Int Rev Econ (2018) 65:1–8
https://doi.org/10.1007/s12232-017-0280-9

RESEARCH ARTICLE

"Better off, as judged by themselves": a comment on evaluating nudges

Cass R. Sunstein[1]

Received: 9 June 2017 / Accepted: 15 June 2017 / Published online: 22 June 2017

Abstract Many nudges are designed to make people better off, as judged by themselves. This criterion, meant to ensure that nudges will increase people's welfare, contains some ambiguity. It is useful to distinguish among three categories of cases: (1) those in which choosers have clear antecedent preferences, and nudges help them to satisfy those preferences (often by increasing "navigability"); (2) those in which choosers face a self-control problem, and nudges help them to overcome that problem; and (3) those in which choosers would be content with the outcomes produced by two or more nudges, or in which ex post preferences are endogenous to nudges, so that without additional clarification or work, the "as judged by themselves" criterion does not identify a unique solution for choice architects. Category (1) is self-evidently large. Because many people agree that they suffer self-control problems, category (2) is large as well. Cases that fall in category (3) create special challenges, which may lead us to make direct inquiries into welfare or to explore what informed, active choosers typically select.

Keywords Nudges · Default rules · Preferences · Behavioral economics

JEL Classification D003 · D10 · D11 · D18 · D60 · D80 · K0 · K2

Some nudges are designed to reduce externalities; consider fuel economy labels that draw attention to environmental consequences, or default rules that automatically enroll people in green energy (Sunstein and Reisch 2014). But many nudges are

✉ Cass R. Sunstein
csunstei@law.harvard.edu

[1] Harvard University, Cambridge, MA 02138, USA

designed to increase the likelihood that people's choices will improve their own welfare. Richard Thaler and I argue that the central goal of such nudges is to "make choosers better off, *as judged by themselves*" (Thaler and Sunstein 2008, p. 5; italics in original; Thaler 2015). Social planners—or in our terminology, choice architects—might well have their own ideas about what would make choosers better off, but in our view, the lodestar is people's own judgments. To be a bit more specific: The lodestar is welfare, and under the appropriate conditions, people's own judgments are a good (if sometimes imperfect) way to test the question whether nudges are increasing their welfare.

The last sentence raises many questions, and it is certainly reasonable to wonder about potential ambiguities in the "as judged by themselves" (hereinafter AJBT) criterion. Should the focus be on choosers' judgments before the nudge, or instead after? What if the nudge alters people's preferences, so that they like the outcome produced by the nudge, when they would not have sought that outcome in advance? What if preferences are constructed by the relevant choice architecture (Lichtenstein and Slovic 2006)? Or what if people's ex ante judgments are wrong, in the sense that a nudge would improve their welfare, even though they do not think that it will (Dolan 2014)? Do we want to ask about choosers' actual, potentially uninformed or behaviorally biased judgments, or are we entitled to ask what choosers would think if they had all relevant information and were unaffected by relevant biases (Goldin 2015)? In an instructive essay, Robert Sugden offers two interpretations of the AJBT criterion (Sugden 2016). He objects that the first, involving self-control problems, applies only to a narrow range of cases, while the second, involving "latent preferences," is "straightforwardly paternalistic," and ensures that the criterion cannot achieve its purpose "of countering the criticism that nudge policies are paternalistic" (ibid.).

It is always hazardous to disagree with Sugden, but I see things differently. My goal here is to explore the meaning of the AJBT criterion and to sort out some of the ambiguities. As we shall see, three categories of cases should be distinguished: (1) those in which choosers have clear antecedent preferences, and nudges help them to satisfy those preferences; (2) those in which choosers face a self-control problem, and nudges help them to overcome that problem; and (3) those in which choosers would be content with the outcomes produced by two or more nudges, or in which ex post preferences are endogenous to or constructed by nudges, so that the AJBT criterion leaves choice architects with several options, without specifying which one to choose. Cases that fall in category (1) plainly satisfy the AJBT criterion, and there are many such cases. From the standpoint of the AJBT criterion, cases that fall in category (2) are also unobjectionable; indeed, they can be seen as a subset of category (1), and they too are plentiful. Cases that fall in category (3) create special challenges, which may lead us to make direct inquiries into people's welfare or to explore what informed, active choosers typically select.

Sugden's question, presented in his title, is simple: "Do people really want to be nudged toward healthy lifestyles?" That is an empirical question, and we have a great deal of evidence about it. The answer is "yes," at least in the sense that in numerous nations—including the USA, the UK, Germany, France, Italy and Australia—strong majorities endorse nudges toward healthy lifestyles (Jung and

Mellers 2016; Sunstein 2016a, b; Reisch and Sunstein 2016; Sunstein et al. 2017). One might object that general attitudes toward nudges do not specifically answer Sugden's question, and that it is best not to ask people generally (1) whether they approve of nudges, but instead to ask more specifically (2) whether they themselves would like to be nudged (on potential differences between answers to the two questions, see Cornwell and Krantz 2014). But general approval of health-related nudges strongly suggests that the answer to (2) is probably "yes" as well, and in any case, the existing evidence suggests that the answer to (2), asked specifically, is also "yes" (Jung and Mellers 2016). To be sure, more remains to be learned on these issues.

Because Sugden does not explore the data, it is not clear that he means to ask an empirical question at all. While he offers some empirical claims, his interest seems to be more conceptual and normative. What does it even mean to say that people want to make the choices toward which they are being nudged? What is the relationship between people's preferences and imaginable nudges? In light of behavioral findings, how confidently can we speak of people's "preferences"? If people want to be nudged, why are they not already doing what they would be nudged to do? Those are important questions (and they do have empirical features).

Thaler and I are interested in "libertarian paternalism," in the form of approaches that preserve freedom of choice, but that steer people in a direction that will promote their welfare. Coercive paternalism is very different (Conly 2014). Reminders, warnings, information disclosure, invocations of social norms and default rules are examples of what we have in mind (Sunstein 2016a, b). Contrary to Sugden's suggestion, the AJBT criterion is not at all designed to counter the charge of paternalism. On the contrary, Thaler and I explicitly embrace (a mild form of) paternalism, and the purpose of the criterion is *to discipline the content of paternalistic interventions*. Consider, for example, a GPS device. It will identify the best route for you, given the direction that you identify; it makes you better off by your own lights. It increases *navigability*. At the same time, it is paternalistic in the sense that it purports to know, better than you do, how to get where you want to go. Its paternalism is one of means, not ends (Sunstein 2014). Means paternalism is unquestionably a form of paternalism, but it will typically satisfy the AJBT criterion, because it is respectful of people's ends. To be sure, we might find cases in which the distinction between means and ends is not straightforward (ibid.), but most cases are easy.

It is not possible to understand the operation of the AJBT criterion without reference to examples. In countless cases, we can fairly say that *given people's antecedent preferences*, a nudge will make choosers better off AJBT. For example:

1. Luke has heart disease, and he needs to take various medications. He wants to do so, but he is sometimes forgetful. His doctor sends him periodic text messages. As a result, he takes the medications. He is very glad to receive those messages.
2. Meredith has a mild weight problem. She is aware of that fact, and while she does not suffer from serious issues of self-control, and does not want to stop eating the foods that she enjoys, she does seek to lose weight. Because of a new

law, many restaurants in her city have clear calorie labels, informing her of the caloric content of various options. As a result, she sometimes chooses low-calorie offerings—which she would not do if she were not informed. She is losing weight. She is very glad to see those calorie labels.

3. Edna is a professor at a large university, which has long offered its employees the option to sign up for a retirement plan. Edna believes that signing up would be a terrific idea, but she has not gotten around to it. She is somewhat embarrassed about that. Last year, the university switched to an automatic enrollment plan, by which employees are defaulted into the university's plan. They are allowed to opt out, but Edna does not. She is very glad that she has been automatically enrolled in the plan.

There is nothing unfamiliar about these cases. On the contrary, they capture a great deal of the real-world terrain of nudging, both by governments and by the private sector (Sunstein 2016a, b; Halpern 2015). Choosers have a goal, or an assortment of goals, and the relevant choice architecture can make it easier or harder for them to achieve it or them. Insofar as we understand the AJBT criterion by reference to people's antecedent preferences, that criterion is met. Note that it would be easy to design variations on these cases in which nudges *failed* that criterion, because they would make people worse off by their own lights.

We could complicate the cases of Luke, Meredith and Edna by assuming that they have clear antecedent preferences, that the nudge is inconsistent with those preferences, but that as a result of the nudge, their preferences are changed. For example:

> Jonathan likes talking on his cell phone while driving. He talks to friends on his commute to work, and he does business as well. As a result of a set of vivid warnings, he has stopped. He is glad. He cannot imagine why anyone would talk on a cell phone while driving. In his view, that is too dangerous.

After the nudge, Luke, Meredith, Edna and Jonathan believe themselves to be better off. Cases like Jonathan's raise the question whether the AJBT criterion requires reference to ex ante or ex post preferences. That is a good question, which might be answered by making direct inquiries into people's welfare; I will turn to that question below. My main point is that as I have described them, the cases of Luke, Elizabeth and Edna are straightforward. Such cases are common (Halpern 2015; Thaler 2015).

Some cases can be seen as different, because they raise questions about self-control (or akrasia):

1. Ted smokes cigarettes. He wishes that he had not started, but he has been unable to quit. His government has recently imposed a new requirement, which is that cigarette packages must be accompanied with graphic images, showing people with serious health problems, including lung cancer. Ted is deeply affected by those images; he cannot bear to see them. He quits, and he is glad.
2. Joan is a student at a large university. She drinks a lot of alcohol. She enjoys it, but not that much, and she is worried that her drinking is impairing her

> performance and her health. She says that she would like to scale back, but for reasons that she does not entirely understand, she has found it difficult to do so. Her university recently embarked on an educational campaign to reduce drinking on campus, in which it (accurately) notes that four out of five students drink only twice a month or less. Informed of the social norm, Joan finally resolves to cut back her drinking. She does, and she is glad.

In these cases, the chooser suffers from a self-control problem and is fully aware of that fact. Ted and Joan can be seen as both planners, with second-order preferences, and doers, with first-order preferences (Thaler and Sunstein 2008; Thaler 2015). A nudge helps to strengthen the hand of the planner. It is possible to raise interesting philosophical and economic questions about akrasia and planner–doer models (Stroud and Tappolet 2003), but insofar as Ted and Joan welcome the relevant nudges, and do so ex ante as well as ex post, the AJBT criterion is met. In a sense, self-control problems require their own GPS devices and so can be seen to involve navigability; but for choosers who face such problems, the underlying challenge is qualitatively distinctive, and they recognize that fact.

Sugden agrees that in such cases, the AJBT criterion is met. His argument is convincing. At the same time, he conjectures, "based both on casual social experience and on general scientific knowledge" (Sugden 2016) that people will not commonly acknowledge that they face a self-control problem. That is an empirical question, of course, and my own preliminary research suggests that Sugden is not right. On Amazon's Mechanical Turk, I asked about 200 people this question:

> Many people believe that they have an issue, whether large or small, of self-control. They may eat too much, they may smoke, they may drink too much, they may not save enough money. Do you believe that you have any issue of self-control?

A whopping 70% said that they did (55% said "somewhat agree," while 15% said "strongly agree"). Only 22% disagreed. (Eight percent were neutral.)

This is a preliminary test, of course. Whether or not Sugden's empirical conjecture is right, the cases of Ted and Mary capture a lot of the territory of human life, as reflected in the immense popularity of programs designed to help people to combat addiction to tobacco (Halpern et al. 2015) and alcohol. We should agree that nudges that do the work of such programs, or that are used in such programs (ibid.), are likely to satisfy the AJBT criterion.

There are harder cases. In some of them, it is not clear if people have antecedent preferences at all. In others—as in the case of Jonathan—their ex post preferences are an artifact of, or constructed by, the nudge. Sometimes these two factors are combined (as marketers are well aware). As Amos Tversky and Richard Thaler put it long ago, "values or preferences are commonly constructed in the process of elicitation" (Tversky and Thaler 1990). If so, how ought the AJBT criterion to be understood and applied?

For example:

1. George cares about the environment, but he also cares about money. He currently receives his electricity from coal; he knows that coal is not exactly good for the environment, but it is cheap, and he does not bother to switch to wind, which would be slightly more expensive. He is quite content with the current situation. Last month, his government imposed an automatic enrollment rule on electricity providers: People will receive energy from wind, and pay a slight premium, unless they choose to switch. George does not bother to switch. He says that he likes the current situation of automatic enrollment. He approves of the policy and he approves of his own enrollment.
2. Mary is automatically enrolled in a Bronze Health Care Plan—it is less expensive than Silver and Gold, but it is also less comprehensive in its coverage, and it has a higher deductible. Mary prefers Bronze and has no interest in switching. In a parallel world (a lot like ours, but not quite identical, Wolf 1990), Mary is automatically enrolled in a Gold Health Care Plan—it is more expensive than Silver and Bronze, but it is also more comprehensive in its coverage, and it has a lower deductible. Mary prefers Gold and has no interest in switching.
3. Thomas has a serious illness. The question is whether he should have an operation, which is accompanied with potential benefits and potential risks. Reading about the operation online, Thomas is not sure whether he should go ahead with it. Thomas' doctor advises him to have the operation, emphasizing how much he has to lose if he does not. He decides to follow the advice. In a parallel world (a lot like ours, but not quite identical), Thomas's doctor advises him not to have the operation, emphasizing how much he has to lose if he does. He decides to follow the advice.

In the latter two cases, Mary and Thomas appear to lack an antecedent preference; what they prefer is an artifact of the default rule (in the case of Mary) or the framing (in the case of Thomas). George's case is less clear, because he might be taken to have an antecedent preference in favor of green energy, but we could easily understand the narrative to mean that his preference, like that of Mary and Thomas, is endogenous to the default rule.

These are the situations on which I am now focusing: People lack an antecedent preference, and what they like is a product of the nudge. Their preference is constructed by it. After being nudged, they will be happy and possibly grateful. We have also seen that even if people have an antecedent preference, the nudge might change it, so that they will be happy and possibly grateful even if they did not want to be nudged in advance.

In all of these cases, application of the AJBT criterion is less simple. Choice architects cannot contend that they are merely vindicating choosers' ex ante preferences. If we look ex post, people do think that they are better off, and in that sense the criterion is met. For use of the AJBT criterion, the challenge is that, *however Mary and Thomas are nudged, they will agree that they are better off*. In my view, there is no escaping at least some kind of welfarist analysis in choosing between the two worlds in the cases of Mary and Thomas. There is a large question about which nudge to choose in such cases (for relevant discussion, see Dolan 2014;

Goldin 2015). Nonetheless, the AJBT criterion remains relevant in the sense that it constrains what choice architects can do, even if it may not specify a unique outcome (as it does in cases in which people have clear ex ante preferences and in which the nudge does not alter them).

If I understand him correctly, Sugden is concerned with problems of this sort; in his view, the AJBT criterion is not helpful and nudging becomes self-evidently paternalistic. Invoking an elaborate and careful paper of which he is coauthor (Infante et al. 2016), Sugden explores an implicit model of choosers, in which an inner rational agent, equipped with stable, context-independent, latent preferences, is interacting with the world through a behaviorally biased, error-prone shell. Sugden urges that this model does not have adequate psychological foundations. He notes that in some cases (akin to those of Mary and Thomas), error-free agents are also affected by choice architecture, as when a cafeteria display leads everyone in a predictable direction. Thus:

> Gretchen enjoys her employer's cafeteria. She tends to eat high-calorie meals, but she knows that, and she likes them a lot. Her employer recently redesigned the cafeteria so that salads and fruits are the most visible and accessible. She now chooses salad and fruit, and she likes them a lot.

By stipulation, Gretchen suffers from no behavioral bias, but she is affected by the nudge. Sugden's conclusions (I think) are that if we are speaking of latent preferences, the AJBT criterion is unhelpful, and nudges turn out to be "straightforwardly paternalistic," in the sense that they are designed to promote people's welfare, whatever they prefer (in advance).

I confess that I am not quite sure what Sugden means by the idea of "latent preferences." Again: Thaler and I embrace paternalism, and so the AJBT criterion is emphatically not designed to defeat a charge of paternalism. It is psychologically fine (often) to think that choosers have antecedent preferences (whether or not "latent"), but that because of a lack of information or a behavioral bias, their choices will not satisfy them. (Recall the cases of Luke, Meredith and Edna.) To be sure, it is imaginable that some forms of choice architecture will affect people who have information or lack such biases. An error-free cafeteria visitor, like Gretchen, might grab the first item she sees, because she is busy, and because it is not worth it to her to decide which item to choose; she picks (Ullmann-Margalit 2017). But in many (standard) cases, behaviorally biased or uninformed choosers will be affected by a nudge, and less biased and informed choosers will not be; a developing literature explores how to proceed in such cases, with careful reference to what seems to me a version of the AJBT criterion (Goldin 2015; Goldin and Lawson 2016).

In Gretchen's case, and all those like it, the criterion does not leave choice architects at sea: If she did not like the salad, the criterion would be violated. From the normative standpoint, it may not be entirely comforting to say that nudges satisfy the AJBT criterion if choice architects succeed in altering the preferences of those whom they are targeting. (Is that a road to serfdom? Recall the chilling last lines of Orwell's 1984: "He had won the victory over himself. He loved Big Brother" (cf. Elster 1983)). But insofar as we are concerned with subjective welfare,

it is a highly relevant question whether choosers believe, ex post, that the nudge has produced a good outcome for them.

Countless nudges increase navigability, writ large, in the sense that they enable people to get where they want to go and, therefore, enable them to satisfy their antecedent preferences. Many other nudges, helping to overcome self-control problems, are warmly welcomed by choosers and so are consistent with the AJBT criterion. Numerous people acknowledge that they suffer from such problems. When people lack antecedent preferences or when those preferences are not firm, and when a nudge constructs or alters their preferences, the AJBT criterion is more difficult to operationalize, and it may not lead to a unique solution. But it restricts the universe of candidate solutions, and in that sense helps to orient nudgers. Even in such cases, solutions are emerging for welfare-oriented choice architects (Goldin 2015).

Acknowledgements I am grateful to the Program on Behavioral Economics and Public Policy for support and to Jacob Goldiu and Lucia Reisch for valuable comments on a previous draft.

References

Conly S (2014) Against autonomy. Oxford University Press, Oxford
Cornwell JF, Krantz DH (2014) Public policy for thee, but not for me: varying the grammatical person of public policy justifications influences their support. Judgm Decis Mak 5:433–444
Dolan P (2014) Happiness by design. Penguin, New York
Elster J (1983) Sour grapes. Cambridge University Press, Cambridge
Goldin J (2015) Which way to nudge? Uncovering preferences in the behavioral age. Yale Law J 125:226–271
Goldin J, Lawson N (2016) Defaults, mandates, and taxes: policy design with active and passive decision-makers. Am J Law Econ 18:438–462
Halpern D (2015) Inside the nudge unit: how small changes can make a big difference. W. H. Allen, London
Halpern SD et al (2015) Randomized trial of four financial-incentive programs for smoking cessation. N Eng J Med 372:2108–2211
Infante G, Lecouteux G, Sugden R (2016) Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. J Econ Methodol 23:1–25
Jung JY, Mellers BA (2016) American attitudes toward nudges. Judgm Decis Mak 11(1):62–74
Lichtenstein S, Slovic P (2006) The construction of preference. Cambridge University Press, Cambridge
Reisch L, Sunstein CR (2016) Do Europeans like nudges? Judgm Decis Mak 11:310–325
Stroud S, Tappolet C (eds) (2003) Weakness of will and practical irrationality. Clarendon Press, Oxford
Sugden R (2016) Do people really want to be nudged towards healthy lifestyles? Int Rev Econ. doi:10.1007/s12232-016-0264-1
Sunstein CR (2014) Why Nudge?. Yale University Press, New Haven
Sunstein CR (2016a) The council of psychological advisers. Ann Rev Psychol 67:713–737
Sunstein CR (2016b) The ethics of influence. Cambridge University Press, Cambridge
Sunstein CR, Reisch LA (2014) Automatically green: behavioral economics and environmental protection. Harvard Environ Law Rev 38:127–158
Sunstein CR, Reich L, Rauber J (2017) Behavioral insights all over the world? Public attitudes toward nudging in a multi-country study. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2921217
Thaler RH (2015) Misbehaving. Norton, New York
Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Tversky A, Thaler RH (1990) Anomalies: preference reversals. J Econ Perspect 4:201–211
Ullmann-Margalit E (2017) Normal rationality. Oxford University Press, Oxford
Wolf F (1990) Parallel universes: the search for other worlds. Simon & Schuster, New York