Review of Behavioral Economics, 2021, 8: 305–326

Libertarian Paternalism: Making Rational Fools

D. Wade Hands[*]

Department of Economics, University of Puget Sound, Tacoma, WA 98416, USA; hands@pugetsound.edu

ABSTRACT

This paper examines criticisms of libertarian paternalism, focusing in particular on so-called knowledge problems: the set of problems associated with the question of whether a choice architect would, or even could, have sufficient information to implement a successful libertarian paternalist policy. The paper builds on arguments presented in Mario Rizzo and Glen Whitman's book *Escaping Paternalism: Rationality, Behavioral Economics and Public Policy* (2020). Although the paper supports Rizzo and Whitman's arguments about knowledge problems, it moves in a different, more social, direction when it comes to the implications of these criticisms for microeconomic-based governmental policy more generally.

Keywords: Behavioral economics, nudging, paternalism, rational choice, welfare

JEL Codes: B4, D6, D9

> *That the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will make him happier, because, in the opinions of others, to do so*

[*]I would like to thank John Davis, Shaun Hargreaves Heap, and Ivan Moscati for helpful comments on earlier versions of this paper.

ISSN 2326-6198; DOI 10.1561/105.00000144

> *would be wise, or even right... The only part of the conduct of any one, for which he is amenable to society, is that which concerns others. In the part which merely concerns himself, his independence is, of right, absolute.* (Mill, 1961 [1859], p. 263)
>
> *The purely economic man is indeed close to being a social moron.* (Sen, 1977, p. 336)

1 Introduction

The initial round of behavioral economics literature – the so-called "heuristics and biases" program sparked by Kahneman and Tversky's (1979) paper on prospect theory – generated a significant amount of methodological debate. The main controversy was whether the heuristics and biases program was a better way to predict and explain individual choice behavior than rational choice theory: the *homo economicus*-based approach that has dominated economic theory since the end of the nineteenth century. While the heuristics and biases program did not provide a unified alternative to rational choice theory, it did produce a vast number of empirical anomalies – including status quo bias, preference reversals, hyperbolic discounting, constructed preferences, framing effects and many others – which demonstrated that actual behavior often deviated from rational choice theory in significant and systematic ways. In Richard Thaler's words: "The approach taken by most behavioral economists has been to focus on a few important ways in which humans diverge from *homo economicus*" (Thaler, 2017, p. 1800).

Despite the challenges that such anomalies posed for standard economics there has not been a paradigm shift in the way economists typically model individual economic behavior. Rather than a revolution, the methodological debate has mostly subsided and there now seems to be a relatively stable equilibrium, a peaceful coexistence, where most economists view behavioral economics as a complement to, rather than a substitute for, traditional rational choice theory (Angner, 2019). However, this stable equilibrium only concerns the positive science of choice theory. In recent years another, and in many ways more controversial, debate has opened up regarding the implications of behavioral economics for normative economics: welfare economics and microeconomic policy. The relationship between behavioral and welfare economics is generally called *behavioral welfare economics*, and the problem of trying to reconcile the various tensions between the two subjects is called the *reconciliation problem* (McQuillin and Sugden, 2012). There are many different aspects to the behavioral welfare economics debate, but the topic that has dominated the conversation more than any other on the policy side is *behavioral paternalism*, the focus of Mario Rizzo and Glen Whitman's *Escaping*

Paternalism: Rationality, Behavioral Economics, and Public Policy (Rizzo and Whitman, 2020, hereafter referred to as RW).

As Rizzo and Whitman explain, the "burgeoning field of behavioral economics has produced a new set of justifications for paternalism" and their book "challenges behavioral paternalism on multiple levels from the abstract to the conceptual to the pragmatic and applied" (RW, i).[1] Their remark that they "have presented a gauntlet of challenges to behavioral paternalism" (RW, 398) is somewhat of an understatement since the book is a relentless critical engagement with behavioral paternalism that attacks it on a wide range of different fronts. Although it is a multi-pronged attack, they do a very good job being systematic about each of the criticisms they discuss. For example, in chapter four they present various concerns about the empirical research on heuristics and biases, research that is the foundation for behavioral paternalism. And yet when they move on to other criticisms in later chapters, they set aside the material in chapter four and argue that even if the empirical research was unproblematic, various other issues would still be a problem. They proceed this way on a topic-by-topic basis. The result is "a series of 'even if' arguments" (RW, 399) that forms a systematic multi-pronged attack coalescing into a broad-based critique of behavioral paternalism.

Although the way Rizzo and Whitman's critical account comes together is impressive, I will draw on only certain aspects of their argument: primarily the first three chapters and chapters six and seven. There are a number of places where slightly different terminology will be used, but for the most part the paper will stick with Rizzo and Whitman's main story line from these chapters – although toward the end of the paper, my argument will turn in a different direction. Their main point is that behavioral paternalist policies are misguided and problematic, and the majority of their arguments are very persuasive, particularly those about the normative role of *homo economicus* and the knowledge problem of policy makers. But as they move through the book, they increasingly emphasize ideas from public choice theory – rational ignorance, government failure, slippery slopes, and such – and it seems that in the process they start to lose sight of the target – behavioral paternalism – and begin to slide into a more broad-ranging criticism of governmental action. In section four the argument will be made that while the vast majority of their critical remarks about behavioral paternalism are correct, such criticisms should lead, not in the direction of less governmental action *in toto*, but rather away from behavioral paternalism and toward increased governmental attention to the things that economists have traditionally considered to be the responsibility of democratic governments: that is *social* policy. In other

[1] RW refers to Rizzo and Whitman (2020) throughout.

words, to spend more time and resources on serious social problems rather than trying to nudge people into being more effective rational fools.[2]

2 Paternalism, Behavioral Economics, and the Inner Rational Agent

Paternalism within economics traditionally involved the use of taxes, subsidies, and regulations to change behavior in ways that would make the relevant individual better off than they would be on their own. For example, smoking cigarettes causes cancer and yet individuals choose to purchase and smoke cigarettes. Given they would be healthier, that is, better off, if they smoked less, a tax or restriction on the sale or use of cigarettes would, *ceteris paribus*, cause them to smoke less and therefore make them better off. Since economists have traditionally assumed that people make choices on the basis of utility/preference maximization, smokers and others who prefer to consume harmful products were not typically viewed as irrational (non-utility maximizing) by economists, but rather, simply as individuals making rational choices given unhealthy preferences. Thus, paternalist economic policies that raise the price, or restrict the sale or use, of products like cigarettes simply incentivize people to consume less of these products even though they are products the individuals prefer. This works in the same paternalistic way as incentivizing a child to consume less candy would make them better off, even though they would prefer to eat more, not less, candy. In the case of adults and products like cigarettes there has been some debate about whether the smoker's behavior is a result of tensions between their short run and long run preferences, or whether it was a simply a matter of maximizing preferences that are not good for them, but in either case the right (paternalistic) thing to do has traditionally been to use standard microeconomic tools to change people's behavior in ways that was good for them in spite of their preferences.

One major impact of behavioral economics was to break the traditional direct linkage between preference and choice. Kahneman and Tversky, following an established psychological tradition in behavioral decision theory, presumed that individuals possess stable and well-ordered preferences, but the overwhelming evidence from their own, and other, psychological research indicated that real people often make cognitive mistakes and fail to act optimally on their preferences. Identifying and helping individuals correct such mistakes was a main goal of this tradition in psychological research.[3] The

[2] The reference to rational fools, here and in the title, is from Amartya Sen's famous paper by that title where he argues that the standard economic view that choice is driven by a stable, highly structured, preference ordering, may make *homo economicus* rational in a very narrow sense, but such a person must also "be a bit of a fool." (Sen, 1977, p. 336)

[3] See Heukelom (2014) for a detailed discussion of the behavioral decision theory and its relation to the work of Kahneman and Tversky as well as later behavioral economics.

various factors – heuristics and biases – responsible for such mistakes is reflected in the laundry list of anomalies that is now associated with behavioral economics. From this point of view, all the various anomaly-creating cognitive mistakes constitute errors in rational decision-making. As Rizzo and Whitman explain:

> "The implicit metaphysical assumption ... is that each person has a neoclassical agent deep inside that is struggling to surface. Decision-making processes are thus deemed to be malfunctioning insofar as they fail to produce choices consistent with the standard preference structure. In other words, malfunction is not independently defined; it is whatever does not make standard choice theory descriptively accurate. This approach thus *assumes away* the possibility of individuals who simply do not have preferences that satisfy the neoclassical axioms ..." (RW, 80)

There are many ways that an individual can make mistakes and act in a non-preference maximizing way, but the general class of errors that has received the most attention are those related to the context of choice: the choice environment or choice architecture. Economists have traditionally viewed the choice space over which individuals maximize utility/preference to be the space of *outcomes* – bundles of various commodities in riskless choice and gambles in choice under risk – and thus in most economic models, the context of choice, say the way the vegetables are arranged in the grocer's display, will have no impact on the choices made. However, as many of the now well-known choice anomalies demonstrate, real people do often allow the choice context to affect their choices, and so alteration in the choice context or choice architecture has become the primary focus in efforts to help people correct their decision-making mistakes.

But choice context is a very general concept that applies to many variations that are quite different from the way things are arranged on a store shelf. For example, Kahneman and Tversky's influential paper on prospect theory focused on loss aversion – where an individual weighs losses more heavily than gains – and loss aversion is a type of endowment effect: the decision-maker values a particular outcome differently depending on their ownership or endowment at the time of choice. Suppose the individual starts at a particular bundle A with the associated utility level U(A). If bundle B is preferred to bundle A, then movement from A $\rightarrow$ B will increase the individual's utility, and the movement back from B $\rightarrow$ A will reduce their utility. But if the individual is subject to loss aversion, then the perceived gain from A $\rightarrow$ B will be less than the perceived loss from B $\rightarrow$ A. This means that the utility associated with bundle A will be lower after the path through B than it was initially. But given the definition of a function, this means that the individual's valuations of various bundles of goods cannot be represented by a stable utility function.[4] In this way the heuristics and biases program has led to (a) disconnecting preference from choice, and (b) opening the door to changing the choice architecture as a way of nudging individuals into making more rational, that is utility-maximizing, choices. Notice that through this lens, decision-making mistakes do not come from having irrational preferences – not transitive, not complete, etc. – but rather in failing to act optimally on one's preferences. Each individual has an inner rational agent, it is just their outer psychological shell prevents them from acting as *homo economicus,* or in short as an *Econ*, would act given such preferences. In other words: "behavioral welfare economics models human beings as faulty Econs" (Infante *et al.*, 2016b, p. 22)

We can now see how behavioral economics opens the door for a new behavioral paternalism that is quite different from traditional paternalism. In traditional paternalism choices are presumed to be consistent with the individual's preferences, it is just that people often prefer things that are not very good for them. Now we have the possibility that because of various cognitive mistakes, the choices that an individual makes need not be consistent with their underlying preferences. If one assumes that individuals are generally self-interested, a common presumption in economics, then such mistakes prevent people from acting rationally in their own self-interest and are thus worse off than they would be if they behaved like Econs. Thus changes in the choice architecture that would nudge individuals away from cognitive mistakes and back into proper utility-maximizing behavior would make them better off. The cognitive mistakes that real people make can be seen as *internalities* – the internal cost to the agent from less than fully-rational behavior[5] – and behavioral paternalism is designed to eliminate such internalities by nudging individuals into more rational choices and thus making them better off. Such a policy of course changes *homo economicus* from being simply a modelling strategy used to predict and explain individual behavior, to being a *normative standard* for behavioral paternalist policy: something that economic agents *ought to do* in order to be rational. In Rizzo and Whitman's words:

> "The behavioral paternalist case hinges crucially on the idea that people deviate systematically from rational choice. Their policies are intended to push people toward more rational behavior. In other words, despite having rejected rationality as a model of how

[4] By the way, this type of 'endowment effect' problem was identified by some early twentieth-century neoclassical economists, Pareto (2014 [1909]) for example, and it was one of motivations for the various efforts to develop a non-integrable demand theory during the 1930s: Allen (1932), Evans (1930), Georgescu-Roegen (1936), and others. See Hands (2011) for a detailed discussion of this literature.

[5] The term internalities was introduced in Herrnstein *et al.* (1993). Also see Loewenstein and Haisley (2018) and Bhargava and Loewenstein (2015).

> people *do* behave, the behavioral paternalist still accept rationality as a model for how people *ought* to behave." (RW, 16)

At this point it is useful to note that the preferences serving as the normative standard for behavioral paternalism cannot be just any preferences, or even just well-behaved (complete and transitive) preferences. They must meet much stricter conditions. First, they must meet all the typical conditions that economists require on preferences; for example, in the case of riskless consumer choice, preferences need to be complete, transitive, monotonic, and convex.[6] But they also need to be stable – remain the same for the time of the analysis, or in the case of behavioral paternalist policy, for the duration of the policy – and be context-independent. They are "the preferences that *would* determine choice in a sanitized or bias-free environment" (RW, 239). Finally, preferences cannot be constructed, revisable, or dynamic; in particular, they cannot be: constructed preferences (Lichtenstein and Slovic, 2006), revised by dynamic learning (e.g., Dold and Schubert, 2018), or emerge from process-based (Rizzo and Whitman, 2018) or inclusive (Rizzo and Whitman, 2020) rationality. The most common term for such preferences is *true* preferences although other terms are used, including: purified, latent, laundered, pruned, and spruced up. The topic has, as one might expect, generated a quite a debate.[7]

All of this brings us to the *libertarian paternalism* (LP) that is the topic of this paper. Although the literature on LP is extensive, the main focus will be on Sunstein and Thaler (2003) and Thaler and Sunstein (2003, 2009).[8] I will focus specifically on LP rather than Rizzo and Whitman's behavioral paternalism, even though LP is a particular type of behavioral paternalism. There are a number of reasons for this. One is that LP is by far the most influential brand of behavioral paternalism. Its influence can be seen in the number of citations, by the numerous examples that are discussed in the popular literature, and by Richard Thaler's Nobel Prize in 2017. But secondly, and more importantly, it is useful to focus on LP since there has been a rapid expansion of behavioral paternalist policies in recent years and many of the defining features of Thaler and Sunstein's original LP have been diluted; this even seems to be the case in later writings of Thaler and Sunstein, for example Sunstein (2016, 2018). Finally, I will focus on LP because even though Rizzo and Whitman use the broader term behavioral paternalism, the most obvious target for most of the chapters in *Escaping Paternalism* is LP.

[6] See the consumer choice chapter of any standard microeconomics textbook.

[7] It is often called the preference purification debate. See for example: Dold (2018), Hausman (2016), Infante *et al.* (2016b), Infante *et al.* (2016a), Rebonato (2012), and Sugden (2015).

[8] Although the *asymmetric paternalism* of Camerer *et al.* (2003) will also be included within the LP rubric.

It is best to start with Thaler and Sunstein's own account of LP. It is an example of the correcting-mistakes-through-changes-in-the-choice-architecture theme, but with its own unique style.

> "... a policy is 'paternalistic' if it tries to influence choices in a way that will make choosers better off, *as judged by themselves.* Drawing on some well-established findings ... individuals make pretty bad decisions – decisions they would not have made if they had paid full attention and possessed complete information, unlimited cognitive ability, and complete self-control." (Thaler and Sunstein, 2009, pp. 5–6)
>
> "Whether or not they have ever studied economics, many people seem at least implicitly committed to the idea of *homo economicus*, or economic man – the notion that each of us thinks and chooses unfailingly well, and thus fits within the textbook picture of human beings offered by economists ... But the folks that we know are not like that ... To keep our Latin usage to a minimum we will hereafter refer to ... Econs and Humans." (Thaler and Sunstein, 2009, p. 7)

The program is paternalist because it makes people better off than they would be on their own. Of course the presupposition is that being better off comes automatically when individuals behave like Econ. It is libertarian not only because it sill allows the individuals to make choices, but because the choice architecture is only re-arranged and the options stay the same. Thus the impact of the policy is always reversible.

LP can be applied to corporations and a wide range of other types of institutions, even families, and does not necessarily involve government policy. That said, I will focus exclusively on government policy since that is where the issues are most clear and where most of the critical attention has been directed.

LP has been explained and its background and motivation have been discussed. The next section moves on to critical concerns, but before moving on there is one additional point to emphasize about LP. The point is that despite the fact that many of Thaler and Sunstein's examples involve things like increasing organ donations and engaging various environmental policies, which seem to be inherently social, LP is in fact deeply pro-self, not pro-social, nudging. When the goal of the nudge is to make someone better off by better satisfying their own preferences, it is about individual rationality and not about social policies in the way economists have traditionally defined them. The concept of paternalism is about internalities, not externalities.

For example, suppose an individual has a preference for consuming goods that lead to a reduction in atmospheric CO_2, but because of various cognitive mistakes, ends up failing, on their own, to purchase goods that accomplish that goal. Now suppose an LP policy, designed to make this person better off by eliminating her/his cognitive mistakes, is successfully implemented. This would mean that the individual will be a more successful utility maximizer, and thus under standard assumptions, be better off, but it also means a lower level of CO_2 emissions, which is a positive social consequence. But this social benefit does not make the LP nudge into a social nudge. It is simply a pro-self LP nudge that coincidentally had a positive social impact. Social policies are aimed at, that is about, social things – social costs and social benefits[9] – and this policy was designed to help a particular individual do a better job maximizing her/his own utility. In other words, it was a nudge to correct for internalities, not externalities.

Given the confusion on this self versus social issue, borrowing some terminology from Barton and Grüne-Yanoff (2015) seems to be useful. In their terminology, the general notion of a nudge is: "an intervention on the choice architecture that is . . . behaviour-steering, but preserves the choice set and . . . does not significantly change the economic incentives" (Barton and Grüne-Yanoff, 2015, p. 343). But nudges come in at least two different forms, pro-self and pro-social. Pro-self nudges aim at steering people's behavior in private welfare-increasing ways, and pro-social nudges which aim at steering people's behavior in ways that internalize externalities or increase public goods (ibid., 344). Finally, LP is a special case of a pro-self nudge: "we suggest to characterize Sunstein and Thaler's (2003) libertarian paternalism as the advocacy of governmental use of pro-self nudges" (ibid., 344). Of course, a single nudge-based policy could have both pro-self and pro-social impacts, but the LP part would be exclusively pro-self.

I will use this terminology for the remainder of the paper. Rizzo and Whitman are close to this terminology, but they seem to be less explicit about the distinction between self- and social-nudges. For example, they note that there are social nudges, but that such policies "are not our primary target in this book" (RW, 22). This may be because some of their criticisms – in particular those I will emphasize – are most relevant to pro-self nudges, but some of their other criticisms can also be applied to pro-social nudges.

3 Knowledge Problems

There has been an extensive discussion of the knowledge problems, associated with LP. A sample of the critical literature that emphasizes these knowledge

[9] There are very good reasons, for certain problems, to move beyond the way that economists typically define the social – as the sum of the individual – but since this traditional approach is shared by the vast majority of economists, even behavioral economists, it is the way that the word 'social' will be used throughout this paper.

issues, in addition to work by Rizzo and Whitman, includes: Barton and Grüne-Yanoff (2015), Berg (2018), Berg and Gigerenzer (2010), Congiu and Moscati (2020), Davis (2011), Gigerenzer (2015, 2018), Grüne-Yanoff (2012, 2016), Grüne-Yanoff and Hertwig (2016), Guala and Mittone (2015), Hands (2020), Hausman (2016), Infante *et al.* (2016b), Infante *et al.* (2016a), Rebonato (2012), Sugden (2015, 2017, 2018), and many others.

In general the knowledge problem is that the policy maker – the choice-architect-in-chief (Rebonato, 2012) – cannot possibly know what they would need to know to do LP policy; they would need to know the individual's true preferences. Of course, standard economic theory is applied to microeconomic policy questions all the time on the basis of welfare-gains-and-losses. There are many different versions: the compensation principle, consumer surplus, special assumptions on preferences which allow economists to use a single representative agent, etc. – but all, at least in standard economics, are based on an individual preference satisfaction definition of welfare. So if standard economics can – to an acceptable degree – measure gains and losses in individual preference satisfaction, then why should it be any more difficult for behavioral choice architects to measure true preferences for LP-based policy? There are many reasons, but most important is that all of the standard techniques use the prices that people are willing to pay, or have paid, for the good to measure the increase or decrease in individual preference satisfaction. Consumer surplus, the old standby from the late nineteenth century, for example, is measured by a certain area under the consumer's demand curve. Such a technique is considered acceptable because economists assume that the consumer's demand curve comes about as the result of utility-maximizing behavior, which implies that price, what the consumer is willing and able to pay for the good, directly reflects the strength of the consumer's preference for the good. But this is not appropriate in the context of a LP policy. LP takes behavioral economics as its foundation, and the cornerstone of behavioral economics is that what people choose is frequently not what they really prefer, and this breaks the link between willingness to pay (i.e., price) and individual valuation (preference satisfaction).[10]

However, there is another, perhaps more serious, set of problems. Not only can economists no longer assume that demand-prices reflect preferences, the preferences that are needed by the choice architect are *true* preferences: the

[10] There is another way of seeing these problems. For example, consumer's surplus is defined by the consumer's demand curve and the quantity of the good. But heuristics and biases create internalities for the individual – differences between the benefits to the consumer if they did not make mistakes, and their benefits when they actually do make mistakes – and this means that what the consumer actually chooses to purchase at any price is off her/his demand curve (see Hands, 2020 for details). Individuals who are making rationality mistakes, either do not have demand curves, or have demand curves that are different from what economists have assumed about demand curves for over one hundred years.

preferences that would be reflected in the individual's choices if they made no mistakes. The choice architect needs not only access to some rough-and-tumble preferences, but preferences that are complete, transitive, insensitive to context effects, not constructed at the point of choice, stable for the duration of the policy period, and so on. It is just not possible to obtain the kind of detailed access to individual true preferences required for LP policy and it is no longer possible to use price as a proxy for the needed information. As Rizzo and Whitman make clear:

> "We are ... challenging the claim that paternalist policymakers could know what the true preference looks like when stripped of all bias. And knowing the true preference is a necessary input into deciding the correct policy intervention." (RW, 243)

But one would need more than true preferences to formulate an effective LP policy. One would also need detailed information about the particular causal mechanisms that are responsible for the cognitive mistakes that lead to sup-optimal decisions (Grüne-Yanoff, 2016). So, at least two necessary conditions are needed to design a LP policy: (a) detailed information about the true preferences of the relevant individuals, and (b) detailed information about the causal mechanisms at work in the relevant individual's outer psychological shell that are responsible for the cognitive errors. And the latter is just as important as the former and almost as difficult to obtain. Again Rizzo and Whitman:

> "Even if behavioral paternalists could discern people's true preferences that's not enough. To craft effective policies, they must also know the extent of people's biases, how much people have self-debased, how their biases interact and offset each other, and how policies may induce compensatory behaviors and substitution effects ... Without such knowledge, behavioral paternalists cannot reasonably hope that their policies will achieve their stated goal of improving the satisfaction of individual preferences." (RW, 18–19)

But, in addition to knowledge problems about both true preferences and the details of error-causing causal mechanisms, there is still another concern. As Rizzo and Whitman and many others have pointed out, just because the information that LP policy would require is not available, does not mean that choice architects are not implementing various LP-based policies. Since they do not actually have the detailed information they would need about either true preferences or the particular causal mechanisms driving mistakes, it seems that they must be using what *they believe* is reasonable to assume about both of these factors in any particular case. This educated-best-guess approach is reflected in the most of the applied LP literature, which almost never starts with details about an individual's preferences or her/his outer psychological shell, and uses this information to characterize an associated problem and propose an appropriate policy. Rather, the process typically involves more presumption and less detailed information. It begins with an 'obvious' concern, like eating fatty food, not saving for retirement, smoking cigarettes, making poor portfolio choices, etc. as a starting point – one that implicitly, but necessarily, makes a number of presumptions about both true preferences and the sources of particular mistakes – and then goes on to directly propose a LP strategy that might nudge the individual in the direction of making fewer of these 'obvious' mistakes. As Daniel Hausman explains:[11]

> "If the object . . . is to satisfy the purified preferences of the inner agent, then economists have to be able to find out what those preferences are . . . when behavioral economists such as Thaler suggest that cafeteria managers should put the cake in the back, they typically have very little detailed evidence. It seems instead that they believe themselves to be wise third parties, who know that fruit is better for almost everyone and who for that reason attribute a purified preference for fruit to most of those served by the cafeteria. But if the object is to satisfy purified preferences rather than to provide consumers with what the behavioral economist judges to be best for them, this is a precarious practice. Behavioral economists who believe that they promote well-being by satisfying purified preferences need to know what people's purified preferences are, not what they should be." (Hausman, 2016, p. 28)

Even if we grant the principle of charity and assume that no opportunism or conscious manipulation is going on – an assumption that Rizzo and Whitman seem doubtful about – such policies driven by the well-intentioned hunches of choice architects, seem to be along way from the science and evidence-based rhetoric of LP advocates.

Although these three knowledge problems constitute a powerful critique of LP, there are many other knowledge-based criticisms scattered throughout the LP literature. There is certainly no reason to try to note them all, but I will close this section by mentioning two other concerns. I am not noting them here because they play a prominent role in the literature, but rather just the opposite. I think both are important and yet they have not attracted as much attention as they should.

The first is the problem of second best. It is noted by Rizzo and Whitman (RW, 260–261), and a few others, Berg (2018) and Besharov (2004) for example, but it is not a common criticism of LP. The second best theorem

[11] Note: purified preferences = true preferences.

came long before the rise of behavioral economics: in Richard Lipsey and Kelvin Lancaster's 1956 paper. Lipsey and Lancaster's second-best result demonstrated that in the context of an optimization problem with multiple constraints "it is *not* true that a situation in which more, but not all, of the optimum conditions are fulfilled is necessarily, or even likely to be superior to a situation in which fewer are fulfilled" (Lipsey and Lancaster, 1956, p. 12). This means that unless the nudges correct for every heuristic or bias that is preventing optimization, then changing one or some of those factors, could move the individual farther away from the optimal choice. Admittedly, the second best issue can arise in any policy context where there are multiple problematic factors, but it seems that it is particularly troublesome in LP policy because choice architects almost never have detailed information about the exact causal mechanisms responsible for the mistakes, or their relative magnitudes, or the various feedbacks they might have on each other.

The second problem concerns the tradeoffs and interactions between LP policies and social policies. We are certainly well aware that individually optimal behaviors may cause externalities that spill over – positively or negatively – on to other individuals in the society. Given this, what is to prevent a pro-self LP policy nudging individuals in ways that lead to negative social impacts? It is not necessary of course that such impacts occur, or that they are negative, but they are always a possibly and generally ignored. Of course, even if such social effects were recognized, it is nearly impossible to know the direction or magnitude of such complex interactive changes. As Rebonato (2012, p. 234) notes, it may even involve LP microeconomic spillovers into macroeconomics. From a Keynesian perspective, an LP nudge that encourages people to save for tomorrow means less consumption and higher unemployment for today (and if the "paradox of thrift" is in effect, it could mean less savings in the future as well).

4 So What is to be Done?

Although this paper has emphasized some things a bit more, and de-emphasized others, most of the previous two sections is broadly consistent with what Rizzo and Whitman say in *Escaping Paternalism* (at least in the chapters I noted in the introduction). It is now time for a quick review and then for a change in direction.

The bottom line of the previous section is that because of fundamental knowledge problems, a successful LP policy is simply not possible if it is designed to nudge people into better maximizing their true preferences.

> "... paternalist policymakers face a severe, and possibly insurmountable, *knowledge problem.* They do not and often cannot possess

> the kind of knowledge needed to craft policy interventions that reliably improve human welfare." (RW, 237)

And many philosophers like Hausman agree:

> "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." (Hausman, 2018, p. 268)

Of course, if how one defines LP or behavioral paternalism becomes more amorphous – downplaying the commitment to *homo economicus*, true preferences, and pro-self nudges; allowing conventional microeconomic tools like taxes and subsides to sneak in; or loosening it up in other ways – it may be possible to dodge at least some of these specific knowledge-based criticisms. And that seems to be what has taken place in the writings of some LP supporters in recent years. But of course what you loose is the originality and uniqueness of the LP program that made it such an important contribution to economics, such a much-discussed policy tool, and worthy of a Nobel prize. But knowledge problems are clearly not all of the concerns that have been raised about LP. There are also autonomy problems, inconsistency/incoherence problems, concerns with the rhetorical style of Thaler, Sunstein and other proponents of LP, public choice criticisms (much discussed by Rizzo and Whitman), as well as many others.[12]

So suppose one is persuaded by all of these arguments that LP is deeply problematic, then so what? What is to be done? Rizzo and Whitman do not provide very much in the way of an answer to this question, although they do suggest a few places where we might begin to look for a better approach. They mention Gerd Gigerenzer's fast-and-frugal heuristics program as an alternative to traditional rational choice theory (Gigerenzer, 2002, 2008). They also note Gigerenzer's ecological rationality, the normative framework associated with the fast-and-frugal heuristics program (Gigerenzer and Sturm, 2012; Gigerenzer and Todd, 2012) as a more flexible evaluative standard than *homo economicus.* These are certainly reasonable resources to consider since Gigerenzer and associates have repeatedly argued for the advantages of fast-and-frugal heuristics over the heuristics and biases program both in general and with respect to LP in particular (e.g., Gigerenzer, 2015, 2018) and also because Gigerenzer's program is often involved in discussions about other non-Econ-based, yet behaviorally sensitive, policy approaches: for example,

[12] For example, Berg (2018) identifies nineteen different types of problems with behavioral paternalism.

"boosting" (Grüne-Yanoff and Hertwig, 2016).[13] It should also be noted that they discuss their own conception of "inclusive rationality" as a substitute for rational choice theory in *Escaping Paternalism* and in more detail in Rizzo and Whitman (2018). They certainly suggest some interesting and important research to consider, but it is discussed fairly briefly and no detailed theoretical or policy alternative to LP is offered.

Toward the end of the book, instead of a theoretical or policy program, one begins to get the idea that for Rizzo and Whitman behavioral paternalism is just one specific, currently fashionable, example of government ineffectiveness and overreach. This is a suggestion that makes appearances throughout the book, but clearly picks up steam in later chapters (particularly chapters eight and nine). The analysis seems to switch from criticisms unique to behavioral paternalism – such as concerns about the empirical foundations of behavioral economics and the inadequacies of *homo economicus* as a normative benchmark – to more generic criticisms of almost any type of governmental policy. The ideas that percolate through the discussion include public sector spillovers, the rational ignorance of voters, government failure, the biases of governmental policymakers, the influence of special interests, slippery slope arguments about the growth of government, and so on. In most cases these are arguments that can be applied to behavioral paternalist polices, but many of these arguments are, unlike the discussion of the knowledge problem, not unique to, or even generally associated with, either paternalism or behavioral economics. Many of them seem to be fairly general boilerplate indictments of government action:

> "In the rough-and-ready world of practical politics, policy is shaped in a maelstrom of idealism, activism, ignorance, time constraints, power struggles, and special-interest pressures. It would be genuinely shocking for real-world policies to resemble those imagined by hopeful academics." (RW, 310)

Although I would say that much of negative indictment of government is overstated, I also realize discussions about such matters run up against individual values – particularly regarding the trade-offs between welfare, justice, and freedom – and are seldom very productive. So instead of arguing against Rizzo and Whitman's position in the closing chapters, I want to go back to a point of common ground, but then take the discussion in a different direction.

[13] Gigerenzer's fast-and-frugal program has a wide following as alternative descriptive approach to individual choice behavior, although concerns have been raised about how his version of ecological rationality might work as a normative framework (Hands, 2014). For a direct exchange on fast-and-frugal heuristics versus heuristics and biases see Gigerenzer (1996) and Kahneman and Tversky (1996). See Lee (2011) for a detailed historical discussion of the similarities and differences between heuristics and biases and fast-and-frugal heuristics, as well as between Gigerenzer's version of ecological rationality and the ecological rationally approach associated with Vernon Smith.

The common ground is Mill *On Liberty* (1961 [1859]). Most of the text of the Mill quote in the opening epigraph to this paper also appears on p. 437 of *Escaping Paternalism.* As Rizzo and Whitman say, the quote is associated with what came to be called the *Harm Principle*: "the idea that we are justified in coercing people only for the purpose of preventing harm to others" (RW, 437). This principle was colorfully captured by Robert Nozick's remark that a person can leave their knife wherever they want "but not in your chest" (Nozick, 1974, p. 171). The problem is that from the two sentences Rizzo and Whitman cite from Mill, one would not necessarily get the importance of the "harm" part of the Harm Principle. Their quote only talks about what others, including government or society, cannot compel an individual to do. There is nothing in their selection of Mill's words about what government *can rightfully* do and that is to coerce individuals "to prevent harm to others." But since it is called the Harm Principle, it seems that the legitimate tasks of governments should be just as important as what it, or society in general, cannot do. Based on the version of the Mill quote in the epigraph, my own response would be to quit trying to do paternalist policy that equates paternalism with successful utility maximization and instead have the government spend its resources on the kind of activities that Mill argued they should be doing: *preventing harm to others.* In modern economics these things – at least in microeconomics – boil down to the traditional subjects for governmental action: positive and negative externalities and public goods.

Perhaps paternalist policies should be abandoned altogether, but if there is broad public support for such policies, make them traditional paternalism: using the best scientific evidence about what harms people the most and encourage them to do less of it. This could be done by traditional taxes or subsidies, direct regulation, social nudges, or some combination of these tools. If paternalism is going to be done at all it should be directed towards making people healthier, live longer lives, and so forth, rather than trying to uncover true preferences and the particular heuristics and biases that are preventing people from making optimal choices on those preferences. This is particularly the case because it would only achieve paternalist goals through the dubious causal chain of utility maximization → people get what they want → what people want is what is really best for them → therefore people get what is really best for them. First, it seems pretty clear that such a strict LP policy is essentially impossible to do, and second, following Mill, it is not something the government should be doing anyway. As the world grows more connected and densely populated, where truly social issues like climate change and pandemics become increasingly important, why should helping people maximize utility be a goal of government? Allowing people to get what they personally want seems to be what markets do best, and governments generally do not do very well. On the other hand, helping control things with massive negative externalities – again, the damage from climate change and pandemics comes readily to mind – is traditionally what governments can do and are supposed to do. My suggestion is thus to follow Mill by: (a) abandoning strict pro-self LP entirely, (b) keeping paternalism at a bare minimum, and when used, do it by targeting improvements in human physiology, not by trying to increase individual preference satisfaction, and (c) reaffirm the commitment to genuinely social concerns which could be approached using traditional microeconomic tools, or pro-social nudges, or both.[14] And remember all of these recommendations are not only based on Mill's vision but also what is and is not possible, or at least practical.

In closing I would like to briefly discuss two of the many recent proposals for improving behavioral economics-inspired policy and see how they compare to what I have suggested. The first approach begins where so many others begin, with the core knowledge problem behind pro-self LP. The second approach does not criticize LP but does suggest broadening and moving beyond narrowness of LP to a more integrated approach to microeconomic policy.

There are several authors who start with the knowledge problem but end up with somewhat different policy proposals. One such case is Francesco Guala and Luigi Mittone's "anti-welfarist" approach (Guala and Mittone, 2015). They interpret LP as welfarist rather than paternalist, because LP policies are evaluated on the basis of individual well-being which is equated, as in standard economics, with individual preference satisfaction. As they explain, LP is concerned with agents who "fail to behave according to their 'true' preferences" and LP interventions "are aimed at improving subjective well-being by replacing irrational agents with well-behaved ones," which makes LP-nudging "an ally of the neoclassical (welfarist) approach" (Guala and Mittone, 2015, pp. 387–388). The problem is, as clear from the above discussion, that such a welfarist approach, runs directly into various knowledge problems. Guala and Mittone's solution is to abandon LP-nudging with its single-minded normative commitment to *homo economicus*, but continue nudging policies with different goals. On one hand, make the paternalism more traditional (drop the goal of making people better subjective utility maximizers and focus on better objective outcomes), and on the other hand, focus on the externalities (note, externalities, not internalities) caused by systematic cognitive mistakes: "prevention of externalities: nudges are a cheap way of preventing later interventions that would be costly for the community" (ibid., 394). As they explain:

> "The idea, roughly, is that choice architects are often justified to intervene to protect other people from the damage that may be caused by irresponsible individuals. Nudge policies are not (or not

[14] Hargreaves Heap (2020) has similar concerns about Rizzo and Whitman's account, but argues for an increased emphasis on universal rules rather than a change in outcomes as a solution.

> only) for the good of the nudged, but for the good of third parties that otherwise are going to be harmed." (Guala and Mittone, 2015, p. 392)

The second set of recommendations I will note comes from George Loewenstein and co-authors, particularly Bhargava and Loewenstein (2015) and Loewenstein and Chater (2017). Loewenstein was an early and influential contributor to the behavioral paternalist literature – including Herrnstein *et al.* (1993) which introduced the idea of an internality and the asymmetric paternalism version of LP: Camerer *et al.* (2003). His recent contributions to the policy debate, unlike almost every author discussed so far, do not start from the position that LP has been a failure in any way. The application of LP "has enjoyed significant success" (Loewenstein and Chater, 2017, p. 27), it is just time to move forward and improve. These authors argue that behavioral economics-based policy "should not limit itself to proposing nano-size interventions that may not significantly address the more basic causes of the magnitude of contemporary policy problems" (ibid.), in particular, the focus on pro-self LP[15] policies has "overshadowed alternative ways in which policy can and should be informed by behavioral economics" (ibid.). In order to bring about this broader application behavioral economic ideas it is necessary to move beyond the narrow pro-self focus of LP and integrate traditional economic tools with pro-social nudging. The bottom line on the argument is thus:

> "Researchers in behavioural economics and practitioners of public policy should exploit a far wider and more nuanced range of ways in which traditional economics and behavioural economics can be combined... Moreover, understanding many of society's problems and formulating policy solutions will involve hybrids between traditional and behavioural economics, rather than pure application of either." (Loewenstein and Chater, 2017, p. 48)

Comparing these two sets of policy perspectives to my suggestions, it seems that Guala and Mittone's strategy is very close to what I have proposed. They are opposed to LP as policy-assisted individual preference satisfaction, and argue for a greater social focus. Perhaps getting enough information to target mistakes that have social externalities may run into some of the same knowledge problems as targeting individual preference satisfaction, but their approach seems to be in the spirit of what I proposed. The proposals from Loewenstein and co-authors also seem to be a movement in the right direction: moving away from pro-self nudging and focusing on social policy informed by a wide range of behavioral tools. I would exclude the congratulatory remarks about LP, because it seems like the policies that have been successful are more examples of what

[15] Note that Bhargava and Loewenstein (2015) use 'nudges' not in the broad sense it has been used in this paper, but rather as the term for LP, that is, pro-self, nudging policies.

these authors are arguing *should be done*, rather than applications of pure pro-self LP nudging, but again, the approach appears to be in the same spirit as my suggestions. The bottom line seems to be that there is (rightfully) less interest in policy designed to promote *homo economicus* – thus making more rational fools – and more interest in bringing a wider range of tools to bear on social concerns. *Homo economicus* is a very good tool for certain tasks in positive economic science, but it is a very poor goal for normative economic policy.

5 Conclusion

This paper has criticized LP theory and policy in general, but in particular, by focusing on knowledge problems: questions about whether a choice architect would, or even could, have sufficient knowledge to implement LP policy. These problems are discussed by Rizzo and Whitman and the paper builds on their discussion of these issues. The paper also draws on a fairly broad base of critical LP literature and also examines the historical origins of many of the relevant ideas. Although there is agreement with Rizzo and Whitman about their primary criticisms of LP, it draws less pessimistic conclusions about the role of ideas from behavioral economics in economic policy, particularly with respect to what economists have traditionally considered the reasons for social policy. The paper is less critical of behavioral economics and the role of government in general, but equally, or perhaps more critical, when it comes to the core idea that public policy should be aimed at nudging people into behaving like *homo economicus* and thus like Sen's rational fools.

References

Allen, R. G. D. 1932. "The foundations of a mathematical theory of exchange". *Economica.* 36: 197–226.

Angner, E. 2019. "We're all behavioral economists now". *Journal of Economic Methodology.* 26: 195–207.

Barton, A. and T. Grüne-Yanoff. 2015. "From libertarian paternalism to nudging – and beyond". *Review of Philosophy and Psychology.* 6: 341–359.

Berg, N. 2018. "Decentralization mislaid: On the paternalism and skepticism toward experts". *Review of Behavioral Economics.* 5: 361–387.

Berg, N. and G. Gigerenzer. 2010. "As-if behavioral economics: Neoclassical economics in disguise?" *History of Economic Ideas.* 18: 133–166.

Besharov, G. 2004. "Second-best considerations in correcting cognitive errors". *Southern Economic Journal.* 71: 12–20.

Bhargava, S. and G. Loewenstein. 2015. "Behavioral economics and policy 102: Beyond nudging". *American Economic Review.* 105: 396–401.

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.

Congiu, L. and I. Moscati. 2020. "Message and environment: A framework for nudges and choice architecture". *Behavioural Public Policy*. 4: 71–87.

Davis, J. B. 2011. *Individuals and Identity in Economics*. Cambridge: Cambridge University Press.

Dold, M. F. 2018. "Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics". *Journal of Economic Methodology*. 25: 160–178.

Dold, M. F. and C. Schubert. 2018. "Toward a behavioral foundation of normative economics". *Review of Behavioral Economics*. 5: 221–241.

Evans, G. C. 1930. *Mathematical Introduction to Economics*. New York: McGraw-Hill.

Georgescu-Roegen, N. 1936. "The pure theory of consumer's behavior". *Quarterly Journal of Economics*. 50: 545–593.

Gigerenzer, G. 1996. "On narrow norms and vague heuristics: A reply to Kahneman and Tversky". *Psychological Review*. 103: 592–596.

Gigerenzer, G. 2002. *Adaptive Thinking: Rationality in the Real World*. New York: Oxford University Press.

Gigerenzer, G. 2008. *Rationality for Mortals: How People Cope With Uncertainty*. New York: Oxford University Press.

Gigerenzer, G. 2015. "On the supposed evidence for libertarian paternalism". *Review of Philosophy and Psychology*. 6: 361–383.

Gigerenzer, G. 2018. "The bias bias in behavioral economics". *Review of Behavioral Economics*. 5: 303–336.

Gigerenzer, G. and T. Sturm. 2012. "How (far) can rationality be naturalized?" *Synthese*. 187: 243–268.

Gigerenzer, G. and P. M. Todd. 2012. "Ecological rationality: The normative study of heuristics". In: *Ecological Rationality: Intelligence in the World*. Ed. by P. M. Todd, G. Gigerenzer, and the ABC Research Group. Oxford: Oxford University Press. 487–497.

Grüne-Yanoff, T. 2012. "Old wine in new casks: Libertarian paternalism still violates liberal principles". *Social Choice and Welfare*. 38: 635–645.

Grüne-Yanoff, T. 2016. "Why behavioural policy needs mechanistic evidence". *Economics and Philosophy*. 32: 463–483.

Grüne-Yanoff, T. and R. Hertwig. 2016. "Nudge versus Boost: How coherent are policy and theory?" *Minds and Machines*. 26: 149–183.

Guala, F. and L. Mittone. 2015. "A political justification of nudging". *Review of Philosophy and Psychology*. 6: 385–395.

Hands, D. W. 2011. "Back to the ordinalist revolution: Behavioral economic concepts in early modern consumer choice theory". *Metroeconomica*. 62: 386–410.

Hands, D. W. 2014. “Normative ecological rationality: Normative rationality in the fast-and-frugal-heuristics research program”. *Journal of Economic Methodology.* 21: 396–410.

Hands, D. W. 2020. “Libertarian paternalism: Taking econs seriously”. *International Review of Economics.* 67: 419–441.

Hargreaves Heap, S. P. 2020. “The ‘problem’ is different and so is the ‘solution’”. *Review of Behavioral Economics.* (forthcoming).

Hausman, D. M. 2016. “On the econ within”. *Journal of Economic Methodology.* 23: 26–32.

Hausman, D. M. 2018. “Efficacious and ethical public paternalism”. *Review of Behavioral Economics.* 5: 261–280.

Herrnstein, R. J., G. F. Loewenstein, D. Prelec, and W. Vaughn Jr. 1993. “Utility maximization and melioration: Internalities in individual choice”. *Journal of Behavioral Decision Making.* 6: 149–185.

Heukelom, F. 2014. *Behavioral Economics: A History.* Cambridge: Cambridge University Press.

Infante, G., G. Lecouteux, and R. Sugden. 2016a. “‘On the econ within’: A reply to Hausman”. *Journal of Economic Methodology.* 23: 33–37.

Infante, G., G. Lecouteux, and R. Sugden. 2016b. “Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics”. *Journal of Economic Methodology.* 23: 1–25.

Kahneman, D. and A. Tversky. 1979. “Prospect theory: An analysis of decisions under risk”. *Econometrica.* 47: 263–291.

Kahneman, D. and A. Tversky. 1996. “On the reality of cognitive illusions: A reply to Gigerenzer’s critique”. *Psychological Review.* 103: 582–591.

Lee, K. S. 2011. “Three ways of linking laboratory endeavors to the realm of policies”. *European Journal of the History of Economic Thought.* 18: 755–776.

Lichtenstein, S. and P. Slovic. 2006. *The Construction of Preference.* Cambridge: Cambridge University Press.

Lipsey, R. G. and K. Lancaster. 1956. “The general theory of second best”. *The Review of Economic Studies.* 24: 11–32.

Loewenstein, G. and N. Chater. 2017. “Putting nudges in perspective”. *Behavioural Public Policy.* 1: 26–53.

Loewenstein, G. and E. Haisley. 2018. “The economist as therapist: Methodological ramifications of ‘light’ paternalism”. In: *The Foundations of Positive and Normative Economics: A Handbook.* Ed. by A. Caplin and A. Schotter. Oxford: Oxford University Press. 210–245.

McQuillin, B. and R. Sugden. 2012. “Reconciling the normative and behavioural economics: The problems to be solved”. *Social Choice and Welfare.* 38: 553–567.

Mill, J. S. 1961 [1859]. “On liberty”. In: *Essential Works of John Stuart Mill: Utilitarianism, Autobiography, On Liberty, The Utility of Religion.* Ed. by M. Lerner. New York: Bantam Books. 257–360.

Nozick, R. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Pareto, V. 2014 [1909]. *Manual of Political Economy: A Critical and Variorum Edition.* Ed. by A. Montesano, A. Zanni, L. Bruni, J. S. Chipman, and M. McLure. Oxford: Oxford University Press.

Rebonato, R. 2012. *Taking Liberties: A Critical Examination of Libertarian Paternalism.* New York: Palgrave Macmillan.

Rizzo, M. J. and G. Whitman. 2018. “Rationality as a process”. *Review of Behavioral Economics.* 5: 201–219.

Rizzo, M. J. and G. Whitman. 2020. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy.* Cambridge: Cambridge University Press.

Sen, A. K. 1977. “Rational fools: A critique of the behavioral foundations of economic theory”. *Philosophy and Public Affairs.* 6: 317–344.

Sugden, R. 2015. “Looking for a psychology for the inner rational agent”. *Social Theory and Practice.* 41: 579–598.

Sugden, R. 2017. “Do people really want to be nudged towards healthy lifestyles?” *International Review of Economics.* 64: 113–123.

Sugden, R. 2018. “‘Better off, as judged by themselves’: A reply to Cass Sunstein”. *International Review of Economics.* 65: 9–13.

Sunstein, C. R. 2016. “People prefer system 2 nudges (kind of)”. *Duke Law Journal.* 66: 121–168.

Sunstein, C. R. 2018. “‘Better off, as judged by themselves’: A comment on evaluating nudges”. *International Review of Economics.* 65: 1–8.

Sunstein, C. R. and R. H. Thaler. 2003. “Libertarian paternalism is not an oxymoron”. *The University of Chicago Law Review.* 70: 1159–1202.

Thaler, R. H. 2017. “Behavioral economics”. *Journal of Political Economy.* 125: 1799–1805.

Thaler, R. H. and C. R. Sunstein. 2003. “Behavioral economics, public policy, and paternalism”. *The American Economic Review.* 93: 175–179.

Thaler, R. H. and C. R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth and Happiness.* London: Penguin.