Review of Behavioral Economics, 2018, 5: 261–280

Efficacious and Ethical Public Paternalism

Daniel M. Hausman*

Department of Philosophy, University of Wisconsin–Madison, 600 North Park Street, Madison, WI 53706; dhausman@wisc.edu

ABSTRACT

People often make bad judgments. A big brother or sister who was wise, well-informed, and properly-motivated could often make better decisions for almost everyone. But can governments, which are not staffed with ideal big brothers or sisters, improve upon the mediocre decisions individuals make? If so, when and how? The risks of extending the reach of government into guiding individual lives must also be addressed.

This essay addresses three questions concerning when paternalistic policies can be efficacious, efficient, and safe:

1. In what circumstances can policy makers be confident that they know better than individuals how individuals can best promote their own well-being?
2. What are the methods governments can use to lead people to make decision that are better for themselves?
3. What are the moral pluses and minuses of these methods?

Answering these questions defines a domain in which paternalistic policy is an attractive option.

Although I am not a psychologist or a behavioral economist, it seems to me obvious that people often make bad judgments both when they seek to benefit themselves and when they have other objectives. (Just look at the results of

*I am indebted for some of the ideas in this paper to Brynn Welch and Luc Bovens. Julian Le Grand offered helpful criticisms of a previous draft. Comments at the Behavioral Economics and New Paternalism workshop at New York University in April 2018 were also extremely helpful. This research was supported by the Ludwig Lachmann Fellowship at the London School of Economics.

ISSN 2326-6198; DOI 10.1561/105.00000090

the last U.S. presidential election.) Unless I am particularly incompetent, I have good evidence from my own experience that we human decision makers are often not very good at the task. A big brother or big sister who was wise, well-informed, and sought my good could make better decisions for me than I make. But can governments, which are not staffed with wise and benevolent big brothers and big sisters, improve on the mediocre decisions individuals make?

Paternalistic policies are justified only if they can make people better off than they would be making their own mistakes. This is only a necessary condition. Paternalistic policies that enhance well-being, may nevertheless be undesirable because they dangerously extend the reach of government into individual lives. Authorizing paternalistic policies gives governments power to do more good, but governments may use that power to do harm. In addition, in shielding individuals from their mistakes, state paternalism may make them less able to manage their own lives.

After defining paternalism in Section 1, Sections 2–4 address in order the following three questions concerning when paternalistic policies can be efficacious and ethical:

1. In what circumstances can policy makers be justifiably confident that they know better than the individuals themselves what actions best promote their well-being?

2. What are the methods or tools by which governments can lead individuals to make the decisions that policy makers are confident will promote the individual's own well-being?

3. What are the moral pluses and minuses of these methods?

In answering these questions, this paper defines a domain in which certain kinds of paternalistic policies are feasible and attractive.

1 What Is Paternalism?

John Stuart Mill's *On Liberty* is often read as the classic text on paternalism, but I think that is a mistake. *On Liberty* is concerned with paternalism only insofar as it bears on individual liberty. Mill is concerned about the power that government and social norms exert over individuals, and he objects to the use of that power – whether for paternalistic or other reasons – except where directed toward the protection of other people.[1] Mill is concerned about freedom, and he does not consider whether there might be paternalism that does not limit freedom. He recognizes that public policies, such as posting

[1] Mill would, I believe, grant that there are other objectives than the protection of others that justify social coercion, such as achieving collective goods, avoiding cruelty to animals, or protecting places of natural beauty and interest.

warnings, may aim to direct individual actions that do not pose any risk of harming others, but since these involve no coercion, he has little to say about them. The term, "paternalism" had not yet been coined, and, as far as I know, there had been no exploration of the concept. Rather than a theory of paternalism, Mill offers an account of one kind of coercion: limitations on an agent's freedom, against her will, with the objective of benefitting her.

Although Mill in *On Liberty* argues against paternalistic coercion, he makes two much-discussed exceptions. First, he maintains that people should not be allowed to sell themselves into slavery. As Mill argues, this prohibition protects their future freedom. Nevertheless, it limits their current freedom; it is an instance of paternalistic coercion that Mill approves of. Second, Mill discusses stopping someone from crossing a dangerous bridge, when it is impossible to warn them effectively. A simpler example of the same sort of justified paternalistic coercion would be physically pushing a pedestrian crossing a road out of the way of a fast-moving truck he or she does not see.

One possible diagnosis of the pedestrian case, which would define a kind of paternalistic coercion that is morally permissible, is that interfering with actions that are not voluntary does not count as coercion or constitutes a morally acceptable form. On this view, it is permissible to push the pedestrian out of the way of the truck because the pedestrian did not voluntarily seek to be hit by a truck. As Feinberg (1971, esp. p. 112; 1986: 15) puts it,

> This seems to lead us to a form of paternalism that is so weak and innocuous that it could be accepted even by Mill, namely that the state has the right to prevent self-regarding harmful conduct only when it is substantially nonvoluntary or when temporary intervention is necessary to establish whether it is voluntary or not (1971: 113).

Richard Arneson copes with cases such as these by maintaining that even though the person crossing the bridge in Mill's example is "seized" and "turned back," these actions are not known to be against his will (1980: 487).

Most philosophical discussion has followed Mill in focusing on paternalistic coercion, and until recently, philosophical discussion of paternalism has taken coercion or a limitation of freedom as a defining feature of paternalism. During the last two decades, several prominent authors have rejected this requirement (for example, Shiffrin, 2000; Thaler and Sunstein, 2003a,b, 2006, 2008; Camerer *et al.*, 2003; Le Grand and New 2014). They deny that coercion is necessary to paternalism. For example, when one student, Martha, refuses to help another, Mary, with a math problem, believing that it is better for Mary to figure out how to solve the problem on her own, she acts paternalistically but without limiting Mary's freedom.[2]

[2]This example echoes one of Shiffrin's (2000, p. 213).

What then is paternalism, and how should one go about deciding how to characterize it? Although researchers are free to define their own technical notions, some of these definitions will be confusing and unhelpful. Richard Thaler's comments concerning the meaning of "paternalism" in his recent book, *Misbehaving* is an egregious example. "By paternalism, we mean trying to help people achieve their own goals. If someone asks how to get to the nearest subway station and you give her accurate directions, you are acting as a paternalist in our usage" (2015, p. 325). Of course, Thaler can use the word "paternalism" however he pleases, provided that he makes his meaning clear, as he does. But this usage makes it mysterious why anyone would have qualms about paternalism. At the same time that it encompasses an enormous range of behavior that few would call paternalistic, it mistakenly denies that actions that reject the goals of the agent in the pursuit of the agent's good count as paternalistic. Consider, for example, giving a blood transfusion to an unconscious Jehovah's Witness in order to save his life.

Seana Shiffrin maintains that "it is worthwhile to assess what is central in our normative reactions to paternalism and to employ a conception of paternalism that complements and makes intelligible our sense of paternalism's normative significance (2000, p. 212). In defining paternalism, Shiffrin points to a general question that encompasses narrower concerns about coercing people for their own benefit. That general question concerns whether the respect that we owe to one another requires that we defer to the judgment of individuals about how to pursue their own objectives, when their actions do not bear on the interests of others:

> The essential motive behind a paternalist act evinces a failure to respect either the capacity of the agent to judge, the capacity of the agent to act, or the propriety of the agent's exerting control over a sphere that is legitimately her domain.... Paternalistic behavior is special because it represents a positive effort by another to insert her will and have it exert control merely because of its (perhaps only alleged) superiority (Shiffrin, 2000, p. 220).

As Shiffrin points out, from such a perspective, refusals to aid as well as intrusive actions may appear to be paternalistic. For example, consider "tough-love" conservatives who would reduce unemployment insurance on the grounds that it discourages work and thereby harms the unemployed. If reducing unemployment insurance is justified in this way, the reduction would count as paternalistic. It would aim to benefit individuals against their will. If, as Shiffrin alleges, "what is distinctive and worrisome about paternalism" is "this substitution of judgment and the circumvention of an agent's will" (Shiffrin, 2000, p. 213), then we should seek a definition of paternalism that encompasses the benevolently intended reduction of unemployment insurance and, of course, a great deal more. As Gerald Dworkin puts it, "What we must ascertain in each

case is whether the act in question constitutes an attempt to substitute one person's judgment for another's, to promote the latter's benefit" (1988, p. 123).

Shiffrin defines a paternalistic action of A with respect to B as

(a) aimed to have (or to avoid) an effect on B or her sphere of legitimate agency

(b) that involves the substitution of A's judgment or agency for B's

(c) directed at B's own interests or matters that legitimately lie within B's control, and

(d) undertaken on the grounds that compared to B's judgment or agency with respect to those interests or other matters, A regards her judgment or agency to be (or as likely to be), in some respect, superior to B's (2000: 218).

In (c) Shiffrin allows for the possibility that a paternalistic action aims to assist B in the pursuit of some non-self-interested objective, but for the purposes of this essay, I shall narrow the notion of a paternalistic action in the standard way as directed toward B's benefit. If one revises (c) to assert that the action aims to make B better off, then one can omit both (a) and (d). If, as condition (c) now demands, the action aims to benefit B, then it must, as condition (a) maintains, aim to have an effect on B. (d) adds to (c) an explanation for why a paternalistic action is undertaken, rather than saying what a paternalistic action is. No such explanation is needed in a definition of what a paternalistic action or policy is. I thus think that the following much leaner definition of government paternalism suffices:

> *A policy is paternalist with respect to some agent if and only if it aims, for the benefit of the agent, to substitute the policy-maker's determination of what the agent should do for what is properly within the agent's own legitimate domain of judgment or action.*

This definition encompasses the traditional examples of sin taxes, prohibition, anti-sodomy laws, and so forth. It defines, however, a much larger domain, including removing barriers in people's way and shaping their behavior in many ways.

2 Meeting the Epistemic Demands of Paternalistic Policy

To devise public policies to influence choices that are properly within the legitimate domain of judgment or action of individual citizens for their benefit requires knowing (a) which, among those actions that are available to the

relevant individuals, they are likely to choose, (b) that there are better choices available to them, and (c) there is some public policy that can lead individuals to make choices that are better for them. Knowing that that there is an opening for a paternalistic policy does not, of course, imply that any specific paternalistic plan is morally permissible or, all things considered, beneficial. For example, before a law banning smoking in public places can be justified on paternalistic grounds,[3] the legislator needs to know that many people smoke, that it would be better for them if they didn't smoke or smoked less, and that there is some way, such as the proposed ban, to get people to smoke less. Knowing these three things is necessary, but it is not sufficient to justify a policy. Among other things, the legislator also needs to know whether a law banning smoking in public places violates moral prohibitions, such as those protecting individual rights, and what other consequences such a law may have.

In what circumstances will necessary conditions (a), (b), and (c) be satisfied?(a) rests on the answer to an empirical question: when faced with a set of alternatives, how do people tend to choose? This question may be hard to answer, but it poses no special problems. Learning how many people smoke or use their seat belts is straightforward. Moreover, behavioral economics has told policy makers a great deal about heuristics and foibles that characterize how people actually choose. Behavioral economics has also helped to answer (c) by identifying alternative methods of steering individuals toward making better choices. Deciding whether policies can improve on what individuals would otherwise choose also requires addressing (b), which is mainly a normative question. What choices make people better off? Policy makers can know that smoking is on average very bad for someone's health, but that fact does not show that smoking is, all things considered, the wrong choice for any particular individual. The pleasure smoking provides may sometimes outweigh the health risks it imposes.

Determining what is good for people is a special problem for liberals, who seek policies that are intended to be neutral among reasonable views of what constitute good human lives. Those who hold illiberal conceptions of the state often have an easier task. If their scripture dictates how people should live, then it is relatively easy for them to identify where people go wrong. It is much harder for liberals to be confident that they know better than individuals by what course of action those individuals can best promote their own well-being.

[3]There are non-paternalistic grounds in support of such a policy, too. It is often the case that policies and actions have multiple justifications. Martha's refusal to help Mary with her math may be intended to benefit Mary, but it also saves Martha trouble. In some cases, the non-paternalistic grounds for a policy are far-fetched, while in others it may be uncertain whether the policy is truly paternalistic or merely happens to benefit those whom it coerces. One might suggest a counterfactual test: a policy is paternalistic only if it would still have been carried out if there were no non-paternalistic reasons for it.

Economists have traditionally coped with this conundrum by supposing that people's choices reveal their preferences and that what satisfies their preferences makes them better off. These assumptions conveniently make paternalism impossible: since people choose whatever is best for themselves, interfering with their choices can only make their lives worse. However, as is obvious, these assumptions oversimplify a complicated reality. When people have false beliefs about the properties and consequences of alternatives or are poor judges of their own interests, they may easily prefer what is worse, and when they do not understand the alternatives among which they are choosing, their choices and their preferences may come apart. The hope of traditional welfare economics was that these complications, which can hardly be denied, could be treated as noise, of interest perhaps to philosophical pedants, but of no systematic importance. Behavioral economics has made this view harder to sustain, because it has identified systematic ways in which people's choices fail to reveal consistent preferences that can reliably indicate well-being.

In response to these findings of behavioral economics, some have hoped to find "true" preferences lurking within the scruffy preferences people express in their choices. Thus, Thaler and Sunstein seek validation of paternalist interventions from the individuals whose choices policy aims to improve. They aim 'to influence choices in a way that will make the choosers better off, *as judged by themselves*" (2008, p. 5, [their emphasis]). "We have no interest in telling people what to do. We want to help them achieve their *own* goals" (Thaler, 2015, p. 325). Le Grand and New maintain that justified "means related paternalism is concerned only with assisting in the achievement of ends that are considered to be fundamentally the individual's own—including the balance between these ends" (2014, p. 29). As Rizzo and Whitman put it (2009, p. 289), "The new paternalists are partially wedded to the principle of standard economics that an agent's welfare ought to be defined in terms of the goals or purposes of the agent himself." Other commentators on behavioral economics, such as Robert Sugden (2018), have suggested that normative economics should instead abandon its focus on welfare or well-being. In any case, it is impossible to conclude that it would be better for people if they choose x rather than y without possessing some way to compare the values of the two alternatives.

Liberal paternalists must judge that some choices are better for individuals than are others, while remaining neutral among comprehensive views of what constitutes a good life, and they cannot take choice as the criterion of well-being. What basis do they then have to judge the value of alternatives? Rizzo and Whitman (2009, p. 703) describe the dilemma as follows:

> The problem with context-dependence is similar to that of hyperbolic discounting: the new paternalist argument relies on an internal inconsistency to justify intervention. There is no theoreti-

> cal basis for choosing which behavior represents the individual's "true" best interest as he sees it. Which better represents a person's real preferences: what he is willing to pay for something or what he is willing to accept to part with it? There is no theoretically correct answer to this question, as Sunstein and Thaler admit: "If the arrangement of the alternatives has a significant effect on the selections the customers make, then their true 'preferences' do not formally exist."[4]

It seems to me hopeless to base public policy on "true" or "real" preferences. Even if these exist and it is possible for some close acquaintance to determine what they are with the help of psychiatric services, policy makers will never be able to determine them. Nor can economists or policy-makers turn to philosophers for enlightenment concerning what is best for individuals, because the theories of well-being philosophers have defended face serious problems and are not detailed enough to do the job.

I believe that those proposing paternalist policies in fact rely on (as they must) a "folk theory" of well-being. This theory lists states of affairs that are generally good and generally bad for people. It is made up of claims such as "Pain is generally bad." "Health is generally good." "Having good friends is ordinarily good." "The suffering of family members is usually bad." It includes some claims about priorities, too, such as, "Ordinarily, health matters more to well-being than sweets." A folk theory of well-being permits no fine discriminations, and it is not well suited to judge what is good for a particular individual. For example, the piece of gooey rich chocolate cake that Caroline, the "choice architect," places inconspicuously near the back of the cafeteria display may be much better today for a particular individual, Mortimer, than the fruit at the front. Having just lost his job, Mortimer may really need the comfort of chocolate. However, because health is *generally* a more important good than gooey desserts, paternalists know that they can make people better off by getting them to eat more fruit and less cake.

There is no feasible, let alone desirable, method whereby policy makers can come to know the detailed alternatives among which individuals choose or the detailed aspirations that direct their choices. Public policy can address failures of judgment only when they are general or systematic. There are three kinds of general or systematic judgment failures. First, there are certain kinds of choices that people are especially prone to flub. On the assumption that it is usually

[4]The quoted phrase is from Thaler and Sunstein (2003b), p. 1164. Le Grand and New (2014, p. 82) offer a partial response: "Put another way, we try to identify situations where an individual would agree to the following statements when she was confronted with a failure of reasoning which she had supposedly committed. Her decision was a non-trivial error that she would probably repeat in similar circumstances; the error was conceptual, not merely a verbal or technical misunderstanding; and she should have known the correct answer or procedure to find it."

better to be healthier, we can tell that people often make bad choices about what and how much to eat, whether to smoke, drink, or take narcotics, whether to take safety precautions such as wearing a seat belt, whether to take prescription medication, and so forth. Second, much of behavioral economics alleges that people possess cognitive foibles that in some contexts lead to systematic judgment failures. If people make different decisions depending on how a problem is framed, if they are lazy and prone to accept whatever is the default, if they are myopic and pay little attention to the distant future, and so forth, then it is possible that policies can help them to do better. The pervasiveness and even the existence of these flaws are however controversial, and this reason to favor paternalistic policies may be weaker than behavioral economists such as Richard Thaler maintain. Finally, although emotions often help people make good decisions, violent emotions interfere with wise decision making. Without knowing what the right choice for Mary might be, third parties can be confident that she is likely to choose less wisely when in a rage or panic. Certain subject matters and certain ways of presenting subject matters can provoke or calm emotions. There is apparently a space for paternalist policy making.

It seems to me that the discussions of ways to influence preferences and choices have perhaps focused too narrowly on cognitive crochets. Policymakers concerned to improve individual choice and preference should also address preferences that are harmful to their possessor, even if they reflect no evident cognitive failure. This thought steps over the line traditional normative economics has drawn, which prohibits economists from questioning the content of people's preferences. But however convenient that line may have been, by allowing choices to indicate preferences and preferences to serve as indicators of well-being, economists have refused to face the fact that people may choose alternatives that they know or should know are worse for themselves over alternatives that provide greater benefits to them. Even when people are rational, well-informed, and good judges of the value of alternatives to themselves and others, they may care about objectives they value more highly than benefitting themselves. Altruism exemplifies this possibility, and paternalistic policy might, in a much more pleasant world than ours, aim to limit it. Unfortunately, people may also sacrifice their own well-being in order malevolently to harm others. There is a bit of Iago in many of us. Even if also objectionable because of the harm they do to others, malevolent preferences should raise paternalistic concerns.

3 How can government help people to choose more wisely?

As discussed in Section 2, it is hard for government officials to know when there is an opening for paternalistic policies. In this section, I shall suppose that the epistemological difficulties have been surmounted and that with respect to

some set of alternatives facing a large portion of the population, policy makers know that people are choosing badly for themselves. The question this section is concerned with is, "What can the policy-maker do about this?" Knowing what people should do, government officials need to figure out how policies can help them to do it. What tools or methods should be used?

To answer this question requires considering what are the ways in which government can influence individual behavior, followed by an ethical appraisal of those methods and an assessment of their efficacy and drawbacks. There are many ways to influence choices. Coercion is perhaps the most obvious.[5] In its investigation of cognitive flaws, behavioral economics has pointed out other ways that do not limit freedom.

In *Nudge* and other works, Richard Thaler and Cass Sunstein call almost all the ways to influence individual decision-making other than coercion "nudges." Thaler and Sunstein are concerned with non-coercive ways of influencing individual choices, whether or not the objective is to benefit the individuals whose choices one is influencing. But most of their examples concern methods of helping people to make better choices for themselves, and I shall consider only nudges that aim to benefit those who are nudged. Thaler and Sunstein's concerns are wider than paternalism both because of the objectives for which nudging is employed and because some of the nudges they consider do not aim to preempt the agent's own judgment, and thus are not paternalist. For example, they take informing people to be a nudge, but providing information does not show disrespect or attempt to substitute for the agent's own agency.

Thaler and Sunstein maintain that a nudge is "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (2008, p. 6). In more recent work, Sunstein enlarges this characterization of nudges as follows:

> Nudges are interventions that steer people in particular directions but that also allow them to go their own way. A reminder is a nudge; so is a warning. A GPS nudges; a default rule nudges. To qualify as a nudge, an intervention must not impose significant material incentives (including disincentives). A subsidy is not a nudge; a tax is not a nudge; a fine or a jail sentence is not a nudge. To count as such, a nudge must fully preserve freedom of choice.... Some nudges work because they inform people; other nudges work because they make certain choices easier; still other nudges work because of the power of inertia and procrastination (Sunstein, 2015, p. 417).

[5] As Al Capone put it, "You can go a long way with a smile. You can go a lot farther with a smile and a gun." https://www.goodreads.com/quotes/272461-you-can-go-a-long-way-with-a-smile-you

Nudging is any way of influencing choices that preserves freedom of choice. So defined, nudges could include slipping LSD into someone's drink, leaving Gideons' Bibles in hotel rooms, or enrolling someone in a course in statistical inference.

It is unhelpful to lump together, as Thaler and Sunstein do, all the different ways of shaping choices other than coercion and "significant material incentives." Decisions about whether and how to carry out paternalistic policies require a more fine-grained taxonomy of freedom-preserving methods. I am far from the first to notice this. Grüne-Yanoff and Hertwig (2014) usefully distinguish nudges from what they call "boosts." Philippe Mongin and Mikaël Cozic distinguish three concepts, "In sum, nudge can mean: (1) an intervention that interferes with the choice conditions minimally; (2) an intervention that uses rationality failures instrumentally; and (3) a welfare-promoting intervention that tries to reduce the negative effects of rationality failures" (2017, p. 2). Although these senses do not exclude one another; neither do they imply one another, and it is perfectly possible for an intervention to be a nudge in only one of these three senses.

After one sets aside coercing people or giving them significant material incentives,[6] many alternative methods of "steering" people remain. Here is a list of eleven:

1. *Encouraging or discouraging them*: Providing weak or non-material incentives
2. *Informing them:* Providing relevant information
3. *Social influencing:* Describing what other people do
4. *Activating or inciting them:* Stimulating emotions to motivate individuals
5. *Cooling them off:* Calming their emotions and encouraging patient deliberation
6. *Educating or "boosting" them:* Influencing their deliberative capacities;[7]
7. *Deceiving them:* Providing false information
8. *Confusing them:* Employing fallacious arguments
9. *Brainwashing them:* Employing means such as subliminal images, drugs, or hypnotism.

[6]Since Thaler and Sunstein maintain that the provision of positive incentives such as subsidies do not count as nudges, I have left them off the list. But I do not understand why they exclude them. Subsidies and other positive incentives expand rather than restrict the choice set.

[7]Grüne-Yanoff and Hertwig call this "boosting".

10. *Focusing them:* Reconceptualizing the context and reframing the choice.

11. *Nudging them* (in a narrow sense): Changing the choice circumstances to neutralize or to exploit deliberative foibles.

All of these count as nudges in Mongin and Cozic's first sense, because they do not interfere appreciably with the conditions of choice, and depending on the motivation, all can be nudges in their third sense. On the other hand, only 8 – 11 and possibly number 4 make instrumental use of limits to rationality and thus count as nudges in Mongin and Cozic's second sense.

Despite some overlap, these methods of influencing behavior differ significantly from one another. Thaler and Sunstein call all of them "nudges," except possibly deception, which may create apparent limitations on the choice set facing an agent. I shall call this broad sense of "nudging" "non-coercive influence" to distinguish it from the narrow sense that I shall be concerned with. Although it risks confusion, I shall use the term "nudge," to refer only to the last of the eleven methods of influencing people to choose some action.[8]

I retain the term, "nudge," because the narrow construal of nudges as shaping the circumstances of choice fits the central examples that Thaler and Sunstein give. However, what concerns me is distinguishing among ways to influence choices, not the usage of the word "nudge." Rather than denying that informing or training counts as nudging, I could instead have distinguished informational nudges and educational nudges from choice-structuring nudges, and so forth.

If Thaler and Sunstein were merely pointing out that there are ways to influence people's behavior without coercing them, their work would be of no particular interest. Mark Antony needed no tutoring from Thaler and Sunstein on the possibility of influencing the actions of Roman citizens without coercing them. Thaler himself comments, "Swindlers did not need to read our book to know how to go about their business" (2015, p. 346). What makes Thaler's and Sunstein's *Nudge* noteworthy is instead their assertion that the flaws and heuristics behavioral economists have identified in human decision-making constitute both problems and opportunities for public policy. Thaler and Sunstein's contribution rests on what they have to say about nudges in the narrow sense. For example, in placing healthy foods prominently in a cafeteria line, Thaler and Sunstein's "choice architect," Caroline, is not encouraging or

[8]Compare this narrow construal of nudges to Hausman and Welch (2010, p. 126): "Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth. They are called for because of flaws in individual decision-making, and they work by making use of these flaws." Similarly, Selinger and Whyte (2011, p. 925) write, "Nudges are changes in the decision-making context that work with cognitive biases, and help prompt us, in subtle ways that often function below the level of our awareness, to make decisions that leave us and usually our society better off."

discouraging, informing, activating, cooling off, boosting, deceiving, confusing or brainwashing individuals. She is instead (in the narrow sense) nudging them. Setting the default in a pension plan nudges employees rather than cooling them off, or informing, activating, calming, boosting, deceiving, confusing, focusing, or brainwashing them.[9]

One size does not fit all. None of these methods will always be better than others. If they are effective and not too expensive, informing and educating individuals are especially attractive, because, unlike the other methods, they are not paternalist at all. Unfortunately, informing and boosting agents are also often ineffective. Tens of millions continue to smoke despite accurate warnings concerning the risks of smoking. The opioid epidemic in the U.S. does not derive from ignorance about the risks, and it cannot be cured merely by providing better information.

Although Thaler and Sunstein espouse nudging, often eliding the difference between influencing choices via choice architecture and the other non-coercive methods, they are offering those who want to carry out paternalistic policies a whole smorgasbord of non-coercive methods, with different advantages and disadvantages. Non-coercive ways to sway individuals are attractive, because coercion is generally undesirable and because legislators and bureaucrats cannot know what is best for each and every agent. Using methods that permit people to go their own way avoids harming those for whom the cigar after dinner is worth a shorter life span or for whom current consumption justifies a less affluent old age. Of course, at the same time, non-coercive policies risk failing to affect the behavior of those for whom smoking is the worse choice or those for whom greater retirement savings would be prudent. But, from a liberal perspective, where possible and not too costly, coercion is to be avoided.

Because smoking is addictive, it is hard to get people to quit. Informing people and training them to process information better are relatively ineffective. On the other hand, preventing smoking by coercive legislation is intrusive and costly. Encouraging, focusing, and nudging are thus the most attractive alternatives. Not wearing seatbelts, in contrast, is not addictive, and it is inexpensive and unintrusive to enforce seatbelt use. Coercion and informing

[9]Pelle Hansen (2016, p. 174) defends a similar but much more detailed concept of a nudge: A nudge is a function of (I) any attempt at influencing people's judgment, choice or behaviour in a predictable way, that is (1) made possible because of cognitive boundaries, biases, routines, and habits in individual and social decision-making posing barriers for people to perform rationally in their own self-declared interests, and which (2) works by making use of those boundaries, biases, routines, and habits as integral parts of such attempts. Thus, a nudge amongst other things works independently of: (i) forbidding or adding any rationally relevant choice options, (ii) changing incentives, whether regarded in terms of time, trouble, social sanctions, economic and so forth, or (iii) the provision of factual information and rational argumentation. I have no serious objection to the additional detail Hansen provides, with one exception. I think it is useful to consider as nudges choice architecture such as cooling-off periods that neutralizes cognitive biases without making use of those biases.

are more attractive methods of getting people to wear seatbelts than to get them to stop smoking.

4 The Moral Appraisal of Non-coercive Influences

In recent work (2015), Sunstein has returned to the question he and Thaler briefly discuss in *Nudge* concerning whether the various methods of influencing individual choices are ethical. More specifically, Sunstein asks whether influencing choices non-coercively but paternalistically promotes or undermines welfare, autonomy, and dignity (2015, p. 413). These questions are not well posed, because they have different answers, depending on which means of influencing behavior are employed. Providing information does not threaten autonomy or dignity. Brainwashing, deceiving, and confusing do.

Sunstein is aware that whether influencing choices non-coercively is ethical depends on the method employed to influence choices. He writes that whether "nudges intrude on autonomy" "depends on what kind of nudge is involved" (2015, p. 437). But Sunstein relies on the heterogeneity of the methods of influencing behavior to evade criticisms of nudging in the narrow sense. Consider, for example, Sunstein's response to the following claims Sarah Conly makes concerning nudges, "Rather than regarding people as generally capable of making good choices, we outmaneuver them by appealing to their irrationality, just in more fruitful ways" (2012, p. 30). Sunstein replies, "But she is making a strong charge, one that is not fairly leveled against most kinds of nudges. Recall that many nudges are educative" (2015, p. 446). But the fact that informing or educating people is unproblematic says nothing about the acceptability of neutralizing or exploiting deliberative foibles, and in his response to Conly, Sunstein does not address the question.

Informing people or educating them to be better deliberators is ethically unobjectionable, although these methods, like any others, can be employed toward iniquitous ends. Other ways of influencing behavior, such as brainwashing individuals or confusing them, raise red flags, regardless of the objectives that they aim to achieve. Is it ethically acceptable for governments to influence choices by nudging in the narrow senses – that is, by neutralizing or exploiting deliberative foibles? My concern is with the appraisal of methods of influencing behavior, not with the objectives to which these methods may be instrumental. The fact that it would be unacceptable to making voting for incumbents the default option (Sunstein, 2015, p. 416), says nothing about whether there is anything generally ethically problematic about influencing behavior by setting defaults.

In this relatively short paper, I cannot offer a careful assessment of each of the ways to influence choices without limiting freedom. I can say a few quick things. Although it requires argument, it seems clear that the government ought not to employ deception, confusion, and brainwashing in efforts to get

people to make better choices. On the other hand, informing, educating, and cooling off are generally unproblematic. Encouraging, activating, and socially influencing are more controversial. There is no sharp boundary between the provision of incentives or sanctions that limit freedom (which, *pace* Sunstein, need not be material – does he know nothing of Jewish mothers?) and those that do not. Emotional appeals can set in motion very destructive behavior. That leaves nudging and focusing. Although nudging structures the alternatives themselves, while focusing structures how individuals think about the alternatives, they raise similar issues, and for the rest of this section I shall discuss the assessment of nudges.

When the objective of nudging is paternalistic (and, as already mentioned, nudges need not be paternalistic), Thaler and Sunstein maintain that for nudges to be ethically acceptable, they must not be secret, and in addition the individuals who are nudged should agree that the actions they are nudged into choosing are better for them than the alternatives. This requirement of retrospective approval is problematic. First, it may have little teeth: As Sunstein notes, individuals may approve of having been nudged to choose alternative A, even though they would have approved of A′, which precludes A, if they had been nudged in a different direction (2015, p. 431). Second, asking people whether they approve of being nudged poses practical problems. Third, it would seem that the flaws in the ways people deliberate that show that what they choose is not always best also undermine the claims of preferences to determine whether paternalistic nudges succeed in making people better off. If people are not good judges of whether an alternative is good for them, why should they be good judges of whether a nudge has been good for them? If policy makers accept the findings of psychology and behavioral economics that make paternalistic nudging feasible, how are they to determine whether it is desirable?

The choice architect who seeks to nudge people takes them as they are. Rather than seeking to improve their deliberative capacities by informing them, boosting them, or getting them to cool off, the choice architect seeks to structure some set of alternatives people face so as to neutralize the effects of their deliberative flaws or to harness those flaws to get individuals to choose an alternative that the architect judges to be better. Provided that the objective passes ethical muster, is this method of influencing choices ethically acceptable?

There are two main arguments in defense of the ethical acceptability of nudging. First, Thaler and Sunstein argue that influencing people by the way that decisions are structured is unavoidable and hence that it must be permissible. However, even if influencing people's choices by one method or another is unavoidable, nudging (especially if it is intentional) is not. If a despicable politician wants to stir up support for anti-immigrant policies, inciting fear and resentment is a potent alternative to nudging. Other alternatives to nudging, such as cooling people off may not even count as steering them, because they do not influence them "in a particular direction." Nudging in

the narrow sense is not inevitable. Moreover, as I will argue below, there is an important difference between the "accidental" nudges that an unplanned choice structure provides and the calculated design of a choice structure in order to get people to choose a particular alternative. Calculated nudging is not inevitable.

A second argument in defense of nudging maintains that when it is advisable to influence people's choices, and unproblematic methods such as informing them, boosting them, or cooling them off are ineffective, then nudging people is better than coercing them, brainwashing them, deceiving them, confusing them, or stirring up their emotions. If it is important to influence choices, nudging may be better than any of the other ways to do so. The objections to nudging discussed below question whether nudging is entirely benign, but sometimes it may be more benign than any feasible alternative.

A number of authors have found nudges problematic (for example, Rizzo and Whitman, 2009, Conly, 2012, Grüne-Yanoff and Hertwig, 2014, Hausman and Welch, 2010, and Waldron, 2014). It seems to me that there are five main concerns about nudging:

1. Nudges put us on a slippery slope to a coercive paternalistic state.
2. Nudges are disrespectful. They treat people like children.
3. Nudges tend to perpetuate and to amplify deliberative flaws.
4. Nudges undermine autonomy.
5. Nudging is condescending and arrogant.

I do not find the first three of these objections particularly powerful. I am not in a position to judge the political risks of nudging. Even if they are serious, as Rizzo and Whitman argue (2009), the political risks do not show that nudging is itself unethical. The disrespect with which the second objection is concerned does not strike me as deeply ethically disturbing. If most people are myopic or subject to framing effects, then policy-makers are not singling anyone out when they structure choices so that people's myopia or their susceptibility to framing either assists them to choose well or does not hinder them from doing so. Respect does not require that policy makers pretend that people do not have deliberative crochets. If certain flaws are characteristic of human beings, is there anything disrespectful in recognizing and making use of them so that people are able to make better choices? Just as we put handrails and abrasive strips on steps, because we recognize that people are prone to fall, so we assign defaults to prevent the deliberative analogue to a fall. The case would be different if government policy singled out some social group (other than children and others with limited competence) as particularly prone to bad choices. Moreover, nudging individuals does not

preclude educating them, informing them, or cooling them off. If it is equally effective and not much more expensive, these other methods are superior, because they expand the competence of individuals. But the fact that there are sometimes better methods is not a general condemnation of nudging.

The third objection, that nudges keep people from learning how to avoid deliberative mistakes and that they encourage bad deliberative habits, rests on an empirical claim, for which there is some evidence (Fishbach and Trope, 2005), but not much. I am inclined to agree with Jeremy Waldron (2014) when he writes "I wish, though, that I could be made a better chooser rather than having someone on high take advantage (even for my own benefit) of my current thoughtlessness and my shabby intuitions." But even if informing and boosting people, if workable, would be better than nudging them, nudging may be a good thing.

The fourth objection – that nudging undermines autonomy – points to a serious worry. By "autonomy" I mean the control individuals have over their own evaluation, deliberation, and choice. Compare the following two policies. (a) An employer sets up a voluntary retirement plan, trains employees on how to make their own evaluations of the alternatives, and then requires that the employees choose. (b) An employer sets defaults and other features of the choice circumstances in order to get employees to contribute to their retirement what the employer believes to be a prudent proportion of their salary. The first employer is enabling the employee's choices without attempting to steer the employee toward a particular option. The second is attempting to control her employees. Her motive is benevolent, and she does not prohibit the choices she thinks mistaken, but she still attempts to control what the employees choose.

The reason why nudges, such as setting defaults, seem to be ethically problematic is that they aim to "push" individuals to make one choice rather than another without engaging in rational persuasion. The employee's freedom, *in the sense of what alternatives are available*, is virtually unaffected, but when this "pushing" takes the form of choice architecture, the autonomy of those who are nudged—the extent to which they have control over their own evaluations and deliberation—is diminished. As Hansen and Jesperson put it "... there seems to be a clear and important distinction to be made between a given context that accidentally influences behaviour in a predictable way, and someone – a choice architect – intentionally trying to alter behaviour by fiddling with such contexts" (2013, p. 10). The employees' actions derive in part from the tactics of the choice architect, rather than exclusively from their own evaluation of alternatives. When a benign employer, Marilyn, congratulates herself on engineering the situation so that the employees chose just the pension plans that she judged to be best for them, Marilyn is celebrating her power over her employees. At the same time (and this is the fifth criticism), nudging suggests the superiority of those who design the nudges or perhaps even their contempt for those whom they nudge. Waldron (2014) puts it this way, "For Sunstein's

idea is that we who know better should manipulate the choice architecture so that those who are less likely to perceive what is good for them can be induced to choose the options that *we* have decided are in their best interest."

This threat to autonomy could be grave. But I suspect that nudges have too little power to threaten autonomy seriously. For example, a recent experiment employed a panoply of nudges to improve people's compliance with medication regimes (Volpp *et al.*, 2017): "This was a kitchen-sink approach. It involved direct financial incentives, social support nudges, health care system resources and significant clinical management" (Carroll, 2017). Yet it failed to improve compliance. Obviously, a single study does not show that nudging has little force. But the results are suggestive. Compared to the enormous power of alternative methods of influencing people, such as social influence and inciting emotions of fear and hatred, which so easily mobilize violent mobs and mass irrationality, nudges are small potatoes.

It seems to me both that there is not that much to be said in general on behalf of nudging, except in those circumstances where the only effective alternative methods of influencing behavior are objectionable and that there is also not much to be said against them, because they pose only very weak threats to deliberative competence or self-respect, and it is unlikely that they will undermine autonomy in any significant degree.

On the other hand, the findings of behavioral economics should be alarming to normative economists, because they threaten the grounds upon which economists have evaluated alternatives. If preferences, especially as manifested in choices, do not reliably indicate what is good for individuals – as there was already good reason to believe before behavioral economics documented the deliberative peculiarities to which people are prone – then normative economics requires considerable rethinking.

References

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. "Regulation for conservatives: behavioral economics and the case for asymmetric paternalism". *University of Pennsylvania Law Review*. 151: 1211–1254.

Carroll, A. 2017. "Don't Nudge Me: The Limits of Behavioral Economics in Medicine". *New York Times*. November 6. URL: https://www.nytimes.com/2017/11/06/upshot/dont-nudge-me-the-limits-of-behavioral-economics-in-medicine.html?rref=collection%2Fbyline%2Faaron-e.-carroll&action=click&contentCollection=undefined&region=stream&module=stream_unit&version=latest&contentPlacement=3&pgtype=collection.

Conly, S. 2012. *Against Autonomy: Justifying Coercive Paternalism*. Cambridge: Cambridge University Press.

Dworkin, G. 1988. "Paternalism: Some Second Thoughts". In: *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press. 121–129.
Fishbach, A. and Y. Trope. 2005. "The Substitutability of External Control and Self-Control". *Journal of Experimental Social Psychology*. 41: 256–270.
Grüne-Yanoff, T. and R. Hertwig. 2014. "Nudge Versus Boost: How Coherent are Policy and Theory?" *Minds and Machines*. 26: 149–183.
Hansen, P. G. and A. M. Jespersen. 2013. "Nudge and the Manipulation of Choice. A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy". *European Journal of Risk Regulation*. 4: 3–28.
Hansen, P. 2016. "The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?" *European Journal of Risk Regulation*. 7: 155–174.
Hausman, D. and B. Welch. 2010. "To Nudge or Not to Nudge". *Journal of Political Philosophy*. 18: 123–136.
Mongin, P. and M. Cozic. 2017. "Rethinking Nudge: Not One but Three Concepts". *Behavioural Public Policy*. DOI: 10.1017/bpp.2016.16.
Rizzo, M. J. and D. G. Whitman. 2009. "Little Brother Is Watching You: New Paternalism on the Slippery Slopes". *Arizona Law Review*. 51: 685–739.
Selinger, E. and K. P. Whyte. 2011. "Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture". *Sociology Compass*. 5: 923–935.
Shiffrin, S. 2000. "Paternalism, Unconscionability Doctrine, and Accommodation". *Philosophy & Public Affairs*. 29: 205–250.
Sugden, R. 2018. *The Community of Advantage: A Behavioural Economist's Defense of the Market*. Oxford: Oxford University Press.
Sunstein, C. 2015. "The Ethics of Nudging". *Yale Journal on Regulation*. 32: 413–450.
Thaler, R. 2015. *Misbehaving: The Making of Behavioral Economics*. New York: W.W. Norton.
Thaler, R. and C. Sunstein. 2003a. "Libertarian paternalism". *American Economic Review*. 93: 175–179.
Thaler, R. and C. Sunstein. 2003b. "Libertarian paternalism is not an oxymoron". *University of Chicago Law Review*. 70: 1159–1202.
Thaler, R. and C. Sunstein. 2006. "Preferences, Paternalism, and Liberty": 233–264. Ed. by S. Olsaretti.
Thaler, R. and C. Sunstein. 2008. *Nudge*. New Haven: Yale University Press.
Volpp, K. G., A. B. Troxel, S. J. Mehta, L. Norton, J. Zhu, R. Lim, W. Wang, N. Marcus, C. Terwiesch, K. Caldarella, T. Levin, M. Relish, N. Negin, A. Smith-McLallen, R. Snyder, C. M. Spettell, B. Drachman, D. Kolansky, and D. A. Asch. 2017. "Effect of Electronic Reminders, Financial Incentives, and Social Support on Outcomes After Myocardial Infarction: The HeartStrong Randomized Clinical Trial". *Journal of the American Medical Association: Internal Medicine*. 177(8): 1093–1101.

Waldron, J. 2014. "It's All for Your Own Good". *New York Review of Books*. October 9. URL: http://www.nybooks.com/articles/2014/10/09/cass-sunstein-its-all-your-own-good/?printpage=true.