is better off as a parent. It is also easy to imagine someone like Ella testifying that she is *not* better off as a parent in any simple hedonic way, and yet, she would not choose any other outcome. In fact, this latter result arises in many complex, real world cases: by many standard metrics, parents testify to being worse off, even if they would not have chosen otherwise.[16]

Similar kinds of situations arise in cases involving physical disability, or for choices where someone chooses to devote their lives, and thus themselves, to an important cause despite the likelihood of its entailing significant suffering and loss. In these situations, a person is worse off along many standard metrics, they do not claim to be better off in any hedonic way, yet they value the life that they actually have over the counterfactual life they do not. And note: this is true despite the fact that their counterfactual selves would also testify that they value the life that they actually have over the counterfactual—i.e., counterfactual from their point of view—life they do not have.

The *welfarist* will argue that: to compare outcomes, we need to determine a person's welfare in each possible state. We need to construct an assessment of a person's welfare in each possible outcome (each world-state), where such an assessment accommodates the sense of meaningfulness, or other less tangible qualities of well-being, that accompanies some of these states.[17] Once we have properly assessed a person's welfare in each possible state, then we can compare them. If, in making this comparison, we find that one world-state involves a higher degree of well-being than the other, nudging towards that state might seem merited. A similar approach could work for a situation where an absence of nudging would lead to a state that had lower well-being.

---

[16] Bloom, P. (2021). *The Sweet Spot: The Pleasures of Suffering and the Search for Meaning*. New York: Ecco, Nelson, S. K., K. Kushlev and S. Lyubomirsky (2014). "The Pains and Pleasures of Parenting: When, Why, and How is Parenthood Associated with More or Less Well- Being?" *Psychological Bulletin*, 140 (3): 846–95, and [redacted for anonymous review].

[17] Bykvist, K. (2022). "Wellbeing and Changing Attitudes Across Time," *Ethical Theory and Moral Practice* (79), https://doi.org/10.1007/s10677-022-10311, Rosati, C. S. (2009). "Self-Interest and Self-Sacrifice," *Proceedings of the Aristotelian Socie*ty 109, 311-325, and [redacted for anonymous review].

Thus, we see the importance of using other criteria, in addition to the AJBT criterion, for assessments in a range of contexts, including some that involve nudging.

**

We have argued that in some contexts, the AJBT criterion is insufficient for the evaluation of nudges, because the interpretation of testimony across transformative, life-changing outcomes requires a different approach from the interpretation of testimony in more standard contexts.

For the committed welfarist, there are two paths forward. Both of them raise complex issues, and we merely flag them here. First, we ask: What choice, and what nudge, really makes people better off? To answer such questions, we need a specification of the right conception of welfarism. Suppose that we place an emphasis on people's subjective experience and assume that we could measure it, or at least come pretty close.[18] A measure of experience might pay attention to subjective happiness; alternatively, it could attend to a sense of purpose or meaning, which might also be measurable.[19]

A challenge is that for transformative experiences, there might be commensurability problems, making it difficult to deal with cases such as those of Ella. People in their situation might endorse a conception of welfare at time t1 that is very different from their conception of welfare at time t2. Does one conception prevail? How are we to prospectively compare the different ways, given the different possible outcomes of their different possible choices, that their welfare could be assessed? Are outsiders permitted to choose between them, or to reject both? Given an agreed-upon conception of welfare, the normative consensus might be sufficient, and if subjective measures are what matter, empirical tools might be able to help make relevant measurements. But in the cases we have in mind, an agreed-upon conception of welfare is difficult to identify, and the measurement issue is daunting.

Second, and as signaled earlier: It may be important to ask about the *process* by which people's preferences are formed. At one pole are cases in which people freely choose to undergo a transformative experience (or otherwise to make a choice that alters their values and preferences). Let us stipulate that no objectionable outside influence is involved (acknowledging that the stipulation raises many questions). If so, there should be no process concern. At another

[18] Paul Dolan, *Happiness By Design* (New York City: Plume, 2015).
[19] Ibid.

pole are cases in which people are coerced into a transformative experience (or otherwise to a situation that alters their values and preferences). A case of coercion might involve a kidnapping; consider "Stockholm Syndrome." Or it might involve an effective mandate or ban. As noted, we might want to adopt a rule or at least a presumption, to the effect that the satisfying the AJBT criterion is no excuse or justification for coercion.

These questions bring out how we need to attend to the process by which the effect was brought about (the means to the ends). If kidnap victims end up admiring their captors and thus being satisfied with their captivity, it is not at all clear (to say the least) that the AJBT criterion captures what matters. We might want to say that the problem here is coercion. If so, we might limit the AJBT criterion to choice-preserving interventions and find it inadequate when coercion is involved.

But the concern about the relevant process might cut more broadly. To take an admittedly extreme case, post-hoc satisfaction of participants might not be decisive if we discover that some people were nudged to choose frontal lobotomies and became highly satisfied merely because of their reduced capacities.[20] Or to take a less extreme case, what about an AI, trained to maximize our past preferences and the preferences of people who are like us in certain ways, that nudges us towards one way of being over another?[21] In such cases, is the problem one involving an objectionable kind of manipulation, or does it involve welfare, rightly understood?

If we are welfarists, we would not have a rigid rule against use of the AJBT criterion, even in cases of coercion. The thought here is that the AJBT captures something of value: post hoc approval. Our discussion of the endogeneity of transformative choice brings out, however, that post-hoc approval is not enough. We need additional criteria in order to be sure that the nudge is justified, one that accommodates the concerns we have raised about process, and one that recognizes the limitations of the AJBT criterion given the possibility of endogeneity.

The hope is that a focus on welfare will provide the needed correction. If third parties are engaging in coercion rather than (say) nudging, we might think that it is most unlikely, in the general run of cases, that those who are coerced will be better off. Moreover, if nudging is based on inadequate data, especially if that data stems from a failure of comparability such that we

[20] Elizabeth Barnes, "What You Can Expect When You Don't Want to be Expecting," *Philosophy and Phenomenological Research* 91(3) (2015): 775-786.
[21] [redacted for anonymous review]

cannot make a legitimate comparison of welfare, we should not endorse the nudge. And yet, we should attach value to an individual's testimony that they are better off as the result of their choice and take that into account for both individual and policy-level decision making.

It should be clear that this claim builds on the view, associated with John Stuart Mill, that individuals are in a unique position to know what will improve their welfare, and that outsiders will often blunder. Mill insists that the individual "is the person most interested in his own well-being," and the "ordinary man or woman has means of knowledge immeasurably surpassing those that can be possessed by anyone else."[22] When society seeks to overrule the individual's judgment, it does so on the basis of "general presumptions," and these "may be altogether wrong, and even if right, are as likely as not to be misapplied to individual cases." If Mill is even broadly correct, a rule or presumption against coercion or against inadequately grounded nudging is justified on welfarist grounds.

[22] John Stuart Mill, *On Liberty* (New York: Dover, 2002), 8.