"As Judged by Themselves": Freedom, Transformative Experiences, Law, and Testimony

L.A. PAUL[†] & CASS R. SUNSTEIN[††]

One way to evaluate various legal interventions in people's lives is to ask whether they make choosers better off by their own lights, or "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. If lawyers, judges, and policymakers give ultimate authority to choosers, this might be taken as respecting choosers' own judgments and promoting 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, which involve life-changing, transformative experiences, the criterion does not offer a unique solution; people might be happy with their choices either way. It is possible that welfarist criteria will resolve the indeterminacy, despite serious questions about incommensurability.

† Millstone Family Professor of Philosophy and Cognitive Science, Yale University.
†† Robert Walmsley University Professor, Harvard University.

TABLE OF CONTENTS

I. HARD CHOICES

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 lies in these kinds of choices, where each life-changing way you might choose is legally, socially, morally, and practically acceptable, and the choice you make will transform who you are, that is, the change it brings will change the kind of person you are. How should we make these kinds of choices? How should such choices be influenced?

It is plausible to think that the answers should be influenced by standard factors emphasized in models of rational choice—such as how happy each option would make you, which would generate the most life satisfaction, or, more precisely, which act would maximize your expected value, given everything that is relevant to what you consider to be a life well-lived. These factors matter. But as we will see, it is not always easy to determine just how they matter.

These issues have implications for law and policy. Suppose that public officials are seeking to promote human welfare through law. How can they do that, if a choice will have transformative effects, and if rational actor models run into trouble? We will get to these issues in due course.

Begin with a simple example, not involving law at all. Ella is faced with the decision of whether to have a child. She is a single woman in her thirties in the midst of a thriving career, ambitious and proud of her accomplishments, and unsure whether she wants to have a baby. She has the means for artificial insemination and to support a child, but she agonizes over whether this irreversible life choice is the right one for her, since, right now, she loves her job, cares much more about career success than parenting, and believes that parenthood is likely to impede her career. She is also unsure whether she would enjoy parenting. Maybe it would be an ordeal? Maybe she would regret it? 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 is filled with admiration for Norwegians. After repeated urging by his friends, he finally relocates to Oslo. After moving, he is very glad he did it. Though he still 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.

Now consider the question of whether to marry. Jon is happy in his life; he has plenty of friends. He is dating a woman named Anne, who wants to marry him. Jon is worried that if he marries, he will become staid and boring, and his life will become routinized. He decided not to marry Anne, who then breaks up with him. Now, he is glad that he did not marry Anne; he loves his freedom, and he loves who he is.

If Ella knew, before she chose to become a parent, that she would end up valuing parenting over her career, would it make sense, before she changed her values, for her to decide to have a baby? Why and in what sense? If Will knew that his choice to move to Norway would make him form new allegiances and lose touch with the political pulse of the United States, why would he choose to leave his native home, the place to which he is so committed? If Jon had married, and if we suppose that he would have been equally happy in his life with Anne, did he make the right decision? More generally, on what grounds should one choose a path that replaces one's current values? On what grounds should a person choose to avoid such a path? And how do the answers to these questions bear on law and policy—for example, those forms of law and policy that may or may not create incentives such as to have children or to marry, or that may or may not nudge people toward particular religious commitments?

Our main 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,[1] 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 demonstrate that their choice was better than the alternative choice. We will have a few things to say about how to evaluate choices when the AJBT criterion fails.

II. THE ARCHITECTURE OF CHOICE

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.[2] Choice architecture, much of it created by law, refers to the background conditions against which people make choices.[3] As an analogy, think of how the way a house is built affects the lives of the people living in it. 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

1. RICHARD THALER & CASS R. SUNSTEIN, NUDGE: THE FINAL EDITION 22 (2021) (highlighting the AJBT criterion).

2. For an overview on choice architecture, see generally *id.*

3. *Id.*

consider the winding paths of a garden with landscaping designed to lead you towards a particularly propitious view.

Similarly, choice contexts embodying choice architecture, made possible or mandated by law, can be designed to lead people towards particular options.[4] Is the choice opt-in or is it opt-out? Whether we are dealing with retirement plans, health care plans, or green energy, our choices might be greatly affected by the answer. Consider a website designed to emphasize Option A, by making it visible and salient, and to downplay Option B, so that fewer people might choose it. Or consider a case where one option is 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.[5] For example, people tend to be especially averse to losses, disliking them far more than they like the corresponding gains.[6] 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 decisionmakers in beneficial directions.[7] Consider a GPS device, which allows people to choose their own destination, and suggests a path, while also permitting people to reject its directions. The ultimate goal of the GPS device is to help people to get to where they want to go. The GPS device nudges, but does not require, users to choose the selected route.

Law can and does operate in the same way as a GPS device. 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, law or policy might frame information, or prescribe default rules, potentially with large effects on outcomes: just as a GPS device may steer more people towards the use of highways, so an opt-out design may significantly increase participation rates for retirement savings plans.[8] Default rules, use of order effects, and particular ways of describing and framing outcomes are all

4. *See* Jon M. Jachimowicz, Shannon Duncan, Elke U. Weber & Eric J. Johnson, *When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects,* 3 BEHAV. PUB. POL'Y 159, 160 (2019).

5. *See* Wändi Bruine De Bruin, *Framing Effects in Surveys: How Respondents Make Sense of the Questions We Ask*, *in* PERSPECTIVES ON FRAMING 303, 314 (Gideon Keren ed., 2015).

6. *See* EYAL ZAMIR, LAW, PSYCHOLOGY, AND MORALITY: THE ROLE OF LOSS AVERSION, at xiv (2014); José Luis Bermúdez, *Rational Framing Effects: A Multidisciplinary Case*, BEHAV. & BRAIN SCIS., Jan. 24, 2022, at 1, 2; Amos Tversky & Daniel Kahneman, *The Framing of Decisions and the Psychology of Choice*, 211 SCI. 453, 456 (1981).

7. *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 6 (2008).

8. *See* Jachimowicz et al., *supra* note 4, at 161.

ways to nudge.[9] Just as you might purchase an open plan house to encourage your teenagers to hang out in a common family space, nudges typically 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 be nudged? A pervasive answer points to maximizing the well-being of choosers. But how do we measure that? There is a great deal of work on this topic.[10] Along with Richard Thaler, Sunstein has argued that the principal criterion is whether people are made better off, *as judged by themselves* ("AJBT").[11] We shall say more about below the AJBT criterion, as we shall call it. Note initially that 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 of the choice.

For example, Omar, newly hired, needs to make a decision about his retirement plan. The default option in his plan has him set aside 8 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 highly likely to be a good one. As we will see, this criterion runs into serious questions when applied to transformative experiences, which are experiences that fundamentally change our self-defining values as measured before and after the choice is made.

III. LAW AND POLICY

The AJBT criterion can be applied in a context of individual decision making, as when an agent is making a choice, and from a more general perspective of law and policy, as from the standpoint of private and public choice architects (including those using artificial intelligence). 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. Policymakers might ask the same question for relevant populations in deciding, for example, which default rule to use for savings plans or health care plans. Artificial intelligence might be enlisted to identify the right nudge, making the AJBT criterion authoritative.[12]

---

9. *See generally, e.g.*, PERSPECTIVES ON FRAMING, *supra* note 5 (discussing research that documents framing effects, their underlying processes, and the conditions that produce differing interpretations of one message).

10. *See, e.g.*, B. Douglas Bernheim & Jonas Mueller-Gastell, *A General Solution to the Problem of Setting Optimal Default Options*, 112 AEA PAPERS & PROC. 131, 131 (2022).

11. THALER & SUNSTEIN, *supra* note 1, at 21.

12. Cass R. Sunstein, *Brave New World? Human Welfare and Paternalistic AI*, THEORETICAL INQUIRIES L. (forthcoming) (manuscript at 1–2) (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4908836.

In many cases, choosers can answer the relevant questions about their own well-being far better than policymakers can. The reason is that choosers frequently know what best fits their situation and how their lives are going—they know their own value framework better than policymakers do. If choosers are glad that they have been defaulted into a certain retirement plan, or a certain health care plan, we have reason to think that the default rule was a sensible one. These arguments point in favor of preserving freedom of choice and the AJBT criterion.

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. They might be ignorant of relevant factors, or they might suffer from some cognitive bias. To address that issue, we might want to know if some such bias or some sort of 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 whether choosers are in fact better off, not merely whether they believe they are. For this reason, giving decisive (as opposed to presumptive) authority to subjective judgments about well-being is a mistake.[13]

Nonetheless, the AJBT criterion has significant pragmatic appeal, perhaps especially when developing policy on which direction to nudge, but also for the ex post assessments of options. The AJBT criterion 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, because it fits their value framework.

For followers of John Stuart Mill, who emphasize the need to allow people to decide for themselves, that reason might be very good indeed.[14] 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, though it cannot do so in all cases. As a small example, consider the default rule for printing: single-sided or double-sided? As a large example, consider the default rule for savings plans: Do people have to opt in, or are they enrolled automatically?

Importantly, the AJBT criterion can claim to draw on strands of the liberal tradition that emphasize the importance of individual agency. If those involved in law and policy want to respect and promote that agency, they should be drawn to it. When people believe that they are better off, we might think that there is a

13. On various conceptions of welfare, see generally MATTHEW D. ADLER, WELL-BEING AND FAIR DISTRIBUTION: BEYOND COST-BENEFIT ANALYSIS (2012).

14. JOHN STUART MILL, ON LIBERTY 8 (Kathy Casey ed., 2002) (1859). For a sustained objection, see SARAH CONLY, AGAINST AUTONOMY: JUSTIFYING COERCIVE PATERNALISM 9 (2013).

sense in which the relevant nudge is consistent with a kind of autonomy that liberal policymakers wish to respect (we will qualify this point below). Finally, the use of the AJBT criterion has significant pragmatic value in how it sharply constrains outsiders—in both government and the private sector—by directing them to attend not to their own concerns, but to those of choosers.

So far, so good. But 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 specified. The problem is that, in an important class of cases, the criterion can fail to provide determinate or unique answers. These cases consist of those in which our choices are *transformative*, in the sense that they alter a chooser's core values between the time of the choice and the time of the measured effect, and by extension, their preferences over these times, by replacing core values at the earlier time with opposing values at the later time.[15] The changes of interest bring about an endogenous change in the chooser's preferences and values, which has implications for our understanding of their post-hoc judgments about the change.

IV. TRANSFORMATIVE EXPERIENCES

Transformative experiences involve large-scale changes in your values, leading to changes in who you are. Small changes in value are routine. They correspond to relatively mundane matters—food preferences, films, clothing, laptops—and do not involve self-transformation. For such cases, the AJBT criterion works well. 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, who is faced with the decision of whether to have a child, and Will, who is deciding whether to expatriate from the United States to Norway.

Such large-scale changes are *epistemically transformative*, in the sense that they change what a person knows, and by extension, their values, through having new kinds of experiences. Having the experience is necessary for the change in knowledge. And 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 change in one's "core" preferences, as in our examples, where a person switches from valuing career over family to valuing family over career, or switches from identifying with one's native country to identifying with one's adoptive country. On an approach where one's core preferences define oneself, the shift can be thought of as a process where

15. 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 2 (2014); *see also* EDNA ULLMANN-MARGALIT, NORMAL RATIONALITY: DECISIONS AND SOCIAL ORDER 69–70 (Avishai Margalit & Cass R. Sunstein eds., 2017).

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.[16]

In personal transformation, one's old self is replaced with a new self by 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. We can define a self at a time by reference to 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 composes 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. In this way, we can think of a persisting person as being realized, over time, by a series of appropriately causally related and psychologically connected selves. When someone's old self is replaced by an appropriately generated, 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 (and thus preferences). As a result of this change in values, the person's ex ante self (the self who makes the choice) is replaced by their ex post self (the self who results from the choice), altering the kind of person they are. A transformative change in a person is a radical change in the self that realizes that person. That change alters who they are through a radical change in their values, from valuing A at the time of making the choice to valuing B at the time of the effect of the choice (where B is inconsistent with A). The change in values creates a corresponding change in preferences. As such, transformative choices amount to choices to change oneself in an especially far-reaching way.

Choices that bring about transformative change have several distinctive features. One 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? By definition, in transformation, the person who is making the choice needs to undergo the transformation before they can appreciate the value—or disvalue—of the person they might become.[17] However, our main focus here is on another property of transformative choices:

16. One's values ordinarily determine one's preferences: if I value A over B, and I am rational, I prefer A to B. The background assumption here is that such a preference determination requires a static preference framework over time.

17. L.A. Paul, *Value by Acquaintance* (forthcoming) (on file with authors).

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*.[18]

Recall: the transformative choice amounts to a choice by a chooser 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 discussions of rational action. The trouble is that 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.

It should immediately be apparent that in most cases of law and policy, we are not dealing with transformative experiences, and for this reason, the AJBT criterion is a good place to start. But in scenarios where choices are transformative, the AJBT criterion is far less useful. If people are automatically enrolled in a health care plan, or warned about the presence of shellfish in their meals, or given information about the fuel economy of vehicle options, their preferences and values are unlikely to shift in any large-scale way. But transformation may nonetheless be influenced by law.

Suppose, for example, that people are nudged or incentivized to marry, perhaps via social norms, perhaps via law. Or suppose that the law nudges or incentivize people to have children, or not to have children. How can we know whether any such intervention is a good idea? How are we to regard the evidence we seem to get when we collect testimony about their satisfaction from transformed individuals?

V. THE ENDOGENEITY OF PREFERENCES

To get clearer on how endogeneity creates a problem for the AJBT criterion and the evaluation of testimonial evidence, let us start by looking at a simple example, one that does not involve transformation at all. 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 should have an operation, which carries with it 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

18. 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 & Kieran Healy, *Transformative Treatments*, 52 NOÛS 320, 320 (2018).

does not. Raul decides to follow the advice, and a year later, he is glad he did. In a different possible world (a parallel, or counterfactual, 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.[19] 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 ex post 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 current preferences. He has not changed anything fundamental to his values or preferences. He has not changed the kind of person he is. It follows not that the AJBT criterion is wrong, but that it is indeterminate. In the abstract, the same could be true of a wide range of legal interventions. People might be satisfied with one health care plan, having been automatically enrolled in it, and also with a very different health care plan, having been automatically enrolled in it. This is more likely if their preferences are endogenous to the default rule. One reason may be that people habituate to the status quo that they have brought about.[20]

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, which, again, makes the AJBT criterion indeterminate, in the sense that it can be satisfied by more than one nudge.

---

19. *Cf.* Gilbert Harman, *Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error*, 99 PROC. ARISTOTELIAN SOC'Y 315, 329 (1999) ("[T]her is no evidence that people differ in character traits. They differ in their situations and in their perceptions of their situations.").

20. *See* TALI SHAROT & CASS R. SUNSTEIN, LOOK AGAIN: THE POWER OF NOTICING WHAT WAS ALWAYS THERE 15–16 (2024). On the phenomenon of adaptive preferences, see JON ELSTER, SOUR GRAPES: STUDIES IN THE SUBVERSION OF REALITY 109–13 (1983).

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 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 are not uncommon, 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 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, because a person's preferences could gradually evolve in more than one way. That problem is a serious one for law and policy. It suggests the need to look elsewhere in deciding which way to nudge. This is an unresolved challenge for behavioral law and economics.

VI. REPLACING PREFERENCES

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. As the case of Raul suggests, that possibility is broader than that of transformative experiences. What then is the special problem with transformation? The difference is a matter of degree that results in a change of state. Transformation involves replacement of one or more core values, and this change is significant enough to change the kind of person one is. In this section, we will explore this question by developing the metaphysical structure of transformative preference replacement and exploring its implications.

Recall that when a person transforms themselves, their values and preferences undergo a seismic change. As outlined 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 that 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 an abortion. The first way she could be nudged, perhaps by law, 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.[21]

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 ended up choosing 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? How are policymakers, or courts, to answer that question?

The endogeneity arises because transformative choices "satisfy a preference" only by replacing a person's self, thus replacing their preferences in such a way as to create the satisfied individual that the person ends up becoming. (Again, habituation may be relevant here.) 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 by not opting-out of the default retirement plan. 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 (*1*) before he made his retirement savings choice, and then assess his satisfaction with his new state (*2*) 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

21. *Cf.* Harman, *supra* note 19 (discussing the effect of situations humans experience and human perception of such situations on differences in actions and decisionmaking).

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.[22] If, when we say, "Omar makes his choice at t1 and he is satisfied with his choice at t2," our use of "he" at t2 picks out the testimony of *somebody else* (for example, 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.[23]

In transformative contexts, including those relevant to law, such as those involving marriage, divorce, and parenthood, 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 a change in state entails a change in who the agent is.

Returning to Ella, if she chooses to have the child, the self that results from the choice 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 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: "Ella" at t1 is a child free person, while "Ella" at t2 is a parent, with all the cares and loves and responsibilities that such a transformation entails.

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. "Ella" at t1 is child free, but open to the possibility of becoming a parent, while "Ella" at t2 is a deeply committed anti-natalist.

---

22. For related methodological discussion, see Paul & Healy, *supra* note 18, at 323; L.A. Paul & John Quiggin, *Real World Problems*, 15 EPISTEME 363, 366–67 (2018).

23. Imagine, when following up with study participants thirty years later, that due to a records mix-up, we interview the wrong person—someone else, from a different study, also named "Omar."

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 you were before you chose, along with who you would be now if you had remained childless.

We may think of this in terms of comparing the testimony from your different possible selves. If you, as you are now, were to compare yourself in a meaningful way with your past self, for example, what you care about most, how you spend your time, who you spend your time with, and so on, you would likely find that your values have 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, tracked in terms of what (and who) you care about most, and thus what it means to have your preferences satisfied has changed at least as much. As a result, 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 at t1 and t2 than measuring a change in testimony across the same person's change of state from t1 to t2.[24] In these kinds of cases, if there is no clear 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 simply use the AJBT criterion to decide which nudge is better. A world in which people are grateful to have been nudged to have children might, superficially, seem like a world full of prudent nudging. But is it?

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 an important 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

24. For an iteration of this worry that translates into empirical comparisons in observational contexts, see Paul & Healy, *supra* note 18, at 323; Paul & Quiggin, *supra* note 22, at 373.

own possibility for self-transformation, if such testimony is endogenously generated from the transformative experience in question?[25]

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 very 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.[26]

We are aware that all this is somewhat abstract, and that we have focused on individual cases. We think that related questions extend to choices that affect individuals in a variety of ways, including those involving transformation on a large scale, such as wars, pandemics, and migration, although we have not developed the argument here. In any case, we have noted that law may or may not encourage or discourage transformative experiences. How and when should it be done? The AJBT criterion might not help. 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 might not know in which direction a person should be nudged. For that reason, it is insufficient in the context of transformative choices and events.

VII. LAW AND WELFARE

We have argued that the AJBT is not sufficient for welfare determination, and thus not sufficient, on its own, for approval of a nudge for an individual. In this section, we will explore the implications of this conclusion, and offer a few words about what to do about it. The exploration bears (1) on what individuals might do and how to evaluate their choices and (2) on what those involved in law and policy might do, and how to evaluate their choices.

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

25. *See* Vladimir Chituc, L.A. Paul & M.J. Crockett, *Evaluating Transformative Decisions,* 43 PROC. ANN. MEETING COGNITIVE SCI. SOC'Y 973, 973 (2021); *see also* Paul, *supra* note 17 (discussing how this relates to reference class problems).

26. *Cf.* Harman, *supra* note 19 (discussing the effect of situations humans experience and human perception of such situations on differences in actions and decisionmaking).

your choice. In this sense, 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 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 wrong to focus only on what will make her happiest in some narrow sense. Consider these words:

> 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.[27]

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

It is important to emphasize that all this is consistent with the AJBT criterion. 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 (we are bracketing the many cases in which parents regret having become parents).[28]

Similar kinds of situations arise 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

27. William Talbott, *Transformative Experience,* 76 ANALYSIS 380, 385 (2016).

28. This debate is highly complex, with varying degrees of supporting evidence for the arguments that have been presented. *See generally* Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz & Arthur A. Stone, *A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method*, 306 SCI. 1776 (2004); S. Katherine Nelson, Kostadin Kushlev, Tammy English, Elizabeth W. Dunn & Sonja Lyubomirsky, *In Defense of Parenthood: Children Are Associated With More Joy Than Misery*, 24 PSYCH. SCI. 3, 6–7 (2013); S. Katherine Nelson, Kostadin Kushlev & Sonja Lyubomirsky, *The Pains and Pleasures of Parenting: When, Why, and How Is Parenthood Associated With More or Less Well-Being?*, 140 PSYCH. BULL. 846, 846–47 (2014).

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 that this can be true despite the fact that their counterfactual selves would also testify that they value the life that they actually have over the counterfactual life—that is, the counterfactual life from their counterfactual point of view—which they do not have.

The *welfarist* will argue that to compare outcomes, we need to determine a person's welfare in each possible state. Welfarism has a strong hold on regulatory law.[29] To assess welfare effects, 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. Once we have properly assessed a person's welfare in each possible state, 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 (or incentivizing) someone toward 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. 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. Welfarists have an approach with which to work, and we think that they have a strong claim for use of their approach in law.[30]

IX. PATHS FORWARD

Our principal argument has been 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.

Our focus has been on the theoretical apparatus, but we have also referred to applications for law and policy. Many policy issues might be assessed in terms of the AJBT criterion, involving (for example) savings policy and health care policy. Policies relating to marriage, gender transition, divorce, adultery, abortion, and (more broadly) whether to have children run into the possibility that the transformative nature of the decision confounds the AJBT criterion. Should the law encourage one or another (bracketing constitutional issues)? What should be done in such cases?

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,

29. *See* Cass R. Sunstein, *The Economic Constitution of the United States,* 38 J. ECON. PERSPS. 25, 25–26 (2024).

30. One of the current authors (Sunstein) likes it a bit better than the other (Paul).

and what nudge, really makes people better off? To answer such questions, we need a specification of the right conception of welfare. Suppose that we place an emphasis on people's subjective experience and assume that we could measure its value, or at least come pretty close.[31] 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.[32] It might also pay attention to diversity of experience, which could point in different directions, and which should also be measurable.[33]

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 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 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 intimated 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 such that they replace their old self with a new self). 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 not merely nudged but also manipulated[34] or even coerced into a transformative experience (or otherwise to a situation that alters their values and preferences such that they replace their old self with a new self). 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 satisfying the AJBT criterion is no excuse or justification for manipulation or coercion.

---

31. *See* PAUL DOLAN, HAPPINESS BY DESIGN: CHANGE WHAT YOU DO, NOT HOW YOU THINK 36–38, 90–91 (2015).

32. *Id*. at 35, 43–46.

33. *See* SHAROT & SUNSTEIN, *supra* note 20.

34. MANIPULATION: THEORY AND PRACTICE (Christian Coons & Michael Weber eds., 2014) provides an excellent overview.

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 kidnapping 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 to appreciate their situation.[35] Or to take a less extreme case, what about 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?[36] In such cases, does the problem involve 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 manipulation[37] or coercion.[38] The only welfarist thought here is that the AJBT captures something of value, post hoc approval, which is consistent with the idea that people's welfare has been increased. 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 the 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 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

35. *See* Elizabeth Barnes, *What You Can Expect When You Don't Want to be Expecting*, 91 PHIL. & PHENOMENOLOGICAL RSCH. 775, 785 (2015).

36. Bloom and Paul REF.

37. Jonathan Baron, *A Welfarist Approach to Manipulation* 1 J. MKTG. BEHAV. 283, 283–84 (2015).

38. *See* CONLY, *supra* note 14, at 3.

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."[39] 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."[40] If Mill is even broadly correct, a rule or presumption against coercion or against inadequately grounded nudging is justified on welfarist grounds. Mill's own form of welfarism raises many questions;[41] it is certainly not crudely utilitarian. But its account of what makes for good lives, and its emphasis on what is hard to quantify, has unmistakable implications for law and policy; and it is of more practical use than it might seem.[42]

39. MILL, *supra* note 14.

40. *Id.*

41. *See* DAVID O. BRINK, MILL'S PROGRESSIVE PRINCIPLES, at x (2015).

42. For evidence, see Sunstein, *supra* note 29 (finding that the OMB Circular A-4 systemized political philosophy yet struggled with assigning monetary value).

***