# "As Judged By Themselves": Transformative Experiences and Endogenous Preferences

L.A. Paul[*] and Cass R. Sunstein[**]

**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 autonomy and also 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 such cases, what people care about shifts, depending on their choice. Some choices change people's preferences and values, and in this sense, change their identity. In these situations, often 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 autonomy are also relevant to 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 will change you. Our interest is in these kinds of choices, where each option, each life-changing way you might choose, is legally, socially, morally, and practically acceptable. How should we make these kinds of choices? Obviously, what you choose should be determined by rational choice factors, such as by how happy each option could make you, or will generate the most life satisfaction, or, more precisely, which act will maximize your expected value. These things matter. But, as we'll see, it's not always so simple.

Consider the choice of whether to have a child. Edith, 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, especially since it could 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

[*] Professor of Philosophy and Cognitive Science, Yale University.
[**] Robert Walmsley University Professor, Harvard University.

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: Frank was born and raised in the United States. He long identified as American. In recent years, he has started to question the direction of his country. He decides to spend time in Norway, where he forms close friendships. His loyalties and values begin to change. He is full of admiration for Norwegians. After repeated urging by his friends, he moves to Oslo, and he is glad that did. Though he feels close to the nation in which he was born and raised, he can no longer imagine thinking of himself as American.

We will argue that, for a certain class of cases with features like these, there is a serious puzzle about whether there is an objectively correct choice to make. One reason for the puzzle is that a central criterion for post-choice evaluation, the "as judged by themselves" 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 the better one.

**

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. People are especially averse to losses; they dislike losses far more than they like corresponding gains.[1] However, whether a change counts as a loss or a gain may depend on how it is framed. The appeal of a choice option can also be affected by framing. If an item is placed first on a menu, consumers will be more likely to select it. If people are informed of the existing social norm, they might well move in its direction.

An understanding of these findings has sparked considerable interest in "nudges," understood as interventions that are intended to preserve freedom of choice while steering decisions in beneficial directions.[2] The guiding idea is that the rules and conventions of the social world can be constructed in ways that preserve freedom of choice while promoting good choices over bad ones. Nudges involve the use of a choice architecture, with defaults set up to encourage better end results. The default rule matters a great deal: for example, with an opt-in design, participation rates are generally far lower than with an opt-out design.[3] Default rules, use of order effects, and particular ways of describing and framing outcomes are all ways of employing nudging.[4]

[1] See Eyal Zamir, Law, Psychology, and Morality: The Role of Loss Aversion (2014).
[2] See Richard H. Thaler and Cass R. Sunstein, Nudge (2008).
[3] See Jon Jachimowicz et al., When and why defaults influence decisions: a meta-analysis of default effects, Behavioural Public Policy (2019), available at https://www.cambridge.org/core/journals/behavioural-public-policy/article/when-and-why-defaults-influence-decisions-a-metaanalysis-of-default-effects/67AF6972CFB52698A60B6BD94B70C2C0
[4] See Perspectives on Framing (Gideon Keren ed. 2010).

But how do we decide whether people have been nudged in a good direction? Which way ought they to be nudged? Along with Richard Thaler, one of us has argued that the principal question 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) – for example, with a warning or a reminder – deem themselves to be better off *ex post* (after the choice) as a result. We can think of the AJBT criterion as applicable from the first-person perspective, when people are making choices, and also from the standpoint of private and public choice architects, deciding in which direction to nudge.

Nudging uses choice architecture to encourage good results, and can be evaluated using ex post testimony. For example: Thomas is overweight. He goes to restaurants a great deal, and he chooses what he thinks he will most enjoy. He would like to lose weight, but he has failed to do so. Because of a new law, he now sees calorie counts at most restaurants. He is surprised to learn that some of his usual choices are very high in calories. He makes different choices, and he is losing weight. He is surprised and delighted.

Use of the AJBT criterion has considerable appeal in this kind of case. The fact that Thomas is glad to have been nudged, ex post, seems to suggest that he is better off. Of course, the ex post judgments of choosers should not always be treated as authoritative. Choosers might believe that they are better off when they are actually worse off. If we are concerned with welfare, the question is whether they are right, not what they think. It's about what's objectively correct for the chooser, not merely subjectively believed.

We think giving decisive authority to subjective judgments about well-being is too strong.[6] However, the AJBT criterion has significant pragmatic appeal, and, at least prima facie, seems to provide testimonial evidence about the value of the nudge for the chooser. With this in mind, let us simply note two points. First, if people believe that they are better off, we have some reason to think that the nudge has, in fact, improved their welfare. For followers of John Stuart Mill, that reason might be very good indeed.[7] This makes the AJBT criterion a useful heuristic. Second, the AJBT criterion can claim to draw on those strands of the liberal tradition that emphasize the importance of individual agency. If people believe that they are better off, we might think that the nudge has respected their autonomy. (We will qualify this point below.) At the very least, 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 criterion is often useful. The problem is that, for an important class of cases, the AJBT criterion can fail to apply for interesting, non-subjective reasons. The important 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

---

[5] Id.

[6] On various conceptions of welfare, see Matthew Adler, Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis (2011).

[7] For a sustained objection, see Sarah Conly, Against Autonomy (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 (2014); Edna Ullmann-Margalit, Normal Rationality (2017).

this change is significant enough to bring about a change in the chooser's understanding of themselves.

**

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 self, as in the cases of Edith and Frank. Such large-scale changes are *epistemically transformative*, in the sense that they change a person's beliefs, or desires, or capacities. But they are more than this. They are also *personally transformative*: the epistemic change is profound, bringing about a seismic shift in a person's values and preferences. The shift can be thought of as a process where one's old self, defined by one's old values and preferences, is replaced by one's new self, defined by one's new values and preferences.

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 mental representation grounded on their conscious beliefs, values, and preferences at that time. When a person's beliefs, values and 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.

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 effects of that choice alters their self-identity, at least in important respects, by altering their core values. As a result of this change in values, the person's ex ante self (the self making the choice) is replaced by an ex post self (the self resulting from the choice). So a transformative change in a person is a radical change in the self that realizes her, changing who she is across the temporal expanse of the choice process. As such, they are choices to construct or determine oneself.

Choices that bring about transformative change have several interesting features, but our focus here is on their endogeneity: when these choices change who a person is, they change the person's preferences *in a way that depends on those choices themselves*.[9]

[9] 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 Paul and Healy (2016), "Transformative Treatments." *Noûs* 52: 320–335. https://doi.org/10.1111/nous.12180

**

To get clearer on the way endogeneity creates a problem, we'll start by looking at a simple example. Consider a non-transformative choice involving an *ex post* evaluative judgment that is endogenous to the nudge: 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 different possible world (a parallel world just like ours right up until Rick consults his doctor for advice, but relevantly different thereafter), Rick's doctor advises him not to have the operation, emphasizing how much he has to lose if he does. In this world, Rick also decides to follow his doctor's advice, and a year later, he is glad he did.

In this kind of case, the AJBT criterion can be satisfied with nudges in (at least) two different directions. In each version of the case, Rick makes a choice, whether to have the operation or not. And in each case, Rick is happy that he followed the doctor's advice. The endogeneity here is that Rick's *ex post* preferences (to have had the operation, to have skipped the operation) are determined by his choice. 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.

Endogeneity like this raises questions. In particular, is nudging the right thing to do? If so, in which direction? We can fairly say that in these cases, a person might be 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 what. And in particular, they consider themselves better off no matter what because of the endogeneity of the process.

In such a case, the AJTB criterion captures one element of a desirable choice architecture: post hoc satisfaction. When participants report they are satisfied with their (choice) results, this supports (without establishing) the claim that there have been positive improvements in welfare, as judged at the time of (post hoc) assessment. But further issues are in play. One question involves a comparison: would the chooser have been more satisfied with a different outcome? And how are we to compare the possible outcomes? Even more importantly, however, what is the ground for the chooser's satisfaction? It is not merely that the chooser's antecedent preferences were satisfied. Something else seems to be involved. But what is that? And is it normatively decisive?

**

As we pointed out above, cases of transformative choices involve a particular kind of endogeneity. In such cases, it isn't merely that the chooser's *ex post* preference is determined by the choice they make. The chooser's *very self* is determined by the choice they make. That is, because the choice is transformative, revising core, self-defining preferences, the chooser

replaces their *ex ante* self with an *ex post* self. In these kinds of cases, people's values and preferences shift as a result of their choices in a way that violates the independence of the *act* from the *state*. The *ex post* state of the chooser depends on the act that the chooser performs.

The cases of Edith and Frank, as described above, involve this sort of transformative self-change. Each chooser is satisfied with the way events developed, and so, as judged by themselves, we may conclude that, *ex post*, they are better off. Note that it is not the particular sequence of events that matters here: rather, it is the way the person's preferences evolved in response to the choice they made. Crucially, there are variations on the stories of Edith and Frank where each chose or was nudged to stick with the status quo, and *each would be glad to have chosen or to have been nudged that way*.

Consider Edith: she chose to have a child, and through this process, she formed a strong preference to have her baby. If she had chosen not to have a child, she could have formed an equally strong preference to live a child-free life, perhaps as the result of follow-up choices she made. She could have been nudged differently, she would be glad to have been nudged that way, *and she would find the alternative outcome to be truly abhorrent*.

Such cases raise several questions. First, what are the implications for the AJBT criterion? For example, given how easily each variation could be brought about, what should we conclude from the fact that, for Edith, the AJBT criterion was satisfied? More broadly, given the endogeneity of transformative choices, how should we interpret the testimony of those who have been transformed by such choices? What is the explanation for why they are *ex post* satisfied with their choice? How much weight should that testimony receive, and what inferences should we draw from it?[10]

Moreover, if the choice people could make would determine their post hoc preferences about that choice in a way that is disconnected from their *ex ante* preferences, how should they weigh post-hoc testimony from others?

Here, then, is the root of the problem that transformative choices raise for the AJBT criterion. Important choices, leading to large-scale alterations in people's lives, can result in endogenous preference change. *If our assessment of the value of such changes is merely that, as judged by themselves, people will be happy ex post, and glad to have ended up as they did, this is not sufficient to distinguish between alternatives. If, for each change we consider, as judged by themselves, people will be happy ex post, we need a further criterion.* Perhaps, as a general rule, the AJBT criterion is (usually) a necessary condition for approval of a nudge, but when that criterion has been satisfied, we will not know in which direction they should be nudged. For that reason, it is insufficient.

**

A transformative choice isn't just a choice about happiness. It's a choice about what kind of person you want to be. This suggests that, for such choices, the AJBT criterion may not be necessary.

[10] For related methodological discussion, see Paul and Healy (2018), and Paul and Quiggin (2018).

In *On Freedom*, reminding us of Huxley's *Brave New World*, one of us (Sunstein) pointed out that sometimes we prefer to choose unhappiness in exchange for the freedom to remain ourselves.

> '"All right then," said the Savage defiantly, "I'm claiming the right to be unhappy."
>
> "Not to mention the right to grow old and ugly and impotent; the right to have syphilis and cancer; the right to have too little to eat, the right to be lousy; the right to live in constant apprehension of what may happen tomorrow; the right to catch typhoid; the right to be tortured by unspeakable pains of every kind."
>
> There was a long silence.
> "I claim them all," said the Savage at last.'[11]

If you choose not to transform, like the Savage, and this involves choosing a path that will bring suffering and loss, you do so because you prefer not to be that kind of self. If you are choosing knowledgably and authentically, you choose knowing that the deep structure of the choice concerns the kind of self you want to become. The choice isn't merely about what would make you happy. It's about which future self you want to be.

Return to the case of Edith, as she considers parenthood. Should she become a parent? The solution is not to have others tell her what to do, or for her to choose merely based on what psychologists (or her mother) tells her will make her happy. In fact, choosing to become a mother may well involve much more suffering than choosing to remain child free. Rather, Edith needs to choose based on what kind of self, *ex post*, she wants to be. If she chooses parenthood, then, like the Savage, she may be claiming the right to be unhappy, but in a particularly meaningful 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." [12]

[11] Aldous Huxley (1932) *Brave New World*, Chatto and Windus.
[12] William Talbott (2016), "Critical Notice: Transformative Experience", Analysis Reviews, p. 6

**

How are we to regard the situation? For transformative experience, as well as for more mundane cases of endogenous preference change, there are two paths forward. Both of them raise complex issues, and we merely flag them here.

First: Committed welfarists would ask: What choice, and what nudge, really makes people better off? To answer such questions, we would need to specify the right conception of welfarism. Suppose that we place an emphasis on people's subjective experience. It is possible that we could measure that, or at least come pretty close.[13] 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.[14] A challenge is that for transformative experiences, there might be commensurability problems, making it difficult to deal with cases such as those of Edith and Frank. People in their situation might endorse a conception of welfare at Time 1 that is very different from their conception of welfare at Time 2. 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 would be difficult to identify, and the measurement issue would be 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 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 can be taken to relate to a concern for autonomy, which requires attention 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

[13] See Paul Dolan, Happiness By Design (2015).
[14] Id.

nudged to choose frontal lobotomies and became highly satisfied merely because of their reduced capacities.[15] In such cases, is the problem one of autonomy, 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. But even on welfarist grounds, we should nonetheless have concerns about process. 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 (AJBT or otherwise). 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 any one else."[16] 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 might have a rule-consequentialist defense, and hence be justified on welfarist grounds.

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

[16] JOHN STUART MILL, ON LIBERTY 8 (Kathy Casey ed., 2002) (1859).