Rev.Phil.Psych.
DOI 10.1007/s13164-015-0244-5

The Problematic Welfare Standards of Behavioral Paternalism

Douglas Glen Whitman · Mario J. Rizzo

Abstract Behavioral paternalism raises deep concerns that do not arise in traditional welfare economics. These concerns stem from behavioral paternalism's acceptance of the defining axioms of neoclassical rationality for normative purposes, despite having rejected them as positive descriptions of reality. We argue (1) that behavioral paternalists have indeed accepted neoclassical rationality axioms as a welfare standard; (2) that economists historically adopted these axioms not for their normative plausibility, but for their usefulness in formal and theoretical modeling; (3) that broadly rational individuals might fail to satisfy the axioms for various reasons, making them unpersuasive as normative criteria; and (4) that even if their violation did constitute irrationality, that would not justify paternalists' choosing among inconsistent preferences to define an individual's "true" preferences.

1 Introduction

> 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*. (Thaler and Sunstein 2008: 5; italics in original)

The case for behavioral paternalism (as we shall call it) is fundamentally *normative* in character; it prescribes or recommends policies to make people better off. As such, its justification cannot rest on positive (i.e., factual) research alone. In critically assessing behavioral paternalism, we must consider the philosophical and ethical assumptions on

D. G. Whitman (✉)
Department of Economics, David Nazarian College of Business and Economics, California State University, Northridge, 18111 Nordhoff St., Los Angeles, CA 91330-8374, USA
e-mail: glen.whitman@gmail.com

M. J. Rizzo
Department of Economics, New York University, 19 W. 4th St, New York, NY 10003, USA
e-mail: mario.rizzo@nyu.edu

Published online: 21 March 2015

which it is based. What welfare criteria justify the use of insights from behavioral economics to support paternalistic policies?

As presented by some of its leading proponents, notably Cass Sunstein and Richard Thaler, behavioral paternalism aims to make people better off *according to their own* (*true*) *preferences*; the quotation above is a typical statement of this goal. In appealing to individual preferences, behavioral paternalism appears to rely on the same normative standards as neoclassical welfare economics, which may be loosely characterized as a form of preference-based utilitarianism. Although utilitarianism has (of course) been subjected to many philosophical criticisms, it seems as though behavioral paternalism raises no *additional* concerns beyond those raised by standard welfare economics.

But in fact, behavioral paternalism raises deep concerns that do not arise in traditional welfare economics. In the traditional approach, agents are simply assumed – as a matter of fact – to meet the definition of rationality used in that approach. Thus, no questions arise regarding how to judge the welfare of agents who do not satisfy the defining axioms of rationality. Behavioral economics challenges the *positive* validity of those axioms in describing human behavior. Nevertheless, behavioral paternalism maintains those axioms as *normative* standards to which agents *ought* to conform. This matters because those axioms – typically referred to as completeness and transitivity – do not simply require that people follow their given preferences, whatever they might be. On the contrary, they impose a particular structure on those preferences.

In this article, we will argue that the rationality axioms adopted as normative criteria by behavioral paternalism are not justified. First, we support our claim the behavioral paternalists have adopted the neoclassical rationality axioms as a welfare standard. Second, we look to the history of these axioms, showing that economists adopted them not primarily for their normative plausibility, but for their usefulness in formal and theoretical modeling. Third, we consider the reasons why individuals who are rational in a broader sense might fail to satisfy those axioms, and thus why those axioms are not persuasive as normative criteria. And fourth, we draw attention to a glaring *non sequitur* at the heart of behavioral paternalism. The paternalists use evidence of internally inconsistent preferences as evidence of irrationality – and then, with little or no justification, they *select* among the inconsistent preferences to determine what someone's "true" preferences must be.

Before we proceed, two caveats are in order.

First, our argument applies to all forms of behavioral paternalism that rely on the notion of better satisfying individual preferences as a welfare standard. Sunstein and Thaler's "libertarian paternalism" or "nudge" is the best-known example, but there are others. Specifically, we have in mind Camerer et al.'s (2003) notion of asymmetric paternalism (which purports to "help people to make better decisions and come closer to behaving in their own best interest" [1218]); Bernheim and Rangel's (2007) discussion of behavioral public economics; Jolls and Sunstein's (2006) concept of debiasing through law; and the sin tax models of Gruber and Koszegi (2001) and O'Donoghue and Rabin (2003, 2006). Conly's (2012) case for coercive paternalism also appears to qualify.

Second, we want to emphasize that behavioral paternalism goes further in terms of policy recommendations than is commonly advertised. Libertarian paternalism allegedly aims to influence choices for the better *without restricting freedom of choice*. As Sunstein and Thaler put it, "Libertarian paternalists want to make it easy for people to

go their own way; they do not want to burden those who want to exercise their freedom" (2008: 5); "To count as a mere nudge, the intervention must be easy and cheap to avoid" (2008: 6). Yet in earlier papers on the topic (Sunstein and Thaler 2003; Thaler and Sunstein 2003), they include under the libertarian paternalist umbrella many policies that do, in fact, burden freedom of choice. Such policies include cooling-off periods, which foreclose the option of concluding a final sale immediately (Sunstein and Thaler 2003: 1188); contractual defaults that cannot be waived without significant effort (1186–7); and labor rules that cannot be waived at all, such as time-and-a-half pay for overtime (1187). As it turns out, the libertarian paternalist *framework* can countenance fairly substantial restrictions: "[A] libertarian paternalist who is especially confident of his welfare judgments would be willing to impose real costs on workers or consumers who seek to do what, in the paternalist's view, would not be in their best interests" (1185–6). And even if Sunstein and Thaler do place substantial weight on freedom of choice, other behavioral paternalists do not. Gruber and Koszegi (2001) and O'Donoghue and Rabin (2003, 2006) advocate sin taxes which, of course, cannot be waived by consumers of the taxed products. Conly (2012) forthrightly advocates bans and mandates.

2 The Normative Standards of Behavioral Paternalism

Neoclassical economics has largely been built on models of rational choice. Behavioral economics is often presented as a refutation of rational choice. This is particularly true in the behavioral paternalist literature. Camerer et al. (2003), for instance, offer this summary:

> The standard approach in economics assumes "full rationality." While disagreement exists as to what exactly full rationality encompasses, most economists would agree on the following components: First, people have well-defined preferences (or goals) and make decisions to maximize those preferences. Second, those preferences accurately reflect (to the best of the person's knowledge) the true costs and benefits of the available options. Third, in situations that involve uncertainty, people have well-formed beliefs about how uncertainty will resolve itself, and when new information becomes available, they update their beliefs using Bayes's law – the presumed ability to update probabilistic assessments in light of new information.
>
> *Behavioral economics challenges all of these assumptions* and attempts to replace them with more realistic approaches based on scientific findings from other social sciences. (1214–15, italics added, footnotes omitted)

Similarly, Sunstein and Thaler (2003) say:

> The false assumption is that almost all people, almost all of the time, make choices that are in their best interest or at the very least are better, by their own lights, than the choices that would be made by third parties. This claim is either tautological, and therefore uninteresting, or testable. *We claim that it is testable and false*, *indeed obviously false*. (1163).

But although behavioral paternalists have rejected the neoclassical notion of rationality as a *positive* description of behavior, they have – perhaps surprisingly – retained it as a normative standard. When they advocate policies designed to improve or correct behavior, what they mean is encouraging behavior that conforms more closely to the neoclassical ideal that they believe is factually false. We will justify this claim later in this section, but first we should be more specific about the requirements of neoclassical rationality.

In this article, we focus our attention on the first of the neoclassical rationality criteria listed by Camerer, et al.: having well-defined preferences. "Well-defined" is economics jargon for preferences that satisfy two axioms: *completeness* and *transitivity.*[1]

Completeness, as usually defined, means that between any two alternatives A and B, the agent must either strictly prefer A to B, strictly prefer B to A, or be indifferent between A and B. Completeness rules out internal contradictions, such as strictly preferring both A to B *and* B to A. In addition, completeness requires a certain *universality*: the agent is assumed to order *any two items in the universe of possibilities* in this way.

Transitivity, as usually defined, means that for any three alternatives A, B, and C, if A is preferred to B and B is preferred to C, then A must be preferred to C. (The preferences involved may be either weak or strict.) Like completeness, transitivity serves to rule out internal contradictions and to impose a form of consistency over the entire set of preferences.

A third axiom, *independence of irrelevant alternatives* (IIA), is sometimes appended to these two. This axiom says that an agent's preference between two options A and B should not be affected by the presence or absence of a third option C. That is, if the agent prefers A to B when C is available, the agent should also prefer A to B when C is not available. As it turns out, IIA is implied by completeness, provided that completeness is defined independently of the choice set (as it is above).[2] We will therefore ignore IIA in this article.

Both completeness and transitivity impose a kind of consistency on preferences. Furthermore, a violation of completeness that takes the form of contradictory preferences – i.e., A is strictly preferred to B *and* vice versa – technically violates transitivity as well: if A is strictly preferred to B, and B is strictly preferred to A, then transitivity would imply A is strictly preferred to itself, which is ruled out by the definition of strict preference.[3] As a result, the axioms are not always clearly distinguished in the literature, nor are they always mentioned by name. Instead, preferences that satisfy all three axioms are typically said to be "well-defined" (as in the Camerer, et al., quote above), while preferences that reveal any form of inconsistency are said to be "not well-defined."

[1] "The hypothesis of rationality is embodied in two basic assumptions about the preference relation…: *completeness* and *transitivity.*" Mas-Colell et al. (1995: 6).

[2] See, for instance, Gintis (2009). Gintis's definition of completeness has the preference relation defined with respect to specific choice sets, so that A might be preferred to B for one choice set and B to A for another. Then he introduces IIA as a third axiom, and notes: "Because of the third property [IIA], we need not specify the choice set and can simply write [that A is preferred to B]" (5, bracketed changes made for consistency of notation).

[3] Or by reflexivity, which is sometimes regarded as a separate axiom. But usually, reflexivity is taken to be implied by the definitions of indifference and strict preference in terms of weak preference: If A is weakly preferred to B and vice versa, then A ~ B; if A is weakly preferred to B but not vice versa, then A is strictly preferred to B. By these definitions, it's simply impossible to have A strictly preferred to A.

Many of the alleged anomalies of choice used by behavioral paternalists to justify policy either violate or appear to violate one or both of these axioms. A short list will suffice:

- *Hyperbolic and quasi-hyperbolic discounting*. When preferences are hyperbolic or quasi-hyperbolic, an agent will prefer (A) a larger reward at time t+1 to (B) a smaller reward at time t, and yet prefer B to A when nothing has changed except the passage of time – that is, t and $t+1$ have both come closer to the present (Frederick et al. 2002). This can be characterized as a violation of either completeness, transitivity, or both, and it has been used to justify imposing sin taxes (among other policies); see Gruber and Koszegi 2001; O'Donoghue and Rabin 2003, 2006. (Alternatively, hyperbolic discounting may be characterized not as a violation of transitivity or completeness, but as a violation of *stationarity*, a lesser-known axiom that requires preferences not to change over time [Manzini and Mariotti 2009: 243–246]. Like the better-known axioms, stationarity imposes a form of consistency on preferences – one that is even harder to justify on normative grounds. In this article, we will stick with the interpretation of hyperbolic preferences as a transitivity or completeness violation.)
- *Hot-cold empathy gaps*. Agents are more inclined to make certain choices such as impulse purchases when in a "hot" state (such as fear, excitement, or sexual arousal) than when in a "cold" state (calm and sober). Thus, the agent prefers A to B when choosing in a hot state, and B to A when choosing in a cold state (Loewenstein 2000). This appears to be a violation of completeness and/or transitivity, and it has been used to justify imposing cooling-off periods on various transactions (Sunstein and Thaler 2003: 1188, Camerer et al. 2003: 1238–40).
- *Framing effects*. Given a choice situation described in two different but logically equivalent ways, the agent's choice differs depending on the description. For instance, a patient is more likely to opt for surgery described as having a 90 % survival rate than surgery described as having a 10 % death rate (Redelmeier et al. 1993). Since framing results in having both A preferred to B and B preferred to A for what the analyst regards as the same choice situation, framing effects are regarded as violations of completeness or transitivity. Framing effects can be difficult to separate from the next two choice anomalies – endowment effects and status-quo bias – and have been used to justify some of the same policies (Sunstein and Thaler 2003: 1179–82; see also citations below).
- *Endowment effects*. Willingness to pay (WTP) is what an agent will pay for an item when he doesn't already own it. Willingness to accept (WTA) is what an agent will accept to sell an item when he does already own it. When WTA exceeds WTP, as has been shown to occur in experimental results (Kahneman et al. 1991), then we can choose a price P in between WTP and WTA and use it to generate a violation of completeness. E.g., if WTP=$4 and WTA=$6, let P=$5. Then the individual will prefer the item over $5 when he already owns the item, but he will prefer $5 to the item when he does not already own the item. Technically, these two situations aren't precisely the same, because ownership of the item increases the agent's wealth. But because wealth effects are negligible – that is, the value of the item is very small relative to the agent's base wealth – the observed behavior is tantamount to a violation of completeness or transitivity (or both). The endowment effect has been used to justify changing default rules of labor contracts by (for instance) assuming

2 weeks paid vacation, for-cause rather than at-will termination, and other allegedly labor-friendly terms (Sunstein and Thaler 2003: 1175–77, 1181).

- *Status-quo bias*. People more often choose an option over an alternative when it is perceived as the default or the status quo – i.e., when they would have to opt out of it rather than opting in (Kahneman et al. 1991). Because what is perceived as the *status quo* can be affected by how the situation is described, *status quo* bias sometimes overlaps with framing effects. Although arguably the choice situations for opting in and opting out aren't identical – because any act of opting out involves some effort – behavioral paternalists have nevertheless treated *status quo* bias as evidence of an inconsistency because "the cost of turning in a form is trivial" (Sunstein and Thaler 2003: 1171). If so, then *status-quo* bias constitutes a violation of completeness, transitivity, or both. This is used to justify some of the same policies as the endowment effect, as well as automatic enrollment in savings plans (1172).

Other explanations of decision-making anomalies, such as switching costs and inference of information from the framing of the problem, have been offered. We take no position (here) on whether those alternative explanations are correct. Instead, we offer an imminent critique. *Assuming that the behavioral paternalists are right* in characterizing these behaviors as demonstrating inconsistencies of choice, what does that imply about their welfare criteria?

Let us clarify the analytical move the behavioral paternalists are making. In keeping with the notion of subjectivism, they have not tried to characterize any preference *in isolation* as necessarily irrational. There is nothing per se irrational about eating unhealthy foods, splurging on the present rather than saving more for the future, buying an expensive car in a moment of excitement, or demanding a high price to part with a recently acquired treasure. The leading behavioral paternalists admit this.[4] How, then, can behavioral paternalists be confident that people are making inferior choices? Their argument relies crucially on demonstrating *inconsistencies* of choice and preference, including those described above. Such inconsistencies supposedly raise a red flag that people are not fully rational – and therefore their behavior could potentially be improved by outside intervention.

The leading behavioral paternalists, particularly Sunstein and Thaler, have not *explicitly* adopted transitivity and completeness as welfare norms. Thaler has said, "A demonstration that human choices often violate the axioms of rationality does not necessarily imply any criticism of the axioms of rational choice as a normative idea" (Thaler 1991: 138). But that was in 1991, many years before "libertarian paternalism," so it is possible that his view has evolved.[5] However, approval of neoclassical

[4] See, for instance, Sunstein and Thaler (2003: 1168): "Of course, rational people care about the taste of food, not simply about health, and we do not claim that everyone who is overweight is necessarily failing to act rationally." Sunstein (2014: 75): "… I am interested in defending paternalists who respect choosers' own views about their ends, and who seek to increase the likelihood that their decisions will promote those ends." Sunstein (2014: 96): "With respect to diet, savings, exercise, romance, credit cards, mortgages, cell phones, health care, computers, and much more, different people have divergent tastes and situations, and they balance the relevant values in different ways."

[5] The view that behavioral models are positive while neoclassical models are normative was common in the early days of behavioral economics. Floris Heukelom observes, "Contrary to [Herbert] Simon, [Daniel] Kahneman and [Amos] Tversky argued that there was nothing wrong with economists' theory of expected utility maximization. It was only that this was the normative theory, and not an accurate description of actually observed human behavior" (2014: 127).

rationality norms is implicit in the form of Sunstein and Thaler's argument, and is also evident in the work of other behavioral paternalists.

First, Sunstein and Thaler repeatedly say that welfare must be judged according to people's own preferences – "as judged by themselves" or "by their own lights" (e.g., Sunstein and Thaler 2003: 1163, 1170). Thus, they eschew an objective notion of welfare that is independent of what people really want; as in the neoclassical approach, their notion of welfare is subjectivist. Second, behavioral economists question "the rationality of many judgments and decisions that individuals make" (Sunstein and Thaler 2003: 1168) in order to argue that people are making "inferior decisions" (1162). As evidence for these claims, they point to behaviors that deviate from the neoclassical model, including inconsistencies. If it were not for these deviations, their argument for paternalism could not even get off their ground, as there would be nothing to fix. Decision-making failure is demonstrated by departures from the neoclassical model.

Third, they advocate policies designed to help people come closer to full rationality. This is stated most clearly by Sunstein (2014): "Some forms of paternalism move people in the directions that they would go if they were fully rational. Paternalism, whether hard or soft, creates 'as if' rationality. Indeed, that is a central point of good choice architecture" (154). If inconsistencies are evidence of the problem, then improvements presumably must involve choices that display fewer such inconsistencies. Ideal choice behavior would conform to some set of subjective preferences that satisfy the neoclassical axioms, including completeness and transitivity.

It is conceivable that Sunstein and Thaler do not fully embrace the neoclassical definition of rationality, but instead have some different – perhaps looser – notion of rationality in mind. But if so, they have not stated it clearly. Furthermore, whatever notion of rationality they have in mind, it must involve some kind of consistency requirement; otherwise, the inconsistencies they point to as evidence of irrationality and inferior decision-making would be irrelevant.

Other behavioral paternalists have also relied on the neoclassical model of rationality as a normative benchmark. Camerer et al. (2003) replicate Sunstein and Thaler's approach almost exactly. After citing inconsistencies of choice (among other things) as violations of rationality, they conclude: "It is such errors – apparent violations of rationality – that can justify the need for paternalistic policies to help people make better decisions and come closer to behaving in their own interest" (1218). Bernheim and Rangel's (2007) use of neoclassical rationality as a welfare norm is even more explicit:

> A natural analytic strategy involves endowing the individual with *well-behaved lifetime preferences*, while simultaneously specifying a decision process (or decision criterion) that does not necessarily involve selecting the maximal element in the preference ordering. To conduct positive analysis, one employs a model of the decision process (or criterion). *To conduct normative analysis*, *one uses a model of lifetime preferences*." (Bernheim and Rangel 2007: 16, italics added)

The key phrase here is "well-defined lifetime preferences." Again, "well-defined" is economics jargon for satisfying completeness and transitivity.

Both pairs of authors who have advocated sin taxes on behavioral grounds, Gruber and Koszegi (2001) and O'Donoghue and Rabin (2003, 2006), adopt as their normative standard an intertemporal utility function with exponential rather than hyperbolic or

quasi-hyperbolic time-discounting. Note that exponential utility functions satisfy the neoclassical axioms (including stationarity) and thus produce no inconsistencies of choice, whereas hyperbolic or quasi-hyperbolic ones do. To impose exponential time-discounting is to impose internal consistency of intertemporal preferences.

Finally, we should note that we are not the first to characterize behavioral paternalists as having adopted neoclassical rationality as a welfare norm. After documenting the increasing willingness of economists to make policy prescriptions based on behavioral research, Berg and Gigerenzer (2010) observe: "This evolution in boldness about looking for prescriptive implications of behavioral economics does not, unfortunately, imply increased boldness about modifying the neoclassical axiomatic formulations of rationality as the unquestioned gold standard for how humans ought to behave" (147).

The behavioral paternalist case rests, then, on the *normative* strength of the neoclassical rationality axioms that are violated by decision-making anomalies – or, perhaps, on some looser criteria that nevertheless impose a similar requirement of consistency.[6]

3 The Origin of Rationality Axioms in Economic Theory

We now turn to the history of the neoclassical rationality axioms. This history is relevant because it helps us to understand why they were accepted in the past, and hence why they are still accepted – for the most part uncritically – by economists today. One might assume, given the decades-long pedigree of these axioms in economic theory, that they had already been fully justified. But it turns out that these axioms lacked a prescriptive justification from their inception. They originally entered economic analysis primarily as positive statements, devoid of prescriptive content. Specifically, they were introduced to provide a logical foundation for microeconomic theory – particularly the existence of utility functions and demand curves.

To the best of our knowledge, the axioms first appeared in economics in the work of Kenneth Arrow on social welfare orderings and his famed Impossibility Theorem (Arrow 2012). That work had an undeniable normative component, as the Impossibility Theorem shows the impossibility of simultaneously satisfying several normative goals for deriving a social welfare ordering from individual preference orderings. But the "rationality" requirements of completeness (which Arrow called "connectedness") and transitivity did not appear among the normative goals (which Arrow called "conditions"). Rather, they appeared among the book's initial *assumptions*, without which the analysis could not proceed: "Throughout this analysis it will be assumed that individuals are rational, by which is meant that the ordering relations R_i satisfy Axioms I [completeness] and II [transitivity]" (19). Completeness and transitivity assured that Arrow could speak unambiguously of an individual's "ordering" of alternatives for the

[6] There is an unfortunate ambiguity in the word "normative" when referring to rules. There are constitutive rules and regulative or prescriptive rules (Searle 1969). The rules of chess may be likened to constitutive rules. They define what it means to be playing chess. Violate them sufficiently and you are not playing chess. The rules of safe driving are prescriptive rules. Violate them and you are still driving but not safely. The axioms of rationality are *constitutive rules*. If you obey them you are a "rational" agent in the sense of standard economic theory. Whether you *should* be such an agent in particular circumstances is a separate matter that needs to be argued for explicitly.

rest of the book. To our knowledge, Arrow's was the first use of these two axioms as the definition of rationality.

The usefulness of Arrow's approach elsewhere in microeconomic theory soon became clear. Debreu (1954) proved the existence of a continuous utility function on the assumptions of completeness, transitivity, and the additional axiom of continuity. Arrow and Debreu (1954) put that proof to immediate use when they assumed existence of continuous utility functions for all consumers (169) *en route* to proving the existence of equilibrium in a competitive economy. Neither of these works used the word "rational" to describe the axioms. Later, Uzawa (1956) and Arrow (1959) merged the axiomatic approach with the revealed preference approach, and in these papers the word "rational" was reintroduced – as a description of various conditions that a choice function (that is, a function that maps sets of alternatives into choices) might satisfy. None of these papers offers any defense of completeness or transitivity as a normative standard.

Revealed preference, in its original form, was born from the popularity of behaviorism in early 20th century social science. Despite the name, behaviorism has no important connection to modern behavioral economics. Behaviorism tried to eschew all reference to the internal mental states of agents, instead limiting itself to descriptions of behavior (in contrast to modern behavioral economics, which attempts to understand behavior as resulting from mental processes). As revealed preference theory was originally conceived, "preferences" were simply descriptions of actual choices, without reference to mental states. This serves to underline that, in the context where rationality acquired its modern axiomatic definition, prescriptive concerns were strictly background; the focus was on objective description.

In short, the neoclassical rationality axioms were originally adopted by the economics profession for their descriptive usefulness, not their normative value. Consequently, if we wish to find their normative justification, we will need to find it elsewhere.

4 The Questionable Normative Status of Rationality Axioms

Do I contradict myself?

Very well then I contradict myself,

(I am large, I contain multitudes.)

– Walt Whitman, *Leaves of Grass* (1855)

Why should completeness and intransitivity be treated as desirable? And relatedly, why should an agent who lacks them be regarded as "irrational"?

Modern texts still offer completeness and transitivity as axioms that guarantee a "total preorder" of alternatives and that, along with continuity, guarantee the existence of a continuous utility function. But the definitions are often accompanied by qualifications; Mas-Colell et al. (1995) is illustrative:

> The strength of the completeness assumption should not be underestimated. Introspection quickly reveals how hard it is to evaluate alternatives that are far from the realm of common experience. It takes work and serious reflection to find out one's own preferences. The completeness axiom says that this task has taken place: our decision makers make only meditated choices. (6)

> Like completeness, the transitivity assumption can be hard to satisfy when evaluating alternatives far from common experience. As compared to the completeness property, however, it is also more fundamental in the sense that substantial portions of economic theory would not survive if economic agents could not be assumed to have transitive preferences. (7)

Note that the argument in the last sentence is strictly a "necessity" defense: transitivity *needs to be* true in order for (descriptive) economic theory to work. It is not a claim that transitivity makes sense as a prescription.

In essence, the axioms require an individual to have *pre-rationalized* his attitudes about the universe of all possibilities before making any actual choices in the world. This is closely akin to the "equilibrium always" assumption that characterizes so many neoclassical models. All relevant forces must have fully worked themselves out so that no further adjustment is necessary. The observed situation thus constitutes a point of rest and balance – like a pendulum that has finally settled into an unmoving vertical position.

In practice, equilibrium-always may be true enough to be useful for some descriptive purposes. But normatively, there is no reason to insist that, in order to be considered rational, every agent should have *already* arrived at fully consistent preferences – no more than an entrepreneur should have *already* created and implemented a full business plan, purchased all inputs, and commenced production. These are processes that play out in real time. A broader notion of rationality, grounded in reasonability and responsiveness to costs and benefits, would not impose such an arbitrary requirement. There are at least three distinct reasons why the preferences of an agent who is rational in this sense might not be fully rationalized in advance:

Preference Discovery A person may not yet know their own preferences. The process of making choices is thus one of gradually learning one's preferences. Implicit here is the notion that a person has a true, underlying set of preferences to be uncovered. The person may make choices that deviate from those as-yet-unknown preferences, including choices that contradict each other. But such mistakes are not necessarily irrational in the broader sense. Just as an entrepreneur may experiment with different business concepts to discover as-yet-unrealized profit opportunities (see Hayek 2002), an individual may experiment with different choices and combinations of choices to discover which best satisfy his underlying preferences.

Preference Formation The notion of preference discovery requires preferences that preexist the act of choice. There is another possibility: that preferences are formed during the process of choice. As James Buchanan puts it, "… I am here advancing the more radical notion that *not even* individuals have well-defined and well-articulated objectives that exist independently of choices themselves" (1979: 111). Choice-making

is a creative process in which the individual changes along with the constraints he faces and the choices he makes. Buchanan again: "Individuals do not act so as to maximize utilities described in *independently existing functions*. They confront genuine choices, and the sequence of decisions taken may be conceptualized, *ex post* (after the choices), in terms of 'as if' functions that are maximized. But these 'as if' functions are, themselves, generated in the choosing process, not separately from such process" (1982).

The leading behavioral paternalists have lent support to this position. Sunstein and Thaler (2003: 1161) state, "[I]n many domains, people lack clear, stable, or well-ordered preferences. What they choose is strongly influence by details of context… These contextual influences render the very meaning of the term 'preferences' unclear." Indeed, the authors say that people may have "ill-formed" or "ill-defined" preferences, or lack "well-defined" or "well-formed" preferences, at least fourteen times throughout the article (1159, 1161, 1164, 1165, 1174, 1177, 1178, 1179, 1181, 1182, 1201).

How does embracing an ongoing process of preference formation affect the behavioral paternalists' normative project? To be blunt, it robs them of the Archimedean point that they would use to judge outcomes. If preferences do not exist independently of the act of choice, then there is no preference set against which to judge the individual's choices as deficient.

Economizing on Cognitive and Non-cognitive Effort Whether preferences are formed in the process of choice or merely waiting to be discovered, another argument explains why we should not expect an equilibrium to emerge in which preferences are fully consistent. As suggested by Mas-Colell, et al. (above), the process of considering (possibly hypothetical) pairs of options, settling on preferences over them, and making sure that all such preferences are mutually consistent (through all possible chains of binary comparisons) is costly. Even if carried out strictly in the mind, the process involves time and cognitive effort. If carried out in real life and real time, as in Buchanan's perspective, it involves non-cognitive resources as well. The expected marginal benefit of discovering and/or forming these preferences presumably declines as the compared options get further from one's likely future experience. Therefore, a rational person (in the broader sense) who compares costs and benefits will not, and *should* not, have complete and transitive preferences.

Before moving on, we should respond to a common argument for the irrationality of intransitivity: the money pump. Agents with intransitive preferences can allegedly be led to pay small sums to move from A to B and then from B to C and finally from C to A, thus returning them to their original position minus some money. If this process is repeated indefinitely, the agent can be drained of money completely, which surely seems irrational. However, this is not an argument for why it would be irrational to have intransitive preferences in one's head – only for why acting upon them might sometimes be problematic in practice. But we are not aware of any evidence in the real world of individuals being pumped of their money in this fashion. The possibility of a money pump depends on transaction costs (including time costs) being low enough to make all of the exchanges worth doing repeatedly. It also depends on both a high degree of awareness and cleverness on the part of the pumper, and a complete lack of global awareness on the part of the pumpee. If we make symmetrical assumptions

about the awareness of pumper and pumpee, the pumping cannot continue indefinitely, as the targeted agent will at some point simply refuse to continue making exchanges.[7]

The simple fact that we are able to reasonably criticize the normativity of the customary axioms for rational preferences is itself quite significant. It shows that the rationality norms behavioral economists have adopted are only a particular manifestation of a more general and profound notion of rationality. Observed preferences may not conform to a "rational" structure because of reasonable – or cost-justified – indecisiveness in the trade-off of values, decision rules applying multiple criteria, the existence of vague or unstable preferences, or variations in the economist's description of possible alternatives (under some descriptions the same behavior may be transitive or intransitive).[8] None of these factors are rooted in preferences or decision rules that fail a reasonableness test.[9]

5 The *Non Sequitur* at the Heart of Behavioral Paternalism

As shown in Part 2, to demonstrate irrationality, behavioral economists often point to inconsistencies of choice that imply the existence of inconsistent preferences. The behavioral paternalists then propose policies designed to correct these inconsistencies – that is, to induce behavior that conforms to a consistent preference ordering.

We argued in Parts 3 and 4 that inconsistent preferences do not necessarily mean the individual is irrational in any normatively significant sense. Here, we assume for argument's sake that such inconsistencies do, in fact, constitute irrationality. Even so, that does not provide any grounds for a third party, such as a behavioral economist or policymaker, to resolve the irrationality by choosing which among the inconsistent preferences should be followed consistently. If an agent shows evidence of having both Preference Set X and Preference Set Y, there is no analytical basis for designating X or Y as the "true" underlying preference set of the agent. Maybe it's both; maybe it's neither. To choose one over the other is simply a *non sequitur*. If Sunstein and Thaler are correct in their oft-repeated claim that agents may simply lack well-defined preferences, then the analyst lacks "true" preferences by which to judge choices. There's no there, so to speak.

A closely related argument has been made by Rizzo and Whitman (2009): that inconsistencies of choice create an insurmountable "knowledge problem" for the paternalist planner, because the planner has no means of determining which of the conflicting preferences reflect the agent's *true* preferences.[10] Here, we focus on a deeper philosophical challenge: that the agent may not even *have* true preferences (i.e., *underlying* preferences, not enacted in choice, that satisfy the neoclassical axioms). Even an omniscient planner cannot have knowledge of what does not exist. Where Rizzo and Whitman's (2009) challenge is epistemic, our challenge here is metaphysical.

To be more specific, consider some of the specific inconsistencies of choice:

[7] Sugden (2004: 1028–9) offers a related but distinct response to the money-pump argument: that competition among potential money-pumpers keeps profits to a minimum, thereby making money pumps ineffective at extracting value from consumers in equilibrium – even if such consumers have incoherent preferences.
[8] For further discussion see Anand (1993).
[9] For a discussion of the general concept of rationality and the reasonableness test, see Rescher (1988).
[10] See Schnellenbach (2012) for more on the knowledge problem as applied to time-discounting.

Hyberbolic or Quasi-hyperbolic Discounting For an agent who seemingly has two rates of time preference, there are at least two possible ways to induce an individual with this kind of discounting to behave consistently: (1) make her follow her *more* patient rate of time preference consistently, or (2) make her follow her *less* patient rate of time preference consistently.[11] Or we could (3) make her consistently follow some intermediate rate. Behavioral paternalists have, with little justification, chosen (1).

Hot-Cold Empathy Gaps For agents who choose differently depending on their mental state, there are at least two ways to "correct" the chooser's inconsistency: (1) make him always behave as he would when in a hot state, or (2) make him always behave as he would in a cool state. Behavioral paternalists have, with little justification, chosen (1).[12]

Framing Effects Take the case of paid vacation time, which we assume the agent tends to prefer when it's presented as the default, but not to prefer when it's not the default. (If it seems obvious that paid vacation is better, keep in mind that wages and other forms of compensation can and probably do adjust to compensate. So the question here is whether the agent is willing to *buy* or *sell* the vacation time.) Again, there are at least two possible fixes: (1) consistently favor the paid vacation or (2) consistently favor the lack of paid vacation.[13] Behavioral paternalists have, with little justification, chosen (1).

In the first two cases, the selected preference seems to be the socially preferred one; after all, "everyone knows" that saving more, eating healthy foods, and resisting impulse purchases is responsible behavior. In the third case, the selected preference seems to derive from a progressive ideology that says employers ought to treat their employees better. But none of these preferences is derived from the individual himself – except in a trivial sense that could equally well justify the opposite policies. By picking and choosing among preference sets, the behavioral paternalists effectively abandon their stated welfare standard of the individual's own preferences – which, to reiterate, may not even exist – in favor of an *external* set of preferences. Even our phrasing of the examples above may be somewhat misleading, as it might suggest the possibility that the paternalists have chosen preference set (1) when they should have chosen preference set (2) or (3). The key issue is that it requires an unjustified leap of logic to choose *any* of these answers, or indeed to assume that there *is* an answer.

Some defenses have been offered for favoring some preferences over others in cases of conflict, but we find these defenses weak:

Verbal Statements and Survey Responses When asked, people may say that they would rather behave differently or have different preferences. For instance, smokers may say they would rather not smoke, and overweight people may say they would like to eat less. It is indeed *possible* that these statements reveal "true" preferences. However, the incentives for speech differ from the incentives for other kinds of action.[14] Behavioral research has cast doubt on many economic principles previously taken as given, but the principle that *talk is cheap* remains intact. Speakers who say one thing while doing

[11] Rizzo and Whitman (2009: 925): "We could just as easily designate the more near-sighted preferences as the correct ones, and then aim to make far-term choices better correspond to them."
[12] See Rizzo and Whitman (2009: 929–931).
[13] See Rizzo and Whitman (2009: 929).
[14] See Rizzo and Whitman (2009. 904–905).

another may simply be expressing what they regard as socially approved attitudes. Or their statements may simply reflect "experienced opportunity cost," i.e., the dissatisfaction that always results from options the agent has forgone.

Regret A person may feel, and express, feelings of regret about the choices they have made: "I wish I had not done that." Although regrets are real, they do not necessarily reflect all costs and benefits associated with an action. Especially for intertemporal choices, such as getting inebriated last night and having a hangover today, the regret is typically experienced while the cost is being experienced in the present and the benefit is already in the past. That does not mean the costs outweighed the benefits *at the moment of choice* – only that the *remaining* costs outweigh the *remaining* benefits. Furthermore, it's worth noting that regret can also be felt about the kinds of choices that behavioral paternalists favor. When approaching death, people often express regret at not having lived a more spontaneous and present-oriented life (Ware 2012). If regret may be experienced regardless of the action taken, then it offers little guidance to the paternalist about which preferences are "true."[15] As with verbal statements, regret can simply reflect the experience of opportunity cost.

Self-Constraint and Commitment Devices People will often use various devices and strategies to try to keep their vices under control: planning automatic deductions for savings, avoiding locations where they will be tempted to smoke or drink, etc. These activities do provide further evidence of conflicting preferences. They do not, however, show which preferences are superior. Commitment devices reveal one set of preferences at work – but other choices show other preferences at work. Furthermore, the outside observer has no means of knowing whether the *right amount* of self-constraint has been performed. The level of self-constraint the person has already chosen might represent a delicate balance between their conflicting preferences. Or there may not be any correct balance to be found, inasmuch as the individual's true preferences could be a chimera. In any case, the analyst is not in a position to know (see Rizzo and Whitman 2009: 919–920).

Planned Versus Unplanned Choices Behavioral paternalists often favor the preferences of a "planning self" over the spontaneous or "acting" self. The idea is that the planning self is more likely to take all costs and benefits into account and render a considered decision. But the planning self does not necessarily represent a disinterested party; rather, the planning self may represent only the longer-term and more self-denying parts of one's personality (Cowen 1991). This becomes most apparent in the case of extreme behaviors like anorexia, where the planning self dominates an acting self that might wish to indulge more often.

Breakdown in Decision-Making Processes According to this argument, all inconsistencies of choice result, not from preferences being genuinely inconsistent or ill-formed, but from errors of judgment or decision-making. From this perspective, "true preferences" must have a standard well-defined structure. Anything that interferes with

[0] See Rizzo and Whitman (2009. 904–905).
[15] See Rizzo and Whitman (2009, 930).

the implementation of these true well-defined preferences constitutes a distortion or malfunction in the relevant psychological decision-making processes.[16] In essence, the advocates of this approach re-impose the neoclassical axioms that guarantee consistency of preferences, and to do so, they place all "blame" for choice inconsistencies on process errors. The implicit metaphysical assumption, nowhere explicitly defended, is that each person has a neoclassical agent deep within himself which is struggling to surface. Decision-making processes are thus deemed to be malfunctioning insofar as they fail to produce choices consistent with the standard preference structure. In other words, malfunction is not independently defined; it is whatever does not make standard choice theory descriptively accurate. This approach thus *assumes away* the possibility of individuals who simply lack "true" preferences satisfying the neoclassical axioms of transitivity and completeness. In other words, it is just another way of committing the same *non sequitur*: identifying an inconsistency, and then resolving it by designating one set of preferences as "true."

6 Conclusion

We wish to emphasize that the present criticism is just one of many hurdles that behavioral paternalism must clear to be justified. First, it must establish a set of normative standards for rational or welfare-enhancing individual behavior. Second, it must provide evidence that real-world behavior significantly and systematically departs from these standards. Third, it must show that policy makers have the requisite knowledge to craft policies that will successfully move individuals toward better, i.e., more welfare-enhancing, behavior according to these standards. And finally, it must show that successful policies can in fact be implemented without unacceptably high costs in welfare, freedom, or other important values. Previous work by ourselves and others has argued that behavioral paternalism cannot satisfy the second, third, and fourth criteria. In this article, we have focused attention solely on the first.

If our argument here is correct, then behavioral paternalism in its current form lacks a normative standard on the basis of which to make judgments about better and worse outcomes. Appealing to the individual's welfare "as judged by their own preferences" provides no guidance if the preferences either don't exist or exist but conflict with each other.

Given this problem, behavioral paternalists may be tempted to abandon subjective preferences and appeal instead to some objective notion of the good. In doing so, they would lose much of what allegedly distinguishes the new paternalism from the old paternalism, which simply imposes values on people without regard to their own good as they see it. Furthermore, objective notions of the good have their own philosophical problems that are perhaps even more daunting than those of ethical theories grounded in the subjective values of individuals.

[16] This is explicitly stated by Bernheim and Rangel (2007: 16): "We assume that people attempt to optimize given their true preferences, but randomly encounter conditions that trigger systematic mistakes…"

References

Anand, P. 1993. The philosophy of intransitive preference. *The Economic Journal* 103: 337–346.

Arrow, K.J. 1959. Rational choice functions and orderings. *Economica, New Series* 26(102): 121–127.

Arrow, K. J. 2012 [1951]. *Social choice and individual values*. New Haven: Yale University Press.

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(3): 265–290.

Berg, N., and G. Gigerenzer. 2010. As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas* 18(1): 133–165.

Bernheim, B.D., and A. Rangel. 2007. Behavioral public economics: Welfare and policy analysis with non-standard decision makers. In *Behavioral economics and its applications*, ed. P. Diamond and H. Vartiainen, 7–84. Princeton: Princeton University Press.

Buchanan, J.M. 1979. Natural and artifactual man. In *What should economists do?* ed. J.M. Buchanan, 93–112. Indianapolis: Liberty Fund.

Buchanan, J. M. (1982). Order defined in the process of its emergence. Reader's forum on Norman Barry's "The tradition of spontaneous order," *The forum* at the online library of liberty, available at http://www.econlib.org/library/Essays/LtrLbrty/bryRF1.html.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. Regulation for conservatives: Behavioral economics and the case for "Asymmetric Paternalism". *University of Pennsylvania Law Review* 151(3): 1211–1254.

Conly, S. 2012. *Against autonomy: Justifying coercive paternalism*. Cambridge: Cambridge University Press.

Cowen, T. 1991. Self-constraint versus self-liberation. *Ethics* 10(2): 360–373.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis, 159–165. New York: Wiley.

Frederick, S., G. Loewenstein, and T. O'Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40(2): 351–401.

Gruber, J., and B. Koszegi. 2001. Is addiction "Rational"? Theory and evidence. *Quarterly Journal of Economics* 116(4): 1261–1303.

Hayek, F.A. 2002. Competition as a discovery procedure. *Quarterly Journal of Economics* 5(3): 9–23.

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

Jolls, C., and C.R. Sunstein. 2006. Debiasing through law. *The Journal of Legal Studies* 35(1): 199–241.

Kahneman, D., J.L. Knetsch, and R.H. Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5(1): 193–206.

Loewenstein, G. 2000. Emotions in economic theory and economic behavior. *American Economic Review* 90(2): 426–432.

Manzini, P., and M. Mariotti. 2009. Choice over time. In *The handbook of rational and social choice: An overview of new foundations and applications*, ed. P. Anand and P.K. Pattanaik, 239–270. Oxford: Oxford University Press.

Mas-Colell, A., M.D. Whinston, and J.R. Green. 1995. *Microeconomic theory*. New York: Oxford University Press.

O'Donoghue, T., and M. Rabin. 2003. Studying optimal paternalism, illustrated by a model of sin taxes. *American Economic Review* 93(2): 186–191.

O'Donoghue, T., and M. Rabin. 2006. Optimal sin taxes. *Journal of Public Economics* 90(10–11): 1825–1849.

Redelmeier, D.A., P. Rozin, and D. Kahneman. 1993. Understanding patients' decisions: Cognitive and emotional perspectives. *Journal of the American Medical Association* 270(1): 72–76.

Rescher, N. 1988. *Rationality: A philosophical inquiry into the nature and the rationale of reason*. Oxford: Clarendon.

Rizzo, M.J., and D.G. Whitman. 2009. The knowledge problem of new paternalism. *Brigham Young University Law Review* 2009(4): 905–968.

Schnellenbach, J. 2012. Nudges and norms: On the political economy of soft paternalism. *European Journal of Political Economy* 28(2): 266–277.

Searle, J.R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Sugden, R. 2004. The opportunity criterion: Consumer sovereignty withou the assumption of coherent preferences. *The American Economic Review* 94(4): 1014–1033.

Sunstein, C.R. 2014. *Why nudge?* New Haven: Yale University Press.

Sunstein, C.R., and R.H. Thaler. 2003. Libertarian paternalism is not an oxymoron. *University of Chicago Law Review* 70(4): 1159–1202.

Thaler, R.H. 1991. *Quasi rational economics*. New York: Russell Sage.

Thaler, R.H., and C.R. Sunstein. 2003. Libertarian paternalism. *American Economic Review* 93(2): 175–179.

Thaler, R.H., and C.R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.

Uzawa, H. 1956. Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics* 8: 35–40.

Ware, B. 2012. *The top five regrets of the dying: A life transformed by the dearly departing*. Carlsbad: Hay House.