Foundations of Libertarian Paternalism: Normativity, Rationality, and Welfare*

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

December 2018
Version 1.6
15,326 Words

Abstract: There is an extensive critical literature within the philosophy of economics analyzing the libertarian paternalism of Richard Thaler and Cass Sunstein. This paper will examine many topics from this literature, but do so from a different perspective than most of the existing research. First, the main focus of the paper will be different. Recalling Thaler and Sunstein's distinction between Econs (those whose behavior will be unchanged by libertarian paternalist policies) and Humans (who will, at least potentially, change their behavior in ways that makes them "better off, as judged by themselves," Thaler and Sunstein, 2009, p. 5), this paper will take Econs seriously by focusing primarily on the particular features of the Econs who are not making mistakes rather than on correcting the behavior of Humans who are. Secondly, although the paper will draw on several ideas from the existing literature, the particular roles that these ideas play in the discussion will be somewhat different. Third, the normative character of the Econ reference point will be examined in closer detail by emphasizing the difference between normativity with respect to rationality, and normativity with respect to welfare/well-being. Finally, in the concluding sections, the question of the role of "the social" will be raised and how it concerns, and ought to concern, nudging and the associated policy prescriptions.

* Versions of this paper were presented at the "Norms and Normativity" conference in Lyon, France, June 27-29, 2018 and in a symposium at the University of Nice, July 3, 2018. Helpful comments were received from various members of these two audiences and as well as from Magdalena Malecka and Ivan Moscati.

> *Economists rarely draw the distinction between normative models of consumer choice and descriptive or positive models. Although the theory is normatively based (it describes what rational consumers should do) economists argue that it also serves well as a descriptive theory (it predicts what consumers in fact do). This … exclusive reliance on the normative theory leads economists to make systematic, predictable errors in describing and forecasting consumer choices. (Thaler, 1980, p. 39)*

0. Introduction

Although the ultimate impact of the behavioral turn that has taken place within economics during the last few decades is still unclear, what is clear is that there has been, so to speak, a disturbance in the force. Not only is behavioral economics now a well-established subfield within economics, various strands of behavioral research have combined with experimental economics, evidence-based economics, and a variety of other developments, to challenge many aspects of what was considered to be the rational core of modern economics only a few decades ago. Although these developments are important and on-going, the impact of the behavioral turn on positive economic science is not the main focus of this paper. The focus here will be on the *normative* aspects of these recent changes, particularly nudging-based behavioral policies and the associated behavioral welfare economics. These topics appear frequently within the contemporary philosophy of economics literature,[1] and have, at least to some extent, displaced the debate about the scientific adequacy of traditional rational choice theory versus behavioral economics.[2]

While trying to improve our philosophical understanding of this normative literature is an important undertaking, this paper will focus on only one part of this larger project: the *nudging* literature associated with so-called soft paternalism: the *libertarian paternalism* of Richard Thaler and Cass Sunstein (Sunstein 2015; Sunstein and Thaler

[1] A sample of this literature includes: Barton and Grüne-Yanoff (2015), Congiu and Moscati (2018), Davis (2011, 2018), Dede (2018), Dold (2018), Fehr and Rangel (2011), Fumagalli (2016), Gigerenzer (2015), Grüne-Yanoff (2012, 2016, 2017), Grüne-Yanoff and Hertwig (2016), Grüne-Yanoff, Marchionni, and Feufel (2018), Guala and Mittone (2015), Hagman, Andersson, Västfjäll, and Tinghög (2015), Hansen (2016), Harrison and Ross (2018), Hausman (2012, 2016), Hausman and Welch (2010), Hédoin (2015), Heidl (2016), Heilmann 2014), Heukelom (2014), Infante, Lecouteux, and Sugden (2016a, 2016b), Lepenies and Malecka (2015), Loewenstein and Chater (2017), Loewenstein and Haisley (2008), McQuillin and Sugden (2012), Mills (2015), Mongin and Cozic (2018), Nagatsu (2015), Rebonato (2012, 2014), Reiss (2013, Ch. 15), Rizzo and Whitman (2009), Schubert (2017), Sugden (2008, 2015, 2017, 2018), Sunstein (2013, 2015, 2018), and Whitman and Rizzo (2015).

[2] This normative turn is sufficiently well-recognized that historians of modern economics have begun to investigate its early origins. For example, Herfeld (2018) notes Cowles documents from the 1950s suggesting that "For actively improving decision making and avoiding inconsistent behavior, deviations from the rationality ideal would have to be detected to subsequently bring people's decision making on the 'rational' track" (p. 41).

2003; Thaler and Sunstein 2003, 2009) and the related work on *asymmetric paternalism* (Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin 2003). But unlike most of the libertarian paternalist literature, which emphasizes the advantages or disadvantages of specific real-world applications of nudging, the focus here will be primarily on theoretical and philosophical foundations. In addition, I will, for the most part, stay away from political and moral concerns such as individual freedom and autonomy, as well as social welfare concerns outside the traditional economic domain of individual preference satisfaction. This is in no way to suggest that applications of libertarian paternalism, or its implications for social and political philosophy, are not important topics – they certainly are and I will make some remarks about them at the end of the paper – it is simply to narrow the focus to a few key issues in the foundations of libertarian paternalism (hereafter LP).

I. Behavioral Economics, Libertarian Paternalism and the Reconciliation Problem

Ideas associated with contemporary behavioral economics have been traced back to Adam Smith (Ashrof, Camerer, and Loewenstein 2005) in the late eighteenth century and Herbert Simon's bounded rationality program of the mid-twentieth century (Camerer and Loewenstein 2004, Sent 2004), but a major impetus for the contemporary literature was Daniel Kahneman and Amos Tversky's paper on prospect theory in 1979. Their approach to individual decision-making was embraced by economists like Richard Thaler, who applied it to economic decision-making, giving birth to the *heuristics and biases program* within contemporary behavioral economics. The defining feature of heuristics and biases (hereafter HB) research has been to provide empirical evidence that actual human decision-makers frequently behave in ways that are inconsistent with rational choice theory; they make *mistakes* and these deviations from rationality are often systematic and repeated.[3] As Thaler recently put it: "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). The result has been a vast number of empirical *anomalies* – including loss-aversion, framing effects, endowment effects, hyperbolic discounting, anchoring effects, and many others – and while it is certainly possible to criticize some of this research, the sheer number of such results suggests they cannot be entirely ignored.

Although the behavioral-inspired debate over the scientific status of rational choice theory clearly raises important philosophical questions, in recent years discussions within the philosophy of economics have increasingly turned toward the *normative* implications of behavioral economics: in particular, the *reconciliation problem* of "how to

[3] This literature is too extensive to provide comprehensive references, but a few classics include: Camerer and Loewenstein (2004), Kahneman (2003); Kahneman and Tversky (2000), and Thaler (1980, 2000, 2018). See Heukelom (2014), Lee (2011), and Sent (2004) for historical discussion of this literature. Heukelom (2014, pp. 110-111) makes it clear that the emphasis on cognitive errors was a key part of Kahneman's research even before his collaboration with Tversky.

reconcile normative and behavioural economics" (McQuillin and Sugden, 2012, pp. 553-554).

One of the main targets for these philosophical discussions is the literature that will be the focus here: *nudging* and in particular *libertarian paternalism*.[4] LP begins from the position that individuals make *mistakes*, cognitive errors, but seeks to find ways to nudge these individuals back to more rational choices *without using coercion* or *incentive-based economic tools*. Since one of the main messages of the HB literature is that the choice context matters to outcomes, it is argued that the individual's choice environment – the choice architecture – can often be changed in ways that will nudge individuals into making more rational choices. As Thaler and Sunstein explain:

> In our understanding, 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 finding in social science, we show that in many cases, 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)

Two frequently discussed examples are the director of food services for a large school system who rearranges the way that cafeteria food is presented so that healthy items are more likely to be selected, and the corporation that changes its opt-in retirement plan to an opt-out system in order to increase plan participation of its workers (Thaler and Sunstein 2009). Notice that such changes in the choice architecture are *paternalistic* – they are changes designed to make students and employees better off – but they are also *libertarian* in that the less healthy food is still available and employees are still free to opt out of the company's retirement savings plan. As Camerer, Issacharoff, Loewenstein, O'Donoghue and Rabin (2003) noted, this is a substantive change from the way that economists have traditionally characterized paternalism:

> The scientific consolidation of psychological finding into a new brand of behavioral economic theory breathes new life into the rationales for

[4] LP is just one part of an extensive literature on *behavioral welfare economics*. There are many, partially overlapping, strands in this literature, but I will just note what appears to be the three most important approaches in addition to LP. The first is the neo-hedonist literature: efforts by Kahneman and others to restore the hedonistic utilitarianism of Jeremy Bentham by grounding welfare/well-being on directly experienced, rather than decision, utility (Kahneman and Thaler 2006, Kahneman, Wakker and Sarin 1997). Chetty (2015) provides an example of how this neo-hedonist approach might be used in a new synthesis with traditional rational choice and welfare economics (see Angner 2018 for criticism). Another strand of the behavioral welfare economics literature begins from a version of revealed preference theory and tries to build a welfare theory grounded in choice and consistency, rather than individual preference satisfaction (Bernheim and Rangel 2009, Bernheim 2016). Finally, and again, not totally disconnected from these other approaches, is the strand of behavioral welfare economics that links itself directly to recent developments in neurophysiology and neuroeconomics (Bernheim 2009, Fehr and Rangel 2011).

> paternalistic regulation … In a sense, behavioral economics extends the paternalistically protected category of 'idiots' to include most people, at predictable times. The challenge is figuring out what sorts of 'idiotic' behaviors are likely to arise … and how to prevent them, while imposing minimal restrictions on those who behave rationally. (p. 1218)

The changes in the choice architecture are designed so that individuals who are prone to HB-type mistakes will be nudged into more rational choices, while those who are not prone to such mistakes will not change their behavior as a result of nudging. Thaler and Sunstein have introduced particular terminology for these two groups; those who do not make such mistakes are called *Econs* (the *Homo economicus* of standard economic theory) and those to who do make such mistakes are called *Humans* (although *Homo Heuristicus* may have been a better choice). Again, Thaler and Sunstein:

> 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.
>
> If you look at economics textbooks, you will learn that *homo economicus* can think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi. Really. But the folks that we know are not like that. Real people have trouble with long division if they don't have a calculator, sometimes forget their spouse's birthday, and have a hangover on New Year's Day … To keep our Latin usage to a minimum we will hereafter refer to … Econs and Humans." (Thaler and Sunstein, 2009, p. 7)

Given these definitions, a LP nudge can be defined in terms of Econs and Humans: "In accordance with our definition, a nudge is any factor that significantly alters the behavior of Humans, even though it would be ignored by Econs." (Thaler and Sunstein, 2009, p. 9)

The exact character of Econs is key to the LP program since it carries both descriptive and normative weight. Descriptively it distinguishes the agents who will be, and those who will not be, affected by LP nudges; and normatively it distinguishes the agents whose decision-making should to be corrected for cognitive errors, and those whose rational decision-making is beyond reproach. Although it is clear that Econs and Humans are the foundations of the LP program, it is also clear they are both caricatures: idealized models of preference-based decision making. Econs are endowed with stable well-ordered preferences that (along with beliefs and constraints) are causally responsible for, and/or can systematically rationalize, the choices of individual Econs.[5]

[5] For the mainstream neoclassical tradition, stable well-ordered preferences define, and provide identity conditions for, economic agents. As John Davis explains:

On the other hand, Humans have Econs deep inside – an *inner rational agent* (Infante, Lecouteux, and Sugden, 2016a, p. 14) – but that inner Econ is seldom responsible for, and/or rationalizes, Human choices, because that inner rational agent is surrounded by a *psychological shell* of heuristics, biases, frames, and other factors which systematically prevent Humans from manifesting the preferences of their *inner rational agent*. As Infante, Lecouteux, and Sugden explain:

> ... ordinary human psychology is being treated as a set of forces that are liable to restrict the inner agent's ability to act according to the implications of its own reasoning. It is as if the inner rational agent is separated from the world in which it wants to act by a *psychological shell*. The human being's behaviour is determined by interactions between the autonomous reasoning of the inner agent and the psychological properties of the outer shell. However, in relation to issues of preference and judgement, the inner agent is the ultimate normative authority. (2016a, p. 14)[6]

In other words, both Econs and Humans have "an ideally rational agent skulking within" (Hausman, 2016, p. 26), but for Humans it is an inner agent "whom their actions betray" (ibid.). Thus even though both Econs and Humans are idealized agents, Econs are foundational since Humans are defined essentially as "faulty Econs" and normative analysis respects "the preferences of the imagined inner Econ" (Infante, Lecouteux, and Sugden, 2016a, p. 22).

Of course it is useful to note, at least in passing, that "Econs or Humans" is an incomplete disjunction; actual living flesh-and-blood humans need not be making decisions as circumscribed by either the Econ or Human model (or for that matter any preference-based model of decision-making). As Robert Sugden points out:

> When individuals' preferences are well ordered, they can be represented by a single unique utility function. According to the standard view, however, neither individuals' utility functions nor their preferences are changed by choices they make … Thus, once an individual is distinguished at one point in time in terms of a particular set of preferences and accompanying utility function, having those same preferences and utility function at a later point in time should in principle distinguish that selfsame individual. Accordingly … the neoclassical pure preferences view of the individual is that individuals have unchanged preferences. (Davis, 2003, p. 49)

[6] Most authors who discuss LP in terms of the tension between the inner rational agent and an error-producing psychological shell call this view the *preference purification account*. For example: Hausman (2012, 2016), Infante, Lecouteux, and Sugden (2016a, 2016b), and Sugden (2015). While the account defended below owes much to the position of these authors, there is one difference that makes the term "preference purification" inappropriate here ("reasoning purification" would be more accurate). This issue will be addressed below; this note is just to explain why the term preference purification was not used here.

> Despite Thaler and Sunstein's label, the decision maker described by this model, is not a Human in the ordinary sense of the word. It is a faulty Econ." (Sugden, 2017, p. 117)

The psychological, sociological, and biological literatures have many different ways of predicting and explaining behavior of individual *homo sapiens* which do not involve preference or utility in any way, and thus are quite different from either Econs or Humans. For example, actual humans may make the choices they do because their behavior is a result of the simple conditioned responses of the early behaviorist literature; or because they are nothing but robotic survival machines being propelled by the replication of their selfish genes (Dawkins, 1976); or perhaps behavior is structurally and culturally determined as with the *homo sociologicus* of traditional social theory; or perhaps it is because of the boundedly rational but not preference-based mechanisms that are the focus of the fast-and-frugal research program. [7] Of course these are just a few examples of the many possible explanations for why real people make decisions the way they do. The point is not to defend non-preference-based ways of explaining real human decision-making, it is simply to note that LP does not start with actual human behavior and try to predict or explain it; rather it starts with the narrow range of behavior defined by two specific types of idealized choosers: Econs whose behavior is the result of successful satisfaction of well-behaved and stable preferences, and Humans who also have such an inner rational agent, but whose behavior fails to achieve preference satisfaction because of interference from their outer psychological shell.

II. Econs, Humans, and Nudges: A Closer Look

It has been pointed out – quite correctly – that advocates of LP are often ambiguous about what exactly is, and is not, a LP-based policy (e.g. Grüne-Yanoff and Hertwig 2016, Rebonato 2012). Some of this is probably a result of the examples-driven style of the LP literature which often starts with a few examples of non-incentive-based and non-coercive policies that change behavior in ways that seem obviously good – better health, longer life, more savings, and so forth – and then conduct the rest of the analysis through the lens of these initial examples. Instead of starting with clear definitions and consistent philosophical commitments, and ending with policies that reflect those commitments, the process is the reverse; the discussion begins with certain (presumed to be obvious) exemplars of good policies and proceeds by identifying various theoretical and philosophical commitments that rationalize those exemplars: "building an 'ostensive' rather than 'axiomatic' definition of libertarian paternalism" (Rebonato, 2012, p. 6). Both approaches are valid of course, but both have potential pitfalls; for the

[7] The literature on the fast-and-frugal-heuristics program – also called the simple heuristics (SH) and the ecological rationality program – is quite extensive (see Gigerenzer 2008, 2015; Gigerenzer and Brighton 2009; and Gigerenzer and Selton 2001 for example). See Grüne-Yanoff and Hertwig (2016) and Grüne-Yanoff, Marchionni, and Feufel (2018) for a discussion of the relationship between boost and nudge policies (boost is to SH as LP is to HB), Hands (2014) for a discussion of the normative implications of SH, and Lee (2011) for a historical discussion.

former, the main difficulty is identifying adequate theoretical and philosophical foundations, and for the latter the main concern is the underdetermination problem (there are generally many different theoretical and philosophical positions that are consistent with the examples).

All this said, the presumption of this paper is that the philosophical discussion of LP would benefit from additional foundations-driven analysis. In particular, the hope is that by shifting the focus from exemplary applications to fairly narrowly defined Econs and Humans, perhaps we will better understand the foundations of LP.[8] The main reason is that *Econs are well-defined* and provide a relatively uncontentious point of departure for the analysis (not uncontentious with respect to either their scientific or normative adequacy of course, but with respect to their identity: what they *are* and what kind of agency and normative guidance they support). There has been relative stability within textbooks and the associated core commitments of economists about Econ agency since roughly the late 1940s; this means *there is a fact of the matter about what it is like to be an Econ*. It has been said that Econs and Humans are a "pleasant but obscure allegory" (Mongin and Cozic, 2018, p. 111), but the position taken here is that *what Econs are* – and thus *what Humans must be* – is the *least obscure* aspect of the LP literature. Given Econs as a starting point, the analysis need not proceed by establishing well-grounded theoretical and philosophical foundations, but rather it starts with Econ and considers the foundations of LP that can be sustained given the restrictions imposed by taking Econ as the normative standard for rational behavior.

The rest of this section and the following section will discuss a number of features of Econ that give us a better understanding of the foundations and limitations of LP. The first of these is the difference between *pro-self* and *pro-social* nudging (Barton and Grüne-Yanoff 2015, Hagman, Andersson, Västfjäll, and Tinghög, Gustav 2015). Pro-self-nudges are nudges designed to make agents more effective individual preference satisfiers and more likely to avoid HB mistakes, while pro-social nudges are designed to better achieve social goals. It is often pointed out that while Thaler and Sunstein's definitions make LP exclusively about (private) individual preference satisfaction, the examples they use often tend to cross the line between pro-self and pro-social nudges:

> A significant part of the nudge literature is directed at using behavioural insights to induce "behaviour change" in situations in which the targeted individuals do not seem to be making mistakes in satisfying their own preferences … they are simply frustrating the achievement of some public policy objective. For example, TS's [Thaler and Sunstein] catalogue of emulation-worthy policies includes nudges designed to reduce littering, to increase registration in organ donation programmes

[8] Although there are a number of papers in the literature that also emphasize foundational and clarification issues – a sample includes Hansen (2016), Heilmann (2014), and Mongin and Cozic (2018) – none start explicitly with the constraints imposed by Econ or end with the particular foundational relationships discussed here.

> and … to reduce the release of potentially hazardous chemical into the environment. (Infante, Lecouteux, and Sugden, 2016a, p. 5)[9]

While the distinction between pro-self and pro-social nudging does get blurred within some of the LP literature, the distinction between these two types of nudging nonetheless provides a clear analytical way to define LP relative to other types of nudging: *LP is purely pro-self-nudging* (Barton and Grüne-Yanoff 2015, p. 344).

Notice that defining LP in this way (as will be done for the remainder of this paper) introduces yet another way that LP involves idealization since any actual nudge, however pro-self the choice architect intended it to be, is almost certainly going to have some social impact. A purely pro-self-nudge – like a perfectly rational consumer – can be modeled, but it will necessarily involve a number of idealizations that will almost never be present in any real target application.

For example, suppose we observe junk food Fred consuming far more unhealthy foods than seems rational (based on the best medical advice). Even if we assume that Fred has well-behaved preferences, the relationship between those preferences and Fred’s junk food consumption is not clear. One case is that Fred is fully-informed about the health effects of such eating, but puts a very high value on the taste of food and has no particular desire to live a long life. So in this case Fred really does prefer to eat junk food and is acting rationally to satisfy his preferences, so there is no room for LP pro-self nudging. But even in this case there might be room for pro-social nudging. We live in a society with a myriad of interdependencies and thus a myriad of possible external effects, both positive and negative. There is a high probability that Fred will be unhealthy and require more medical expenditure than the average citizen over his lifetime. That extra cost will be paid in part by other citizens, either through higher insurance costs or higher taxes (or both). There are of course many other possible negative externalities (and there may even be a few positive externalities, say to the stockholders of companies that manufacture junk food). The point is that nudge-type policies might be used to change Fred’s junk food consumption for pro-social reasons, but if so it would not be LP-nudging.

Of course it is also possible that Fred’s inner rational agent really *does not want* to be eating so much junk food, but for whatever HB-based reasons, he keeps acting sub-optimally, and so in this case, he could be helped by LP nudging. The point is that in a purely observational case of imposing the nudge and observing Fred eating less junk food there is no way to know which of these processes was at work, and in addition, there is also the possibility that a bit of both, pro-self and pro-social nudging, were responsible. There is a way to make a clear distinction between pro-social and pro-self-nudges, but it is not by observing either pre-nudge or post-nudge behavior; it involves the specific goals of the nudger. If it is (solely) to make Fred a more efficient utility

[9] Also see Sunstein (2016) where survey questions include policies to "reduce pollution" and "encourage water conservation."

maximizer, then it is an example of LP; on the other hand if it is (solely) about externalities or public goods, then it is a pure pro-social nudge. But either way, the matter cannot be decided simply observing outcomes, one would need information about the intentions of the nudger.

This way of thinking about nudges also allows us to differentiate not only between the cases of LP and pro-social nudges, but it also provides some insight into *traditional, or "hard paternalist,"* nudging. So now consider Sally. Suppose Sally is fully informed, acts rationally, and really does prefer junk food, but now suppose she lives alone as a hermit and her eating habits impose no externalities on anyone else in society. In this case *neither* a LP-nudge (because she is acting rationally) nor a pro-social nudge (since there are no externalities) is needed,[10] but we still might want to change her eating behavior because eating junk food is – based on our best scientific evidence – *not good for her*. This is the *traditional paternalist motivation* for policy or other intervention; the motivation is what is really good for the person and has nothing to do with the individual's preferences or whether they are acting rationally given those preferences. In this case it may be possible to introduce a nudge that would change Sally's behavior in the direction of what is really good for her; such a nudge would make Sally objectively better off, but it would not be a LP-nudge, since it would change her behavior in a way that is contrary to the desires of her inner rational agent.

The bottom line seems to be that there are at least three kinds of interventions: LP pro-self-nudges, pure pro-social nudges, and traditional paternalist nudges. And of course there can be combinations of all three, as well as the possibility of various outcomes and motivations that are not identical to any of these. This means that a pure LP nudge will be a very difficult, if not impossible (see section V below) intervention to even identify, much less execute, if one is consistent with the Thaler and Sunstein definitions of Econ, Humans, etc. This means that LP policies will constitute a *very narrow class of interventions*: a class that is often inconsistent with what those sympathetic to LP say about the range of LP policy applications.

In closing this section it is useful to introduce the language of *internalities* – or internal externalities – originally introduced in Herrnstein, Loewenstein, Prelec, and Vaughn (1993). The idea of an internality mirrors the traditional idea of an externality. An externality is a difference between private and social cost or benefit. For example, a polluting firm incurs a particular private cost, but the negative externality of its pollution imposes costs on someone else in the society and the traditional solution has been to *internalize the externality*; in the case of the polluting firm, to make the firm pay the full social cost of producing the good. When the externality remains external they are acting rationally given their own private costs and benefits, but when it has been internalized, they are acting in a *socially rational* way.

[10] At least given traditional economic definitions of social costs and benefits (as sums of the private costs and benefits of the relevant agents) although some type of nudge or other policy might be available with some broader conception of the social.

Transferring this idea over to the behavior of an individual agent, the internality – the "within-person externality" (Bhargava and Loewenstein, 2015, p. 396) – is the cost associated with not behaving in a fully rational way. The mistakes that individuals make have costs to the individuals themselves and these costs are internalities. LP-based policies will nudge the agent into fully rational action, thus internalizing these internalities in precisely the same way that a tax or other environmental regulation would induce the polluting firm into fully rational social behavior. As George Loewenstein and Emily Haisley explain:

> Paternalistic policies have the goal of benefiting people on an individual basis … Whereas the conventional justification for government regulation is to limit *externalities* – costs people impose on other people that they don't internalize – to promote the public good, the justification for paternalism is to limit *internalities* – costs that people impose on themselves that they don't internalize … (Loewenstein and Haisley, 2009, p. 212)

Returning to Econs and Humans, it seems that Econs are internality-free Humans (or Humans are internality-plagued Econs) and LP-nudges are various ways to help Humans internalize their internalities and unleash their inner rational agent.[11]

III. What's It Like to be an Econ?[12]

Econ behavior is clearly behavior consistent with rational choice theory, but what exactly is rational choice theory? Rational choice has traditionally been seen as a particular version of *instrumental rationality* (using the most appropriate means to achieve given ends) that constrains instrumentally rational action in at least three specific ways. First, the ends or goals are *given* and remain constant throughout the analysis. The goal in the nudging literature is satisfaction of the preferences of the relevant economic agents. Secondly, the content of the given ends is entirely open. An agent can have the goal of killing or maiming others and set about to accomplish that goal in a perfectly (and hideously) rational way. It is also possible, as was the case for junk food Fred and Sally the hermit, that agents have perfectly rational preferences that do not coincide with what is really good for them or increases their objective well-being. This topic will be discussed in more detail below, but here the point is simply that

[11] One way to think about internalities and the differences between Econs and Humans is in terms of Thaler and Sunstein's two systems (two selves): the Automatic System (System I) and the Reflective System (System II) (Thaler and Sunstein, 2009, pp. 21-24). In this case internalities are the costs the Automatic System imposes on the Reflective System: costs that would be eliminated by successful LP-based nudges which "defend the *rational* System-II self from the damages (internalities) imposed by prevaricating and *irrational* System-I self" (Rebonato, 2012, p. 35).

[12] With apologies of course to Thomas Nagel (1974).

rational choice theory *alone* implies neither ethically normative behavior nor behavior that coincides with the objective well-being of the individual agent.

Thirdly, while the content of preferences is wide open, the structure of those preferences is not. Since preferences are the goal of instrumentally rational action, they must have sufficient structure so that the "most appropriate means" exist and can be identified. The core structural restrictions on preferences are completeness and transitivity. These are minimal conditions, traditional demand theory for example, adds restrictions such as convexity and monotonicity so that the resulting demand function is well-behaved. These assumptions will obviously vary from application to application, but the point is that they are restrictions on the *structure* of preferences and not the *content* of preferences. Having intransitive preferences makes one irrational, while having well-behaved preferences that are heavily weighted toward candy and fried food may be perfectly rational (just unhealthy). As Daniel Hausman and Michael McPherson put it: "People's preferences are rational if they are complete and transitive, and people choose rationally if their choices are determined by their preferences" (2006, p. 60). Econs and Humans both satisfy the first condition, but Humans often fail to satisfy the second.

Thus far we have been discussing the preferences of Econ as stable as well as complete and transitive. However there is a substantial amount of behavioral literature that suggests that preferences are not (even locally) stable, but rather are constructed in the context of specific choice situations.

> There are two major approaches in the literature on preferences. The first, dominant in mainstream economics, it that people have well-defined preferences … The second, dominant in psychology, is that preferences … are often constructed – not merely revealed – in the generation of a response to a judgment or choice task." (Grüne-Yanoff and Hertwig, 2016, pp. 170-171)

The first approach involves the stable well-ordered preferences of Econ, while the second approach is the product of the extensive behavioral literature on *constructed preferences*.[13] In general, preference construction is a complex and path-dependent process that is contingent on details of the particular choice situation. As Paul Slovic and Sarah Lichtenstein explained:

[13] See Lichtenstein and Slovic (2006b) for an collection of the most important research on constructed preferences. The constructed preference literature originated in the psychological research on *preference reversals* from the early 1970s (Lichtenstein and Slovic 1971, Lindman 1971). See Seidl (2002) for a survey or the preference reversals literature, Guala (2000) and/or Hausman (1992, Ch. 13) for philosophical analysis, and Heukelom (2014) for historical discussion. Preference reversals also produced a derivative debate over the so-called *discovered preference hypothesis* (see Cubitt 2005, Guala 2005 and Plott 1996).

> … the preferences themselves are determined not only by our knowledge, feelings, and memory but also by many aspects of the decision environment, including how the preference objects are described, … The variability in the ways we construct and reconstruct our preferences yields preferences that are labile, inconsistent, subject to factors that we are unaware of, and not always in our own best interests. Indeed … the very notion of a 'true' preference must, in many situations, be rejected. (Lichtenstein and Slovic, 2006a, p. 2)

Constructed preferences are indeed a challenge to rational choice theory and therefore to much of traditional economic analysis, but even if its supporters are entirely correct about preference construction, there seems to be no direct implications for LP, because *Econ, by definition, do not have constructed preferences*. Real people may well have constructed preferences – or no preferences whatsoever – but such people are not Econ and they cannot be nudged into being Econ, because there is no coherent way of talking about mistakes in rational decision-making unless there are stable well-ordered preferences to serve as the normative reference point. Nudging people to better fulfil their preferences when the preferences get constructed in the process of choice embroils the choice architect in an insoluble chicken-and-egg problem. If a Human's preferences were constructed within the context of choice there would be no way to design a nudge that would move them into better satisfaction of their preferences since their preferences would not come into existence until the agent was engaged in the choice process itself. No one can help you correct a mathematical error when the mathematical problem you are trying to solve only comes into existence when you begin the process of solving it, and keeps changing as a result of you working on it. But this argument extends to *any* type preference change, not just constructed preferences. LP-nudging is nudging toward the satisfaction of, or at least rationalizability by, a target set of stable preferences.

Both Econs and Humans have stable rational preferences and it is important to emphasize that those preferences *are stable* for the period of analysis in at least *four ways*: i) in the traditional way that economists have assumed stable preferences, i.e. they are not changing with respect to new information, interaction with other agents, advertising, etc., ii) preferences are *context independent* (they do not change with the choice context), iii) each agent has a single stable preference order (in particular the agent's preferences do not change as a result of the interactions of multiple selves within the inner rational agent) and iv) preferences are *not constructed* in the act of choice. The mistakes of Humans do not come from having something wrong with their preferences, but rather from their outer psychological shell that leads them to the wrong choices, given their preferences. This is just a result of what it is like to be an Econ, and in turn, what it is like to be a faulty Econ (i.e. a Human). The reference point of stable well-behaved preferences – often called latent, or true, preferences – is necessary for Econs to play the proper normative roll with respect to the mistakes of Humans. As

Sugden notes this "is why Thaler and Sunstein need the concept of latent preference – with all its problems" (Sugden, 2018, p. 11).[14]

So the conclusion is that while constructed preferences may well be an issue for real people making real decisions, LP's commitment to Econs as the proper normative baseline means that constructed preferences *play no role in LP theory or practice*. Since constructed preferences are often considered to be the most powerful critique that has emerged out of the behavioral economics literature, this means that LP – which supposedly puts behavioral economics to work in a serious way – turns a completely blind eye to one of behavioral economics most challenging insights.

Although it is clear that Econ preferences should have sufficient structure to make rational action possible – i.e. they should be complete and transitive – these restrictions, even when combined with stability and context-independence, do not seem to be sufficient for the preferences of the Econ that provide the standard of rationality for LP nudges. If public policy is going to nudge someone into being more rational in the sense of making choices that better satisfy their preferences, then it seems that, at least in some cases, the content, as well as the structure, of the preferences must be considered. Suppose someone is a sociopath or sadist that prefers to inflict pain and suffering on others. It hardly seems that nudging such a person into being more efficient at satisfying his/her preferences would be a good idea. In this case individual preference satisfaction and social welfare are in conflict. But while such issues are important they will be reserved for section V which brings social welfare into the discussion. At this point what Econ exhibit that Humans do not is simply *rationality* – rational preferences and the ability to behave in ways consistent with the satisfaction of those preferences – and that need not be the same as contributing to social welfare, or even (recall the examples of junk food Fred and hermit Sally) contributing to their own objective well-being. But even setting aside questions about well-being, it still seems that in order to better understand nudging that tries to "make choosers better off, *as judged by themselves*" (Thaler and Sunstein, 2009, p. 5) we need some restrictions on preferences that are more than the bare bones of completeness and transitivity. Different authors approach this issue in various ways, but the most common characterization is that Econs are endowed with *true preferences* and that pure LP-based nudging is nudging to move people in the direction of better satisfying their true or latent preferences.

To this end it seems that something needs to be added to link Econ's true preferences directly to being "better off as judged by themselves." This missing link can be filled by

[14] By the way, this is the reason for the comment about preference purification in footnote 6. The problem that Humans have is not that their preferences are impure in some way, it is that they make mistakes in act sub-optimally given their (rational and stable) preferences. What is impure is their *reasoning* or *optimizing effectiveness,* not their preferences. This means that their *preferences* can be *context independent* – that is, not be changed by their context, framing, endowment, etc. – at the same time that their *choices* are *context dependent*. They make context-induced choice mistakes, but they are mistakes in how they *behave/choose*, not in what they *prefer*.

self-interest: preferences that are exclusively self-regarding. If agents prefer that which they believe makes them better off and if what they believe makes them better off is what they prefer, then having such (rational) self-interested preferences and acting rationally on those preferences would mean that what people prefer is what makes them "better off as judged by themselves." As Hausman and McPherson explain:

> Start with the theory of rationality and add a common assumption of positive economics: that individuals are exclusively self-interested. If nothing but self-interest affects S's preferences, then S prefers x to y if and only if S believe that x is strictly better for S than is y. Rational and exclusively self-interested individuals always prefer that they believe to be better for themselves over what they believe to be worse. (2006, p. 64)

So the bottom line for this part of the story is that Econ have true preferences which are rational, stable, and context-independent, but also self-interested. If such an agent acts rationally on such preferences they will choose that which they believe will make them better off. Thus Econ are fully rational and make no HB mistakes in decision-making that would motivate or justify (pro-self) nudging.[15] Not only is this characterization of Econ preferences consistent with standard economics and much of the philosophical literature on LP, it is also consistent with most characterizations of why pro-self-nudging is needed: to "counteract cognitive and emotional barriers to genuine self-interest" (Loewenstein and Haisley, 2008, p. 215).

Of course accepting this characterization of Econ preferences is not the full story of what it is like to be an Econ. The missing piece – that which Humans lack – is to act rationally on those preferences. To make optimal decisions, the decisions they would have made "if they had paid full attention and possessed complete information, unlimited cognitive ability, and complete self-control." (Thaler and Sunstein, 2009, p. 6). But unlike specifying the necessary restrictions on Econ preferences, it is essentially impossible to document what exactly needs to be done to act optimally given those preferences. Mistakes can happen in an infinite number of ways – literally, the consumption of a particular good could be incorrect by 1 unit, or 75 units, or 103.765 units – but mistakes are not just about incorrect outcomes, they also involve incorrect beliefs, probabilities, miscalculation (for many reasons), i.e., because of all of the various HB mistakes that behavioral economists have identified. As a result, the ways that a Human can have the true preferences of an Econ but fail to act rationally on those preferences is extremely complex. Of course the number of ways that real humans can

[15] The assumption of self-interest also avoids all the thorny problems associated with altruism and/or malevolence. If A is altruistic toward B but is irrational, then a nudge that makes A more rational will make B better off. But this means that a LP nudge – supposedly purely pro-self – makes a Pareto improvement and also produces a positive externality. But maybe not. Maybe A is altruistic toward B but since this is based only on A's subjective judgements about what would make B better off and may not in fact do so, perhaps B is actually worse off. And such complexities go on and on. Real people are altruistic and malevolent and positive rational choice theory often needs to address such issues, but that is not the case for LP which is, as the name suggests, about paternalism and not about third party effects.

go astray is even greater. Real humans might not have preferences that are complete, or transitive, or stable, and in fact they might not have preferences at all. Even assuming that a real human being has preferences that along with beliefs and constraints determines behavior, those preferences could be altruistic or malevolent, or of the on-the-spot kind of constructed preferences, or involve many other factors that would make their behavior quite different from that of Econ, but could not be corrected by LP strategies from even the most well-informed choice architect. The next section offers a model that will allows us to explore some of these possibilities in a relatively controlled environment, but even in this restricted case, specifying all of the possible mistakes is not feasible.

IV. A Simple, but Clarifying, Special Case

As noted above, one aspect of the existing LP literature is that it often jumps from an analysis of the various parts of the LP argument – Econ, Humans, rationality, and such – to particular policy applications which are so complex that these analytical distinctions often get lost in the inevitable messiness of actual practice and real human behavior. The result is often a mangle (in the sense of Pickering 1995) rather than an increase in analytical clarity. In this section I will try to take an approach more associated with traditional economic analysis. Instead of looking to real world policy applications, I will look at a particular idealized model of Econ and Human decision-making where these points are clear and straightforward. The model that will be employed is in many ways the best exemplar of *homo economicus*; it is the backbone of twentieth century microeconomics and played a key role in textbooks and economists' intuition since the 1940s. It is the standard utility-maximizing budget-constrained consumer choice model. Granted most of the discussion of rational choice theory focuses on risky choice and expected utility theory, but consumer choice theory is a simpler case with cleaner analytical distinctions.

One reason is that decision-making under certainty simplifies the ways that mistakes can be made; using the standard terminology of behavioral anomalies, there are no judgement mistakes and all mistakes are choice mistakes,[16] but sheer reduction in complexity is not the only benefit of focusing on consumer choice theory. The second reason is that the type of mistake that is excluded (judgement) is actually less relevant with respect to the philosophical problems raised by LP since mistakes in calculating probabilities are, in general, more clearly *mistakes* in the common sense meaning of the term than the "mistake" associated with having true preferences, but not acting optimally on them. For example, mistakes in recognizing the relevance of the conjunction rule when making risky judgements – as in the case of the famous "Linda"

[16] "The field of behavioral decision research, on which behavioral economics has drawn more than any other subfield of psychology, typically classifies research into two categories: judgment and choice. Judgment research deals with the processes that people use to estimate probabilities. Choice deals with the processes people use to select among actions, taking account of any relevant judgments that they may have made." (Camerer and Loewenstein, 2004, p. 9)

example (Tversky and Kahneman 1983) – can be recognized and corrected in much the same way as correcting a mistake in addition or subtraction. This is not to say that getting real people to make such corrections is particularly easy – as many experiments attest – but the difference between correctly applying the conjunction rule and correctly adding numbers is simply a difference in *degree*; the difference between such judgement mistakes and the myriad of possible types of choice mistakes is a difference in *kind*. Thirdly, while judgement mistakes do not apply to consumer choice theory, all of the *choice mistakes* in the behavioral literature – reference-dependence, loss-aversion, framing, etc. – certainly do.[17] The carry-over to risk-free consumer choice is most clear in the extensive literature on endowment effects and loss aversion (e.g. Kahneman, Knetsch, and Thaler 1991; Knetsch 1989, 1992; Thaler 1980; Tversky and Kahneman 1991). Finally, and most simply, we can use consumer choice for the same reason that economists generally employ idealized models: to strip away the complexity of the situation in order to better identify the fundamental relationships and mechanisms.

In the certainty case, an Econ purchasing a set of goods $x = (x_1, x_2, \ldots x_n)$, facing fixed (competitive) prices $p = (p_1, p_2, \ldots p_n)$ and fixed money income (M) would satisfy his/her true preferences by solving the following, well-defined constrained optimization problem:

$$\text{Max } U(x)$$
$$\text{subject to: } \textstyle\sum_i p_i x_i = M,$$

where the utility function represents the agent's true preferences. Let's call this the Econ Consumer Choice Problem (ECCP). The solution to ECCP is a set of *n consumer demand functions*:

$$h_i = h_i(p, M) \text{ for all } i = 1, 2, \ldots, n.$$

Econs solve this problem perfectly while Humans have the utility function U(x), but fail to solve the problem correctly; they make mistakes. In the general case these demand functions will satisfy certain potentially observable comparative statics conditions[18] and making mistakes reduces to either not having demand functions or having functions that do not possess these properties. But in the case where the utility function is specified explicitly what is and what is not a mistake becomes even more clear.

For example, consider an extremely simple, two-good example with the utility function $U(x_1, x_2) = x_1x_2$ which generates the demand functions $x^*_1 = M/2p_1$ and $x^*_2 = M/2p_2$. In

[17] In fact, some of the anomalies identified in contemporary behavioral economics were recognized by early contributors to consumer choice theory (See Hands 2011).

[18] There are slightly different characterizations in the literature, but the h_is being homogeneous of degree zero, having negative substitution effects, symmetric cross-partial derivatives with respect to all prices, and a negative semi-definite substitution matrix are the most common (see any advanced microeconomics textbook).

such a simple case, making a mistake is crystal clear. If M = 10 and $p_1 = 1$ then the consumer is acting rationally/optimally if they choose $x^*_1 = 5$. On the other hand, with this price and money income, any other "choice," any $x^*_1 \neq 5$, is a *mistake* and the consumer is acting irrationally. This particular utility function is complete and transitive, but if we also presume it is self-interested and context-independent, then ECCP fully characterizes Econ behavior in this particular, highly idealized, case. So in this example, it is extremely easy to characterize Econs, Humans, and what LP would need to do. Econs choose $x^*_1 = 5$ and Humans choose any quantity of x_1 other than 5.[19] A Human with this utility function is making a mistake if they are not correctly maximizing utility by choosing the wrong quantity of the good.

But now let's generalize away from this particular two-good example, but stay within the ECCP framework. In this case Econ behavior is quite clear; *Econs will always "be on their demand functions"*; in other words, for any particular vector of prices and money income (p, M), Econs will choose $h_i = h_i(p, M)$ for all i = 1, 2, …, n and each of the h_i functions will satisfy the standard restrictions. Humans making mistakes, on the other hand, *will "be off their demand functions"*; in other words, in the world of this model, the condition that "each of us thinks and chooses unfailingly well, and thus fits within the textbook picture of human beings offered by economists" (Thaler and Sunstein, 2009, p. 7) simply means that each agent would always be *on their demand functions*. So given this, what does a LP-nudge do in this textbook world? *They move Humans on to the demand functions that they would have if they were Econs*.

Although this way of thinking about LP has not been a part of the recent discussions, it is not without precedent. While Thaler and Sunstein never frame the LP problem in this way, it was the way that it was framed in the original asymmetric paternalism paper (Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin 2003). They framed mistakes in terms of internalities and used the analogy of nudges getting individuals back to their optimal demand curves from their mistaken demands.

> When consumers make errors, it is as if they are imposing externalities on themselves because the decisions they make (as reflected by their demand) do not accurately reflect the benefits they derive. The goal of asymmetric paternalism is to help boundedly rational consumers make better decisions and align their demand more closely with the true benefits they derive from consumption. (ibid., p. 1221)

Not only did they frame the nudging problem in terms of being on the fully rational demand curve, they also emphasized, as above, that the problem to be solved by nudging is a mistake (i.e. in the decision-making process) and not the irrationality or instability of the agent's true preferences. They stress that not everything that appears to be irrational is irrational (the choice could be what the agent actually prefers like junk

[19] And assuming there is no mistake with respect to x_2, the internality cost is $|U(5, x^*_2) - U(x_1, x^*_2)|$.

food Fred or hermit Sally). The authors use the example of extended warranties. It may be that people make mistakes when they buy such warranties and they do not realize how unlikely such expenses are, but it may be that even fully-informed they would still do it (i.e. it is not a mistake for them), they just put a high value on peace of mind. These authors, unlike Thaler and Sunstein who seem to present LP-nudging as straightforward, note that such policies must be preceded by a careful investigation which sorts out these two possibilities: "in order to properly assess asymmetrically paternalistic policies, we must carefully address whether patterns of apparently irrational behavior are mistakes or expressions of stable preferences" (ibid., p. 1254).

Thus, it seems that thinking in terms of internalities and getting back to individual demand curves is not only a useful way to think about LP-nudging, it is also an approach that is more likely to inject a note of caution into the discussion of LP policies. But even with all this there is still a missing part to the characterization of Econ, Humans, and LP given here. It is clear what rational choice is, and clear what Econ and Humans are – and the distance both have (or at least could have) from real human beings – but at this point Econ choices are based only on what they prefer. This guarantees that Econ (and post-nudged Humans) will make fully rational choices "as judged by themselves," but it does not guarantee that those choices will make them objectively "better off." But questions about what makes people – even idealized agents – actually better off, and/or what makes the society better off, involves questions about welfare or well-being. It is finally time to turn to that issue.

V. Normativity, Rationality, and Welfare

This section will address two important questions that have been avoided to this point. These two topics are generally front and forward in philosophical discussions of LP and they were deliberately avoided here until the foundations were in place with respect to Econs, Humans, and LP. These two issues and associated questions are:

1. Econs act rationally on the basis of true preferences – they make no HB mistakes – but their choices only reflect what they prefer and not necessarily what really makes them better off, i.e. what increases their welfare or enhances their well-being. And yet, in several places, Thaler and Sunstein suggest that LP-nudging does more than simply make people behave rationally with respect to their own preferences; they also suggest that nudged Humans would be objectively better off (i.e. have increased welfare): "it is legitimate for choice architects to try to influence people's behavior in order to make their lives longer, healthier, and better" (Thaler and Sunstein, 2009, p. 5). So the question is: What is the relationship between behaving rationally with respect to true preferences, and actually being objectively better off?[20]

[20] The word "objectively" has been used several times above – as in "objectively better off" – but one should be careful with such expressions. I take a fairly common sense approach to the way expressions like "objectively better off" are being used. We all know that smoking cigarettes is objectively bad for us,

2. LP is generally viewed as one approach to behavioral welfare economics and welfare economics is about public policy related to social welfare. But LP only seems to be about making Humans more rational – making fewer mistakes in decision-making – and it doesn't say anything about the welfare of those who are nudged, much less about the distribution of resources, Pareto Optimality, or anything else that might be concerned with social welfare. So what is the relationship between LP-nudging and social welfare?

There is a lot to say here, but let's start with the answer to these two questions and then work through the various arguments supporting those answers. The answer to the first question is that LP-nudging doesn't *necessarily* make those nudged objectively better off, but it could make them better off, and the conditions under which that would be the case are the same conditions that also support the main results of Paretian welfare economics. The answer to the second question is that there is no necessary connection between LP-nudging and social welfare. LP-nudging is exclusively pro-self and thus, at best, about individual welfare, while social welfare is fundamentally comparative and social. As noted above, nudging techniques can be used to achieve traditional social goals, but that would be social nudging and not libertarian nudging.

Once the discussion turns from rationality to welfare or well-being, the question of how welfare is defined and/or measured immediately arises. Modern Paretian welfare economics is committed to an *individual preference satisfaction* based view of welfare "which assesses outcomes, policies, and institutions exclusively by how much they enhance or diminish welfare, as measured by the extent to which preferences are satisfied." (Hausman, McPherson, and Satz, 2017, p. 147). This individual preference satisfaction view has its origins in 18th and 19th century hedonistic utilitarianism, but differs from utilitarianism in several respects. Perhaps the most significant difference is that hedonistic utilitarianism is a *substantive* theory of welfare which defines explicitly what welfare *is* – it is hedonistic *feelings of pleasure and pain* – while individual preference satisfaction is a *formal* theory of welfare that "does not say what things are good for individuals, instead it says how to find out: by seeing what people prefer" (Hausman and McPherson, 2006, p. 119). Of course there are many other substantive theories of welfare – John Rawls' "primary goods" (Rawls 1971), the "capabilities" view of Amartya Sen (Sen 1992) and/or Martha Nussbaum (Nussbaum 2001), Robert Sugden's "opportunity" approach (Sugden 2004, 2010), views based on various lists of measurable outcomes (life expectancy, infant mortality, etc.), and a variety of others – that challenge the dominant individual preference satisfaction view, but at this point

but we also must realize this fact rests on our current scientific knowledge. Even given the massive amount of empirical evidence linking lung cancer to smoking, it is very unlikely, but possible, that it will be discovered that all of these studies contained a previously unrecognized problem that completely undermines this accepted scientific relationship. So objective, as it is used here means based on the best available scientific evidence, but with full recognition that, i) it has no necessary relationship to what a person may or may not prefer, and ii) what does and does not make someone objectively better off may change over time.

they remain minority positions. Given this, the rest of this discussion will assume the individual preference satisfaction view of welfare/well-being.

So returning to the first question above, how does one get from having true preferences and acting rationally on them, to choosing in such a way as to make oneself objectively better off? There may be many ways to close the circle connecting individual preference with individual well-being, but I will employ a slightly modified version of the approach taken by Hausman and McPherson (1996, p. 42; 2006, pp. 64-65). The key assumption to close the circle is *perfect knowledge*: agents have perfect knowledge about what does and does not make them better off. So putting together the various parts from previous sections with the assumption of perfect knowledge we have the following:

First Rational Choice Theory:

> (R1) Agents have true preferences
>
> (R2) Agents act rationally/optimally/in an instrumentally rational way given those true preferences

So (R1) + (R2) = Rational Choice Theory

Add two additional assumptions:
Self-interest (SI) and Perfect Knowledge (PK):

> (SI) Agents prefer x to y iff they believe x is better for them than y
>
> (PK) Agents have perfect knowledge about what does and what does not make them better off

Now putting (R1), (R2), (SI), and (PK) together we have:

> Agents choose what they most prefer and they prefer x to y iff x really makes them better off than y

So this completes the preference-individual-welfare identity. If agents have true preferences and act rationally on them their choices will be rational. If they are self-interested those choices will reflect what they believe is best for them (choices will make them "*better off, as judged by themselves*"). So Econs make choices that satisfy (R1), (R2), and (SI). If we add (PK) then the preference/utility maximizing behavior of Econs becomes the *own-welfare maximizing* agent of traditional welfare economics. Such agents have well-ordered preferences and act rationally on them by choosing what they prefer, but under (PK) these choices will actually make them better off. Such a world is a world where agents *never make mistakes in either rationality or their own welfare*; correspondingly they need no help in decision-making and are perfect judges of their own best interest. Now such an agent may still not be doing the right thing socially – they may be generating externalities (costs and benefits on others) and/or may be free riding on public goods – and they might need taxes, subsidies, or even a social nudge, to be behaving in a socially optimal way, but they are acting in an individually rational way and doing what is in fact best for their own well-being.[21]

So now let's start dropping assumptions. First let's drop (PK); the agent is still rational and self-interested but although they are making choices that rationally satisfy their true preferences, they may not be doing what is really best for them. *This is an Econ*; they are acting in a fully rational way and making no HB mistakes, but they might actually like to eat fatty foods or smoke cigarettes. They are the choosers in microeconomic textbooks; they may not take account of the social costs they impose on others, and they may have some preferences that most of us do not share and even find repugnant, *but this is what it is like to be an Econ*. Such Econ do not need LP-nudging, although they may need to pay some taxes or get nudged into more socially responsible behavior. And while they are doing what they prefer and what is rational given their preferences, they may not be doing things that actually make them better off (junk food Fred and hermit Sally being such cases).

Finally let's drop (R2); the agent has rational and self-interested preferences but does not choose optimally; they make mistakes in their decision-making. This is the Human, the agent who's outer psychological shell is preventing fully rational decision-making. This is an agent who could be LP-nudged into behaving more rationally.

So this provides the answer to the first question above. Successful LP-nudging doesn't necessarily make the agent objectively better off, but if they had perfect knowledge they would be better off after successful LP-nudging, and in addition they would be the type of agents that inhabit traditional welfare economics. Regarding the second question, the answer was provided in the last few paragraphs. Individual rationality and even adding (PK) so that agents not only choose rationally but also in a ways that makes them better off, has no direct relationship to socially optimal behavior. Socially optimal behavior involves consideration of social costs and benefits and there is no necessary reason for the combination of (R1), (R2), (SI), and (PK) to guarantee the agent takes such social considerations into account. Of course the agent may have a preference for (say) not littering, but while that is a possibility, it is certainly not a necessity. Individual rationality and welfare is one thing, while social rationality and welfare is something else.[22]

[21] Some readers may ask "Why not make Econs satisfy (R1), (R2), (SI), *and (PK)*, rather than just the first three? It may be splitting hairs since even the first three conditions would put extremely demanding restrictions on the decision-making – but (PK) is still too much to ask of even *Homo economicus*. Rational decision-making involves acting optimally given (R1) using all of the available information, not using all of the information there is. A consumer behaving optimally in the sense of ECCP need not know all of the information there is about the goods he/she is purchasing, only all of the information available (or perhaps all that it is rational to obtain); and profit-maximizing (and thus rational) firms need not know everything there is to know about the market, the inputs, the technology, etc., only all of the information available (or is rational to obtain).

[22] Of course in the case of a one-person economy then nudging Robinson Crusoe would not only make him better off, it would do so without making anyone else worse off, and thus it would be a Pareto

What all this boils down to is that if we take Econs and Humans seriously, LP-nudging is an extremely weak policy tool. In fact, as discussed below, it may even be impossible. But it is weak with respect to individual behavior in part because even if entirely effective (an issue discussed below) it only deals with an extremely small set of ways that the behavior of agents – either idealized agents or actual human beings – could deviate from the rationality of Econ: that is (R1), (R2) and (SI). LP-nudging is exclusively pro-self-nudging and strictly about rationality and not necessarily welfare, and most importantly, it only corrects for the narrow class of mistakes identified within the HB behavioral and psychological literature. Perhaps a large portion of human decision-making is driven by factors and mechanisms that are not based at all on beliefs, desires, and instrumental rationality. But even if folk-psychological beliefs and desires are behind much of real human decision-making – even perhaps consistent desires and epistemically warranted beliefs – the relevant causal mechanisms as well as the outcomes could still be quite different from those of Econ. Perhaps choice is driven by beliefs and desires, but preferences are intransitive, unstable, or constructed. These concerns emphasize the point that successful LP-nudging doesn't even correct for many of the anomalies identified within the behavioral economics literature. In addition, even in the case of a fully-equipped Human with well-ordered true preferences and making only HB-based mistakes, successful LP-nudging would only increase their objective well-being – "make their lives longer, healthier, and better" – under the heroic assumption of perfect knowledge. Finally, add to the fact that LP-nudging is exclusively self-nudging and need not have any direct connection with the traditional social issues that motivate most microeconomic-based social policy, and we have a very weak policy tool indeed.

But in addition of all the things noted in the previous paragraph, there is an extensive philosophical literature critical of LP that is concerned with issues that are often quite different from the ones discussed in this paper. There are far too many criticisms to attempt to cite or even classify, but let me note just two broad classes of concerns that have been emphasized within the existing literature. One of these categories can be called *Autonomy* problems, those generally associated with issues in moral and political philosophy: freedom, power, liberal values, manipulation, etc. This philosophical research includes: Barton and Grüne-Yanoff (2015), Fumagalli (2016), Grüne-Yanoff (2012), Guala and Mittone (2015), Hagman, Andersson, Västfjäll, and Tinghög (2015), Hausman and Welch (2010), Heilman (2014), Lepenies and Malecka (2015), Mills (2015), Mitchell (2005), Nagatsu (2015), Rebonato (2012, 2014), Rizzo and Whitman (2009), Schubert (2017), Whitman and Rizzo (2015), and others.

The second category of concerns can be called *Epistemological* problems, those generally associated with issues in epistemology, philosophy of science, and cognitive

improvement. But this is a very special case. In general, Pareto improvements put restrictions not only on an individual agent, but on what is happening to others as well.

psychology.[23] This research includes: Berg and Gigerenzer (2010), Grüne-Yanoff (2012, 2016), Grüne-Yanoff and Hertwig (2016), Gigerenzer (2015), Guala and Mittone (2015), Hausman (2016), Heilman (2014), Infante, Lecouteux and Sugden (2916a, 2916b), McQuillin and Sugden (2012), Rebonato (2012, 2014), Rizzo and Whitman (2009), Sugden (2008, 2015, 2017, 2018), Whitman and Rizzo (2015), and others.

Since the epistemological problem that gets the most attention was lurking in the background of the above discussion, it is useful to draw attention to it. It is what has been called the *interpersonal intelligibility of preferences* problem (Rebonato 2012): the problem that the nudgers/social planners simply cannot know what they would need to know – particularly the agent's true preferences – to design effective LP-nudges. As Hausman explains:

> If the object … is to satisfy the … 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 … preference for fruit to most of those served by the cafeteria. But if the object is to satisfy … 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 … preferences need to know what people's … preferences are, not what they should be. (Hausman, 2016, p. 28)

This problem – in some ways a LP-induced version of Lionel Robbins's problem of interpersonal utility comparisons (Robbins 1935) – is a fundamental barrier to the application of such policies.

Although a very wide array of concerns have been raised in both of these critical literatures, and some arguments certainly seem stronger than others, it is fair to say that the majority of this research is in general quite *consistent with the account of Econs, Humans, and LP-nudges* provided in this paper. Not only is the account given here consistent with the majority of these criticisms, it also identifies some new concerns such as emphasizing how few of the decision-making errors that are possible would be corrected by LP-nudging even if the epistemological problems could be overcome, as well as how few of the important insights of behavioral economics are actually

[23] It is important to note these two that these two categories are neither exhaustive nor mutually exclusive. They are not exhaustive because the critical literature is vast a comes at LP from many different directions. They are not exhaustive because may authors offer criticisms that involve both autonomy problems and epistemological problems (note the overlap within the two samples of research listed above).

addressed by LP-nudging. It also clarifies many of the important distinctions that are often blurred within the existing literature, such as: rationality versus welfare, preference satisfaction versus being objectively better off, Thaler and Sunstein's Humans versus real human beings, and individual rationality versus social rationality.

Finally, although there were various comments about pro-social nudges in the above discussion, it was always in reference to what they are not – that is, they are not LP pro-self-nudges – rather than any serious discussion of what they are, could, or should be. I will not attempt that here, but I would note that nothing said in this paper should be interpreted as a criticism of using nudge-like policies to address traditional social concerns, either as new tools or in combination with exiting taxes, subsidies, and regulations. And, I would add, individual nudges may be quite effective with people who have revealed that they are struggling with certain types of decision-making (doing things they would in fact prefer not to do) by say, purchasing things to help them stop smoking, or joining weight watchers, or going to a therapist who addresses such problems. In other words, the account of LP offered here is consistent with recent arguments for a more integrated view of both social policy and individual decision-making that includes various types of nudging along with other more traditional policies and solutions (Bhargava and Loewenstein (2015), Guala and Mittone (2015), Loewenstein and Chater (2017), and others). The problems associated with LP are not about the idea of nudging in general, they are about the idea, to put it bluntly, that the goal of policy is to help people get back on their individual demand curves *rather than* for addressing genuinely *social* problems.

Let me close this section with a specific example. Consider a relatively low impact environmental problem like littering. If we think of the problem solely in LP terms, it is only a problem if the people really prefer not to litter and their littering is the result of making various HB-type mistakes that lead them to generate non-utility maximizing levels of litter. If the problem is viewed strictly in LP terms the role of government would be to i) try to find out what the individuals in question *really preferred* with respect to litter, ii) discover what particular heuristic was preventing them from producing the rational amount of litter, and then iii) designing a particular nudge that would change the choice environment in such a way that it would lead them to generate less litter. By the way, if, during (i), the preference examination phase, it was discovered that the individuals in question really do *like to litter*, then as a pure LP-nudger, there is absolutely nothing that can be, or should be, done to change their behavior. Now consider the problem in a more traditional way. Litter is a negative externality, it imposes costs on others in the society, and since the cost is not paid by the people who litter, they tend to overproduce it unless there is some sort of disincentive to do otherwise. In this case the government has a direct reason to reduce litter – it imposes social costs on others in the community – and there is no serious epistemic problem – one *can see who is, and who is not, littering*. So the answer is simply to put a fine or tax on those who litter and the litter will consequently be reduced. And, it should be noted, the litter is reduced regardless of whether those who were taxed genuinely liked to throw trash around or whether they didn't really want to do it, but couldn't stop themselves because of the interference of their outer psychological shell. So which one of these sounds like a more reasonable policy approach to litter?

VI. Conclusion and Some Broader Remarks

Rather than simply summarizing the various arguments offered in this paper, I will close by responding to two potential criticisms of what has been said.

Some readers may find that my austere interpretation of Econs (and Humans) is unfair to those who support LP policies and approaches. After all, I have characterized the goal of LP-nudging quite narrowly, and yet the LP literature is shot through and through with stories about nudges that: achieve important social goals (not just satisfy individual preferences), increase objective well-being (not just promote a particular version of rationality), benefit a wide range of real human beings (and not just narrowly defined Humans), and seem warm and helpful (rather than cold and analytical). Shouldn't I pay more attention to the good they are trying to do and pay less attention to the specific things they say about Econs, Humans, and such? Well no, not really. In the previous section I endorsed the use of nudging techniques to address social issues and made it quite clear that the critical points of the paper were only directed at the way LP was originally characterized by Thaler and Sunstein and not at the idea of nudging in general. While the world might be a better place if nudging techniques were broadly applied to getting people to act more in the public interest or in ways that were actually good for them (whether they prefer it or not), the fact is that this would no longer be *libertarian paternalist* policy, and LP wouldn't be a new innovative approach to public policy. In later work Thaler and Sunstein, particularly Sunstein in survey and response-pieces like Sunstein (2016 and 2018), have often sounded like they didn't really mean what they said about Econ being the sole normative standard for LP-nudging. The fact is that using some fairly narrow notion of *homo economicus* as the standard for non-coercive and non-incentive-based paternalist policy *is what differentiates LP from all other approaches to policy*; it is one of the things that made it academically successful and politically popular, and if it becomes a more generic set of policy tools it disappears as a novel, or even specific, approach to microeconomic policy.

My second point is that although some readers may interpret my discussion as a general criticism of the use of rational choice models in positive economics, that need not be the case. It may be quite reasonable to characterize individual behavior in terms of acting optimally on stable well-behaved preferences for certain kinds of behavior, and yet not embrace rational choice theory as the sole universal standard for rationality. Recall that for most of the early neoclassical economists, as was the case for John Stuart Mill before them, economic theory was not applicable to all types of decision-making, only that in a particular domain of human action. The point of such remarks is not to make a case for a particular historical interpretation of rational choice, or to try to specify when exactly rational choice theory is, or is not, an adequate positive theory of individual decision-making, but rather simply to note that one could very well accept rational choice as a being an appropriate scientific framework for predicting and explaining the behavior of individuals and/or institutions in *certain contexts*, and yet not be willing to accept that turning people into *homo economicus* is the proper goal of public policy and thus treating those who live by any other normative standard as being fundamentally faulty and in need of corrective nudging.

References

Angner, Erik (2018), "We Are All Behavioral Economists Now," *Journal of Economic Methodology* (forthcoming).

Ashrof, Nova; Camerer, Colin F. and Loewenstein, George (2005), "Adam Smith's Behavioral Economics," *Journal of Economic Perspectives*, 19, 131-45.

Barton, Adrien and Grüne-Yanoff, Till (2015), "From Libertarian Paternalism to Nudging – and Beyond," *Review of Philosophy and Psychology*, 6, 341-359.

Berg, Nathan and Gigerenzer, Gerd (2010), "As-If Behavioral Economics: Neoclassical Economics in Disguise?" *History of Economic Ideas*, 18, 133-66.

Bernheim, Douglas (2009), "On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal," *American Economic Journal: Microeconomics*, 1, 1-41.

Bernheim, B. Douglas (2016), "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics," *Benefit Cost Analysis*, 7, 12-68.

Bernheim, B. Douglas and Rangel, Antonio (2009), "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*, 124, 51-104.

Bhargava, Saurabh and Loewenstein, George (2015), "Behavioral Economics and Policy 102: Beyond Nudging," *American Economic Review*, 105, 396-401.

Camerer, Colin; Issacharoff, Samuel; Loewenstein, George; O'Donoghue, Ted; and Rabin, Matthew (2003), "Regulation for Conservatives: Behavioral Economics and the Case for 'asymmetric Paternalism'," *University of Pennsylvania Law Review*, 151, 1211-1254.

Camerer, Colin F. and Loewenstein, George (2004), "Behavioral Economics: Past, Present, Future," in *Advances in Behavioral Economics*, C. F. Camerer, G. Loewenstein and M. Rabin (eds.), New York: Princeton University Press, 3-51.

Chetty, Raj (2015), "Behavioral Economics as Public Policy: A Pragmatic Perspective." *American Economic Review*, 105, 1-33.

Congiu, Luca and Moscati, Ivan (2018), "Message and Environment: A Framework for Nudges and Choice Architecture," *Behavioural Public Policy*, https://doi.org/10.1017/bpp.2018.29.

Cubitt, Robin (2005), "Experiments and the Domain of Economic Theory," *Journal of Economic Methodology*, 12, 197-210.

Davis, John B. (2003), *The Theory of the Individual In Economics: Identity and Value*. London: Routledge.

Davis, John B. (2011), *Individuals and Identity in Economics*. Cambridge: Cambridge University Press.

Davis, John B. (2018), "Behavioral Economics and the Positive-Normative Distinction: Sunstein's *Choosing Not to Choose* and Behavioral Economics Imperialism,' *Éthique et économique/Ethics and Economics*, 15, 1-15.

Dawkins, Richard (1976), *The Selfish Gene*. Oxford: Oxford University Press.

Dede, Çağlar (2018), "Behavioral Policies and Inequities: The Case of Incentivized Smoking Cessation Policies," *Journal of Economic Methodology* (forthcoming).

Dold, Malte (2018), "Back to Buchanan? Explorations of Welfare and Subjectivism in Behavioral Economics," *Journal of Economic Methodology*, 25, 160-178.

Fehr, Erst and Rangel, Antonio (2011), "Neuroeconomic Foundations of Economic Choice – Recent Advances," *The Journal of Economic Perspectives*, 25, 3-30.

Fumagalli, Roberto (2016), "Decision Sciences and the New Case for Paternalism: Three Welfare-related Justificatory Challenges," *Social Choice & Welfare*, 47, 459–480

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

Gigerenzer, Gerd (2015), "On the Supposed Evidence for Libertarian Paternalism," *Review of Philosophy and Psychology*, 6, 361-383.

Gigerenzer, Gerd and Brighton, Henry (2009), "*Homo Heristicus*: Why Biased Minds Make Better Inferences," *Topics in Cognitive Science*, 1, 107-43.

Gigerenzer, Gerd and Selten, Reinhard (eds.) (2001), *Bounded Rationality and the Adaptive Toolbox*. Cambridge, MA: MIT Press.

Grüne-Yanoff, Till (2012), "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles," *Social Choice and Welfare*, 38, 635-645.

Grüne-Yanoff, Till (2016), "Why Behavioural Policy Needs Mechanistic Evidence," *Economics and Philosophy*, 32, 463-483.

Grüne-Yanoff, Till (2017), "Reflections on the 2017 Nobel Memorial Prize Awarded to Richard Thaler," *Erasmus Journal for Philosophy and Economics*, 10, 61-75.

Grüne-Yanoff, Till and Hertwig, Ralph (2016), "Nudge versus Boost: How Coherent are Policy and Theory?," *Minds and Machines*, 26, 149–183.

Grüne-Yanoff, Till; Marchionni, Caterina; and Feufel, Markus A. (2018), "Toward a Framework for Selecting Behavioural Policies: How to Choose Between Boosts and Nudges," *Economics and Philosophy*, 34, 243-266.

Guala, Francesco (2000), "Artefacts in Experimental Economics: Preference Reversals and the Becker-DeGroot-Marschak Mechanism,: *Economics & Philosophy*, 16, 47-75.

Guala, Francesco (2005), "Economics in the Lab: Completeness vs. Testability," *Journal of Economic Methodology*, 12, 185-196.

Guala, Francesco and Mittone, Luigi (2015), "A Political Justification of Nudging," *Review of Philosophy and Psychology*, 6, 385-395.

Hagman, William; Andersson, David; Västfjäll, Daniel; and Tinghög, Gustav (2015), "Public Views on Policies Involving Nudges," *Review of Philosophy and Psychology*, 6, 439-453.

Hands, D. Wade (2011), "Back to the Ordinalist Revolution: Behavioral Economic Concerns in Early Modern Consumer Choice Theory," *Metroeconomica*, 62, 386-410.

Hands, D. Wade (2014), "Normative Ecological Rationality: Normative Rationality in the Fast-and-Frugal-Heuristics Research Program,' *Journal of Economic Methodology*, 21, 396-410.

Hansen, Pelle Guldorg (2016), "The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove? *European Journal of Risk Regulation*, 7, 155-174.

Harrison, Glenn W. and Ross, Don (2018), "Varieties of Paternalism and the Heterogeneity of Utility Structures," *Journal of Economic Methodology*, 25, 42-67.

Hausman, Daniel M. (1992), *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.

Hausman, Daniel M. (2012), *Preferences, Value, Choice, and Welfare*. New York: Cambridge University Press.

Hausman, Daniel M. (2016),"On the Econ Within," *Journal of Economic Methodology*, 23, 26-32.

Hausman, Daniel; McPherson, Michael; and Satz, Debra (2017), *Economic Analysis, Moral Philosophy, and Public Policy*. 3rd edition, Cambridge: Cambridge University Press.

Hausman, Daniel M. and McPherson, Michael (1996), *Economic Analysis and Moral Philosophy*, Cambridge: Cambridge University Press.

Hausman, Daniel M. and McPherson, Michael (2006), *Economic Analysis, Moral Philosophy, and Public Policy*, 2nd Edition, Cambridge: Cambridge University Press.

Hausman, Daniel M. and Welch, B. (2010),"Debate: To Nudge or Not to Nudge," *Journal of Political Philosophy*, 18, 123-136.

Hédoin, Cyril (2015), "From Utilitarianism to Paternalism: When Behavioral Economics Meets Moral Philosophy," *Revue De Philosophy économique*, 16, 73-106.

Heidl, Stefan (2016), *Philosophical Problems of Behavioral Economics*. London: Routledge.

Heilmann, Conrad (2014), "Success Conditions for Nudges: a Methodological Critique of Libertarian Paternalism," *European Journal for Philosophy of Science*, 4, 75-94.

Herfeld, Catherine (2018), "From Theories of Human Behavior to Rules of Rational Choice: Tracing a Normative Turn at the Cowles Commission, 1943-54," *History of Political Economy*, 50, 1-48.

Herrnstein, R. J.; Loewenstein, George F.; Prelec, Drazen; and Vaughn, Willian Jr. (1993), "Utility Maximization and Melioration: Internalities in Individual Choice," *Journal of Behavioral Decision Making*, 6, 149-185.

Heukelom, Floris (2014), *Behavioral Economics: A History*. Cambridge: Cambridge University Press.

Infante, Gerardo; Lecouteux, Guilhem; and Sugden, Robert (2016a), "Preference Purification and the inner Rational Agent: A Critique of The Conventional Wisdom of Behavioural Welfare Economics," *Journal of Economic Methodology*, 23, 1-25.

Infante, Gerardo; Lecouteux, Guilhem; and Sugden, Robert (2016b), "'On the Econ Within: A Reply to Daniel Hausman," *Journal of Economic Methodology*, 23, 33-37.

Kahneman, Daniel (2003), "Maps of Bounded Rationality: A Perspective on Intuitive Judgment," *American Economic Review*, 93, 1449--1475.

Kahneman, Daniel; Knetsch, Jack L. and Thaler, Richard (1991), "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, 5, 193-206.

Kahneman, Daniel and Thaler, Richard H. (2006), "Utility Maximization and Experienced Utility," *Journal of Economic Perspectives*, 20, 221-34.

Kahneman, Daniel and Tversky, Amos (1979), "Prospect Theory: An Analysis of Decisions Under Risk," Econometrica, 47, 263-91.

Kahneman, Daniel and Tversky, Amos (eds.) (2000), *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Kahneman, Daniel; Wakker, P. P.; and Sarin, R. (1997), "Back to Bentham? Explorations of Experienced Utility?," *Quarterly Journal of Economics*, 112, 375-406.

Knetsch, Jack L. (1989), "The Endowment Effect and Evidence of Nonreversible Indifference Curves," *American Economic Review*, 79, 1277-1284.

Knetsch, Jack L. (1992), "Preferences and the Nonreversibility of Indifference Curves," *Journal of Economic Behavior and Organization*, 17, 131-39.

Lee, Kyu Sang (2011), "Three Ways of Linking Laboratory Endeavors to the Realm of Policies," *European Journal of the History of Economic Thought*, 18, 755-76.

Lepenies, Robert and Malecka, Magdalena (2015), "The Institutional Consequence of Nudging – Nudges, Politics, and the Law," *Review of Philosophy and Psychology*, 6, 427-437.

Lichtenstein, Sarah and Slovic, Paul (1971), "Reversals of Preference Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, 89, 46-55.

Lichtenstein, Sarah and Slovic, Paul (2006a), "The Construction of Preference: An Overview," in *The Construction of Preference*, S. Lichtenstein and P. Slovic (eds.), Cambridge: Cambridge University Press, 1-40.

Lichtenstein, Sarah and Slovic, Paul (eds.) (2006b), *The Construction of Preference*. Cambridge: Cambridge University Press.

Lindman, Harold R. (1971), “Inconsistent Preferences Among Gambles,” *Journal of Experimental Psychology*, 89, 390-397.

Loewenstein, George and Chater, Nick (2017), “Putting Nudges in Perspective,” *Behavioural Public Policy*, 1, 26-53.

Loewenstein, George and Haisley, Emily (2008), “The Economist as Therapist: Methodological Ramifications of ‘Light’ paternalism,” in *The Foundations of Positive and Normative Economics: A Handbook*, A. Caplin and A. Schotter (eds.), Oxford: Oxford University Press, 210-245.

McQuillin, Ben and Sugden, Robert (2012), “Reconciling the Normative and Behavioural Economics: the Problems to be Solved,” *Social Choice and Welfare*, 38, 553-567.

Mills, Chris (2015), “The Heteronomy of Choice Architecture,” *Review of Philosophy and Psychology*, 6, 495-509.

Mitchell, Gregory (2005), "Libertarian Paternalism is an Oxymoron," *Northwestern University Law Review*, 99, 1245-1277.

Mongin, Philippe and Cozic, Mikael (2018), "Rethinking Nudge: Not One But Three Concepts," *Behavioural Public Policy*, 2, 107-124.

Nagatsu, Michiru (2015), “Social Nudges: Their Mechanisms and Justification,” *Review of Philosophy and Psychology*, 6, 481-494.

Nagel, Thomas (1974), “What Is It Like to Be a Bat?” *The Philosophical Review*, 83, 435-450.

Nussbaum, Martha (2001), *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press.

Pickering, Andrew (1995), *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Plott, Charles R. (1996), “Rational Individual Behaviour in Markets and Social Choice Processes: The Discovered Preference Hypothesis,” in K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (eds), *The Rational Foundations of Economic Behaviour*. London: Macmillan. 225-250.

Rawls, John (1971), *A Theory of Justice*. Cambridge, MA: Harvard University Press.

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

Rebonato, Riccardo (2014), "A Critical Assessment of Libertarian Paternalism," *Journal of Consumer Policy*, 37, 357-396.

Reiss, Julian (2013), *Philosophy of Economics: A Contemporary Introduction*. London: Routledge.

Rizzo, Mario and Whitman, Douglas Glenn (2009), "The Knowledge Problem in the New Paternalism," *Brigham Young University Law Review*, vol. 2009, issue 4, 905-968.

Robbins, Lionel (1935), *An Essay on the Nature & Significance of Economic Science*. 2nd edition, London: Macmillan.

Schubert, Christian (2017), "Exploring the (Behavioural) Political Economy of Nudging," *Journal of Institutional Economics*, 13, 499-522.

Seidl, Christian (2002), "Preference Reversal," *Journal of Economic Surveys*, 16, 621-655.

Sen, Amartya (1992), *Inequality Reexamined*. Cambridge, MA: Harvard University Press.

Sent, Esther-Mirjam (2004), "Behavioral Economics: How Psychology Made Its (Limited) Way Back Into Economics," *History of Political Economy*, 36, 735-60.

Sugden, Robert (2004), "The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences," *American Economic Review*, 94, 1014-1033.

Sugden, Robert (2008), "Why Incoherent Preferences do not Justify Paternalism," *Constitutional Political Economy*, 19, 226-248.

Sugden, Robert (2010), "Opportunity as Mutual Advantage," *Economics & Philosophy*, 26, 47-68

Sugden, Robert (2015), "Looking for a Psychology for the Inner Rational Agent," *Social Theory and Practice*, 41, 579-598.

Sugden, Robert (2017), "Do People Really Want to be Nudged Towards Healthy Lifestyles?" *International Review of Economics*, 64, 113-123.

Sugden, Robert (2018), "'Better Off, as Judged by Themselves': a Reply to Cass Sunstein," *International Review of Economics*, 65, 9-13.

Sunstein, Cass R. (2013), "The Storrs Lectures: Behavioral Economics and Paternalism." *Yale Law Journal,* 122, 1826–1898

Sunstein, Cass R. (2015), "Nudges, Agency, and Abstraction: A Reply to Critics," *Review of Philosophy and Psychology*, 6, 511-529.

Sunstein, Cass R. (2016), "People Prefer System 2 Nudges (Kind Of)," *Duke Law Journal*, 66, 121-168.

Sunstein, Cass R. (2018), "'Better Off, as Judged by Themselves': A Comment on Evaluating Nudges," *International Review of Economics*, 65, 1-8.

Sunstein, Cass R. and Thaler, Richard H. (2003), "Libertarian Paternalism Is Not an Oxymoron," *The University of Chicago Law Review*, 70, 1159-1202.

Thaler, Richard H. (1980), "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, 1, 39-60.

Thaler, Richard H. (2000), "From Homo Economicus to Homo Sapiens," *Journal of Economic Perspectives*, 14, 133-41.

Thaler, Richard H. (2017), "Behavioral Economics," *Journal of Political Economy*, 125, 1799-1805.

Thaler, Richard H. (2018), "From Cashews to Nudges: The Evolution of Behavioral Economics," *American Economic Review*, 108, 1265-1287.

Thaler, Richard H. and Sunstein, Cass R. (2003), "Behavioral economics, Public Policy, and Paternalism," *The American Economic Review*, 93, 175-179.

Thaler, Richard H. and Sunstein, Cass R. (2009), *Nudge: Improving Decisions About Health, Wealth and Happiness*. London: Penguin.

Tversky, Amos and Kahneman, Daniel (1983), "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review*, 90, 293-315.

Tversky, Amos and Kahneman, Daniel (1991), "Loss Aversion in Riskless Choice," *Quarterly Journal of Economics*, 106, 1039-1061 [reprinted as chapter 7 of Kahneman and Tversky 2000].

Whitman, Douglas Glenn and Rizzo, Mario J. (2015), "The Problematic Welfare Standards of Behavioral Paternalism," *Review of Philosophy and Psychology*, 6, 409-425.