RHETM
Review of the History of Economic Thought and Methodology

Hands, D. Wade. 2024. "On the (non) History of Preference Purification in Modern Economics." *Review of the History of Economic Thought and Methodology* 1(1): 1–42. DOI: https://doi.org/10.16995/rhetm.17321

Open Library of Humanities

On the (non) History of Preference Purification in Modern Economics

D. Wade Hands, Professor Emeritus, Department of Economics, University of Puget Sound, Tacoma, WA, US, hands@pugestsound.edu

Economists have typically viewed an individual's economic choices as being tightly linked to their preferences, and in turn, their preferences being tightly linked to the welfare associated with those choices. But behavioral economics clearly drove a wedge between choice and preference, and thus, in turn, between choice and welfare. Trying to reconcile the choice-preference-welfare relationship came to be called the reconciliation problem and one of the main approaches to the problem has been called preference purification. But the presumption in the literature has been that preference purification only became an issue with the rise of behavioral economics. This paper will argue that is not the case. During the first part of the twentieth century when the ordinal utility theory of consumer choice was still in the early stages of development, there were many economists who thought about problematic preferences in ways that were quite similar to the way that preferences have been characterized in recent debates about preference purification. This paper will discuss the history of this literature in a way that emphasizes the difference between the situational context of this early research on ordinal utility and the quite different situational context of the recent debates on preference purification. The conclusion suggests how these differences in situational context prevented important similarities between the two literatures from being recognized.

> *... we will call this approach "preference purification." The essential idea is that when an individual's decisions are inconsistent with defensible assumptions about rational choice, those decisions can be treated as mistakes. The task for welfare economics is then to reconstruct the preferences that the individual would have acted on, had her reasoning not been distorted by whatever psychological mechanisms were responsible for the mistakes, and to use the satisfaction of these reconstructed preferences as a normative criterion. (Infante et al. 2016a, 1)*

1. Introduction

The Infante, Lecouteux and Sugden (2016a) paper on *preference purification* seems to have set off what became a substantial debate within certain areas of economics and the philosophy of economics.[1] The debate about preference purification is in turn a subset of a broader discussion about reconciling behavioral economics with Paretian/mainstream welfare economics. This wider debate is often called the *reconciliation problem*[2] and it has played a key role in the development of the (even broader) literature on *behavioral welfare economics* (hereafter BWE). The BWE literature attempts to re-conceptualize welfare in a way that accommodates the results of behavioral economics: particularly the evidence that individuals frequently do not behave in the way that standard economic theory suggests (i.e. they do not behave like *homo economicus*).

As one might expect, efforts to overhaul the way that economists have typically thought about welfare and the foundations of microeconomic policy since early in the twentieth century is an extremely difficult project. So far, the main consequences have been that many different alternatives have been proposed, there has been (and still is) significant disagreement, and no consensus seems to be anywhere in sight.[3]

---

[1] See for example: Beck (2023), Bernheim (2016; 2021), Dold (2018), Grill (2015), Grüne-Yanoff (2016; 2022), Hausman (2012; 2016; 2022), Infante, Lecouteux, and Sugden (2016b), Lecouteux (2021a; 2023), Rizzo and Whitman (2020), Sugden (2015; 2021), Thoma (2021), and Whitman and Rizzo (2015). Note that both economists and philosophers have contributed to this literature.

[2] This label was popularized by McQuillin and Sugden (2012). It is important to be clear that the reconciliation problem concerns the relationship between behavioral economics and Paretian welfare economics, and not the relationship between behavioral economics and neoclassical economics in general. There was certainly tension between neoclassical theory and behavioral economics in the early years – Grether and Plott (1979; 1982) for example – but it has faded more recently.

[3] For a sample of research that reflects the diversity within BWE and the related literature see: Bernheim (2009; 2016), Bernheim and Taubinsky (2018), Dold and Schubert (2018), Dold and Stanton (2021), Grüne-Yanoff and Hertwig (2016), Gul and Pesendorfer (2007), Hargreaves Heap (2013), Harrison and Ross (2023), Lecouteux (2021b), and Sugden (2010; 2019). Curiously, preference purification has been called the "new consensus" (Sugden 2019; Thoma 2021) even though there is no real consensus on the topic.

Although there are many different approaches to BWE it is possible to identify two fairly distinct strategies for addressing the controversy. One preserves the link between preference satisfaction and well-being[4] and attempts to reconcile the two by making various adjustments to the preference side of the problem; this relatively revisionist strategy typically involves the concept of, if not the term, preference purification.[5] One such revisionist approach is the version of BWE that has received the most attention in both the academic and public policy literature: the libertarian paternalism of Cass Sunstein and Richard Thaler.[6]

The second, more revolutionary, approach dissolves the reconciliation problem by severing the link between individual preference satisfaction and welfare. If preference satisfaction does not constitute welfare, then the behavioral economics research demonstrating that individuals often behave in ways that are inconsistent with preference satisfaction has no welfare implications and thus no reconciliation problem exists. Of course, this means that well-being must be defined differently than how neoclassical economists and standard textbooks have traditionally defined it, but many alternatives are available.[7]

It is important to note that despite the disagreement and lack of consensus, preference purification, the reconciliation problem, and BWE are extremely important topics. Research on these topics has clearly: helped to rekindle discussion about the nature of well-being and the foundations of welfare economics among economic theorists; demonstrated that experimentally derived empirical evidence can initiate substantive reconsideration of long-established and highly formalized economic theory, even welfare theory; encouraged an increase in the depth and breadth of interdisciplinary dialogue among the fields of economics, psychology, and philosophy;

---

[4] For the purposes of this paper welfare and well-being will be used interchangeably.

[5] As with purification in general, there are different purification processes and different names for the various purified products. In the case of preference purification, the most common term for the product is *true* preferences, but some of the other terms used in the literature include: clean, considered, corrected, informed, latent, laundered, pruned, rational, spruced-up, underlying, and welfare-relevant.

[6] See Sunstein and Thaler (2003) and Thaler and Sunstein (2003; 2009) for the original version, although the concept has evolved over time. The asymmetric paternalism of Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin (2003) is a similar, but not identical, approach.

[7] Some of these have emerged fairly recently and with the reconciliation problem in mind; one example is the opportunity-based framework of Robert Sugden (2010; 2019). Other approaches draw on ideas that are critical of the link between welfare and preference satisfaction, but do not focus directly on behavioral economics as the source of the schism; many of these are associated with the capability literature (Nussbaum 2011; Nussbaum and Sen 1993; Sen 1979; 1999; 2002). There are also non-preference approaches based on identity (personal persistence) and adaptive capabilities (Davis 2024).

and played a significant role in a variety of different public policy discussions around the world.

Given these remarks about the importance of such ideas, one might expect this paper to offer a new solution to the reconciliation problem and/or provide a new account of BWE. But that is not the case. Although debates about preference purification, the reconciliation problem, and BWE are all interrelated, this paper will, as much as possible, focus on preference purification. It will also take a historical approach to the topic, rather than examining preference purification from the perspective of economic theory, economic policy, or the philosophy of economics. The core motivation is that while research associated with preference purification is receiving a significant amount of attention, there is almost no historical examination of the role that preference purification-like concerns played in economics prior to the rise of behavioral economics.[8]

In the interest of clarity and manageability, let me briefly note a few things that will be presumed throughout the paper and/or that readers might expect to be discussed, but will not be.

- Although there will be very little discussion of the specific anomalies that have been identified within the behavioral economics literature,[9] it will be presumed that such anomalies are generally empirically reliable. There are certainly debates about the reliability of some behavioral anomalies, but they are not relevant here. Preference purification and the reconciliation problem are topics of debate precisely because the scholars involved accept that behavioral economics has

[8] When I say that there has been almost no historical research on topics related to preference purification, I certainly do not mean that there has not been historical research on behavioral economics in general. There has been a significant amount of historical research on early precursors to various behavioral economic ideas, for example: Ashraf, Camerer, and Loewenstein (2005) on Adam Smith, Bruni and Sugden (2007) on William Stanley Jevons and Francis Edgeworth, Hands (2023) on Frank Knight, Sugden (2021) on David Hume, and others. There is also an extensive historical literature on scholars from the mid-twentieth century who anticipated various behavioral insights such as James Duesenberry, Ward Edwards, George Katona, James March, Tibor Scitovsky, Herbert Simon and many others (see for example Camerer and Loewenstein 2004, Heukelom 2014, and Sent 2004). There is historical work on the behavioral decision theory that influenced Daniel Kahneman and Amos Tversky's early research on individual decision-making and rationality (e.g., Davis 2011; Heukelom 2014), as well as on many other specific topics. So no, the issue is not a general lack of historical research on behavioral economics; it is the lack of historical research directly related to preference purification-like concerns.

[9] There are many such anomalies, but most can be reduced to some version of context-dependency: framing, reference dependence, loss aversion, status quo bias, sunk costs, and to some extent constructed preferences. Others that are less easily defined in terms of context-dependency include social preferences and, in the case of risky choice, various judgment errors involving miscalculation or misperception of probabilities. See Camerer and Loewenstein (2004) for an early, but still quite useful discussion.

demonstrated that individuals often do not behave like *homo economicus*. If most of the experiments that led to the various behavioral anomalies were widely discredited, the literature on the reconciliation problem would not exist.

- In discussions concerning welfare economics, it will be presumed that individuals have preferences that causally determine – along with other factors like beliefs and constraints – the decisions that individuals make. While there is an extensive literature that argues for non-causal (instrumentalist, as-if, or black-box) accounts of preferences for the purposes of positive economics (the prediction and explanation of actual behavior), such a position is much more difficult to defend for the preferences involved in welfare economics.[10] The causal account is also appropriate here because it is typically a presupposition of the contemporary research on preference purification as well as the historical literature discussed below.
- Since the paper's main focus is preference purification there will be no further discussion of BWE that completely severs the link between preference and welfare. The literature on non-preference-based welfare is quite extensive and seems to be gaining ground. In fact, it is quite possible that at some point down the road a version of preference-free welfare will end up completely transforming how most economists think about well-being, but that said, this particular paper is not directly concerned with the future of welfare economics. It is concerned with the history of ideas related to preference purification and commitment to any type of preference purification implies that one believes, at least to some degree, that preferences are welfare relevant.
- Although the vast majority of research in behavioral economics is concerned with risky choice – expected utility theory and other approaches – the historical discussion that follows concerns rational choice under certainty, particularly budget-constrained utility-maximizing consumer choice (i.e., demand) theory.[11]

[10] For recent discussion of these issues, see Moscati (2024) for a defense of as-if modeling in decision theory and Grüne-Yanoff (2022) for the argument that causal preferences are required for BWE.

[11] The history of consumer choice theory and expected utility theory in economics is complicated. Here is a very abridged version. The change from cardinal conceptions of utility to ordinal utility (better or worse rather than quantitative magnitudes) in consumer choice theory took place during the first third of the twentieth century and focused on risk-free rational choice: particularly consumer choice theory. Many different economists played a role in the development of ordinal utility theory: Pareto (1909 [2014]), Slutsky (1915 [1952]), Hicks and Allen (1934), and many others. After the representation theorems in Debreu (1954) – which proved that any well-behaved (complete, transitive, and continuous) preference ordering could be represented by an ordinal utility function – the terms utility and preference became relatively interchangeable in consumer choice theory. The theory of decision-making under risk began much earlier and has taken many different forms: expected utility theory being historically the most influential. For a recent historical discussion of these topics see Moscati (2019; 2023). It should also be noted that the relationship between expected utility

> There are several reasons for this, but the main one is that even though expected utility theory has been around for much longer than consumer choice theory, the fact is that expected utility did not become the influential characterization of *homo economicus* that it is today until relatively late in the twentieth century, and consequently the economists who were thinking about theoretical issues related to preference purification during the first half of the twentieth century were doing so in the context of consumer choice theory. As Nicholas Georgescu-Roegen put it in 1958: "The *raison d'étre* of the theory of choice as a chapter of economics is *above all* the simplification it brings to the theory of demand." (157, emphasis added)

To summarize, the approach of this paper is to draw attention to some of the economic theorizing from the first half of the twentieth century that, although not using the term, touched upon issues directly related to the concept of preference purification. The general goals of the paper are to provide a better historical understanding of the various forces that contributed to the relevant debates (both historical and recent) and to emphasize that context matters, not only in behavioral economics, but also in the reception of what counts as interesting theory and relevant evidence in economic research. The more specific goal, and most original contribution to the literature, is to make the case that the problem of preference purification was not simply a product of the rise of behavioral economics – an empirical experiment-based research program that focused on individual decision-making in risky choice environments – but was also present in earlier neoclassical consumer choice theory, which was more deductive, typically involved continuous (often differentiable) functions, and concerned constrained individual choice under conditions of certainty.

With this introduction it is now possible to move forward into the historical discussion. The rest of the paper is organized as follows. Section 2 is a review of the many historical forces leading up to the current debates about preference purification and related issues; it also provides clarification of a number of the terms that are used in the literature. Sections 3–5 contain the main historical discussion. It begins with an introduction to the historical material in section 3 and is followed by two sections (4 & 5) on the work of economists who made contributions to ordinal utility theory, but also

---

theory and behavioral economics is quite involved and is still being debated. The dominant version of the story is that behavioral economists are generally critical of expected utility theory as a descriptive (positive, predictive, explanatory) theory, but consider it to be the appropriate normative benchmark for *how rational decisions ought to be made*. Or as one critic put it: "they rather uncritically accept the rules of axiomatic decision theory as the *norm* for all rational behavior, and blame mortals for not living up to this ideal" (Gigerenzer 2015, 365).

worked to improve the theory by modifying its psychological foundations in ways that, with hindsight, look similar to preference purification. Section 6 is a brief conclusion.

2. Normative, Descriptive, Preference Purification, Welfare, and All That

Perhaps the best way to begin this section is with a quote from the paper that popularized the reconciliation problem: McQuillin and Sugden (2012). It is a long quote, but a good starting place and one that we will come back to frequently in the following discussion.

> For at least the last three quarters of a century, both descriptive and normative economics have been based on assumptions about individual rationality. In descriptive economics, economic agents have been assumed to act as if seeking to satisfy preferences that are … stable, consistent, and context-independent. In normative analysis, economic institutions, projects or policies have been treated as justified to the extent that their outcomes are ranked highly in the preference orderings that agents have been assumed to possess … however, there have been increasingly evident signs that economics might be changing direction, towards what has come to be called the *behavioural* approach. There has been an accumulation of work which tests rationality assumptions at the individual level, often in controlled experiments, and finds systemic "anomalies" (that is, deviations from received theory) … These developments pose severe problems for normative economics. Standard theoretical results – most obviously, the fundamental theorems of welfare economics – assume that individuals have coherent preferences. (McQuillin and Sugden 2012, 553–54)

This quote provides a very nice summary of both the forces behind, and the implications for, the reconciliation problem. Nevertheless, I believe that a number of the terms used in the quote need to be examined before moving on to section 3.

The term "descriptive economics" in the second sentence probably seemed curious to many economists when the paper first appeared. Modern economists have traditionally used terms like "positive economics" or "economic science" for economic analysis that is aimed at explanation and/or prediction of observed economic behavior or events. "Description" seems to be a rather puzzling word since so much of neoclassical economics is extremely idealized and does not, in any obvious way, describe real-world phenomena. Of course there are extreme cases like Gerard Debreu's (1959) Bourbaki-inspired axiomatization of general equilibrium theory – a book the historian of economics Mark Blaug (2002, 27) once called "the most arid and pointless book in the entire literature of economics" – but even less formalized modeling in economics, including some at the introductory level, is often highly idealized and in many cases the

idealized aspects cannot be de-idealized and the model still provide the results it was constructed for.[12]

The fact is that until very recently the majority of economists would not have used the term "descriptive" for economic theorizing, but things have changed. There has been a "quiet revolution" (Hausman 2018, 196) and many parts of economics have taken an empirical turn[13] and are now more likely to describe real world economic phenomena. Some of this is perhaps a sign of the times with our vast quantities of empirical data and powerful computational capabilities, but some has also come about as a result of a slow but steady advance in experimental and other empirical techniques. That said, it should be noted that the research of Kahneman, Tversky, and the other psychologists who contributed to the development of behavioral economics considered "description" to be the primary cognitive goal of their research (Heukelom 2014). Given all this, the terms positive economics and descriptive economics will be used interchangeably in this paper. There may be some contexts where a distinction between these terms is important, but that does not seem to be the case for the topics examined here.

Continuing on in sentence two, we are told that traditional (i.e. pre-behavioral) economics assumed that economic agents act, or act as if, they have preferences that are: "stable, consistent, and context-independent." It is certainly possible that some economists said this, but it is not how consumer preferences were "traditionally" characterized in neoclassical economics. Examining the relevant sections of the influential theoretical texts on exchange and general equilibrium – for example Debreu (1959, Ch. 4), Arrow and Hahn (1971, Ch 4), the popular graduate level microeconomics text Mas-Colell, Whinston, and Green (1995, Ch. 3), or even the popular undergraduate text Varian (2014, Ch. 3) – we find preferences generally characterized by completeness, transitivity, and continuity as core assumptions with monotonicity (or nonsatiation) and (some version of) convexity added for purposes of deriving demand functions from budget-constrained utility-maximization.

So, why is this? Why is it that informed and well-respected contemporary behavioral economists would say that preferences in traditional consumer choice theory were typically assumed to be "stable, consistent, and context-independent"? My suggestion – supported in more detail below – is that it is because such assumptions would be sufficient to guarantee rational individual choice behavior (and thus prevent choice

[12] The idealization problem is an ongoing topic of discussion within the philosophy of economics literature. See Aydinonat (2018), Hoover (2023), Knuuttila and Morgan (2019), Mäki (1994), Marchionni (2017), Reiss (2012), Sugden (2009), and Ylikoski and Aydinonat (2014) for a sample of some of the various positions that have been taken on the issue.

[13] See for example Backhouse and Cherrier (2017a; 2017b), Biddle and Hamermesh (2017), Davis (2007), and Sugden (2008).

anomalies) in the experimental choice situations that matter to contemporary behavioral economists – that is to solve *their* problem – but they are projecting their conceptual concerns, in this case empirical experimental concerns, back on twentieth century neoclassical economists. But choice anomalies were not *the* problem for the neoclassical economists who standardized and stabilized consumer choice theory during the first half of the twentieth century. Those economists were not working in empirical and experimental environments; they assumed that economic agents maximized continuous, and generally differentiable, utility functions subject to a linear budget constraint and their primary goal was – again to be supported in more detail below – the derivation of well-behaved individual demand functions.[14]

The context of traditional theory is quite different than the context of contemporary behavioral economics. Not only are the standard restrictions on preferences different, but so is the whole setup for the analysis of individual choice. While the choice context of behavioral economics can be any experimental setup where individuals make discrete choices in a controlled environment, the choice context associated with neoclassical consumer choice theory was much more abstract/idealized. In addition to the different restrictions on preferences noted above, the analytical setup also included very specific constraints – particularly a linear budget constraint and competitive market prices (consumers being price-takers not price-makers).

On the other hand, while "stable, consistent, and context-independent" are relatively weak restrictions that are not *sufficient* to derive well-behaved consumer demand functions, they are structural features that stabilize preferences/utility functions in ways that accommodate the derivation of such demand functions – and it was that derivation, not circumventing behavioral anomalies, that motivated the traditional assumptions on consumer preferences. As Paul Samuelson (1947, 97) put it at the end of his discussion of consumer demand in *Foundations*: "their derivation is the whole end and purpose of our analysis of consumer behavior." Perhaps an example would help make this more clear.

Consider context-independence which is the assumption that (rightly) gets the most attention in behavioral economics and BWE. If a consumer acts rationally and actually "has," or always acts as if she/he has, a continuous utility function, then context-independence is automatic. Consider a very simple case. Suppose a consumer has well-behaved preferences that generate the utility function $U(x_1, x_2) = x_1x_2$ for two goods x_1

[14] This is certainly not to suggest that these neoclassical economists were entirely unconcerned about the empirical application or empirical foundations of consumer choice theory; it is just that the epistemic context – what counted as adequate empirical evidence to the majority of neoclassical economists — changed over the course of the twentieth century. This issue is discussed in more detail in Section 4.

and x_2, and maximizes that function subject to the standard linear budget constraint; in this case they will *always purchase* the two goods in the utility-maximizing quantities $x_1 = M/2p_1$ and $x_2 = M/2p_2$ (where p_1 and p_2 are prices and M income). If this consumer really has (or always acts as if she/he has) this utility function, and acts rationally on it, the mathematical structure of the consumer's problem prevents any kind of deviation from the optimal solution caused by context-dependency. The optimal consumption bundle *is the solution to a mathematical optimization problem*, and the solution to a math problem is context-independent; it doesn't matter where the choice is made, how x_1 and x_2 are arranged on the shelf, what endowment the individual started with, or anything else.

Of course, once one moves away from this extremely tight mathematical structure into a world where "choice" has much more latitude – such as the experimental world of behavioral economics – then consistent *homo economicus* behavior would require McQuillin and Sugden's restrictions on preferences: stability, consistency, and context-independence. So despite the surface mismatch, these assumptions were implicitly met in traditional modeling, since the austere mathematical structure of neoclassical choice theory eliminated the need for explicitly assuming context-independence because the structure of the theory automatically implied that such conditions would hold.

The bottom line, as behavioral economics has taught us, is that *context matters*, but it matters just as much about the choices that economists make in economic theorizing as it does in other forms of decision-making. The recent debate about preference purification was inspired by developments in behavioral economics, economic research that is experimental and based on discrete empirical data – a context in which assumptions like stability, consistency, and context-independence would need to be imposed to guarantee rational behavior and prevent the emergence of behavioral anomalies. On the other hand, the economic theorists working on ordinal utility-based consumer choice theory during the first half of the twentieth century were not starting from discrete empirical data, or thinking in terms of what rational choice would need to assume in order to avoid behavioral anomalies; they were thinking in terms of finding the minimal restrictions on preferences that would guarantee that budget-constrained utility maximization would support the derivation of well-behaved consumer demand functions and accommodate various comparative statics exercises. These are two *fundamentally different situational contexts* based on different motivations, as well as on the quite different technologies of derivation and inference available at the time. We should not be surprised that two quite different theoretical and epistemic contexts generate different reference points and thus different characterizations of what

"traditional" theory was, or must have been. It is a simple historical point, but one that doesn't get the attention it deserves in contemporary discussions about behavioral economics. This argument will be elaborated in more detail below.

So now consider the term "normative economics" in the third sentence. This is also a term whose typical usage among economists has changed from what it was during the middle of the twentieth century. Like "descriptive," it is a case where economists have slowly adopted the terminology of experimental psychologists, but the terminology was also well-established within other fields, particularly the interdisciplinary field of decision theory.[15] While many of the important contributions to decision theory were made by economists, it has only been during the last few decades that the majority of economists have started thinking about normative economics in the way it has typically been characterized in normative decision theory. So to help clarify these different meanings of the term "normative," it is again useful to start with some quotes: in this case quotes from two very influential twentieth century economists (although from different halves of the twentieth century).

The first is from Lionel Robbins's famous book on the nature and significance of economic science:

> Economics deals with ascertainable facts; *ethics with valuations and obligations*. The two fields of inquiry are not on the same plane of discourse. Between the generalisations of positive and normative studies there is a logical gulf fixed which no ingenuity can disguise and no juxtaposition in space and time can bridge over. (Robbins 1935, 148, emphasis added)

The second is from Richard Thaler's influential 1980 paper that played an important role in helping to persuade economists that Kahneman and Tversky's research, particularly Kahneman and Tversky (1979), directly addressed issues at the heart of neoclassical economics (like consumer choice theory). As Kahneman put it in his Nobel Lecture: "The core idea ... became useful to economics when Thaler (1980) used it to explain *riskless choices*" (Kahneman 2003, 1457, emphasis added).

---

[15] Even though risk-free consumer choice theory is a kind of decision theory, the term "decision theory" has traditionally meant risky rational choice theory: often, but certainly not exclusively, expected utility theory. Decision theory has deep philosophical and mathematical roots, but the explosion of research on decision theory that still covers the scholarly landscape happened in the decades following WWII. It was/is a very interdisciplinary field involving economists, philosophers, psychologists, mathematicians, and scholars from other fields. Although expected utility theory probably remains the cornerstone, it is also a very diverse field with many different approaches. See Binmore (2009), Davidson and Suppes (1957), Jeffrey (1965), Luce and Raiffa (1957), Savage (1954), and Suppes (1961) for a sample of the variety of different approaches to decision theory.

> 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 paper argues that exclusive reliance on the normative theory leads economists to make systematic, predictable errors in describing or forecasting consumer choices. (Thaler 1980, 39, emphasis added)

Notice that these two quotes employ entirely different uses of the term "normative." In the Robbins quote normative means *ethics* – moral valuations and obligations – while in the Thaler quote normative is about *rationality*: what rational consumers should do. Both are about what individuals ought to do, but the grounding of the obligation is quite different. In one case the term normative refers to what one *ought to do in order to be ethical/moral*, while in the other case normative refers to what one *ought to do in order to be rational*.

But not only is the term "normative" used in quite different ways, there is also a difference in the information provided about the particular obligation of concern. Robbins does not provide any details about the particular ethical obligations he is referring to, undoubtedly because he took it as given that ethics should be completely avoided in economic science and thus the question of which particular ethical view is involved is irrelevant. On the other hand, Thaler does provide information about the particular conception of rationality that grounds the normative obligation. It is the rationality of *homo economicus*: having well-ordered preferences and acting optimally on those preferences subject to the relevant constraints. It is not the rationality of Aristotle, Kant, Hegel, or even the rational prudence of Adam Smith, but the rationality of constrained-optimization-based utility theory.

Notice that unlike Robbins, Thaler is framing a vision of economic science in which both positive and normative economics play a role. But, and this is a very important step down the road to the reconciliation problem and preference purification, Thaler is also criticizing mainstream economics for its long-held practice of placing *homo economicus* at the heart of descriptive economics. He is arguing – pointing toward the extensive research in behavioral economics that will come in the decades that follow – that the utility maximization-based consumer choice theory that began with the neoclassical revolution in the 1870s has *not been successful* in predicting and explaining individual economic behavior and thus, by implication, should be replaced. Now while this seems to be a radical stance to take in 1980, it had long been the position of many heterodox economists, philosophers, and various critics of neoclassicism from the

other social sciences. Thaler also makes a surprising move by retaining *homo economicus* as the normative standard for individual choice behavior: "it describes what rational consumers should do." For Thaler, normative economics – not ethics, but normative with respect to (a particular version of) rationality – has an important role to play in economics. The mainstream theory that Thaler argued was a failure came to be harshly criticized, but at the same time promoted to the position of an ideal standard to be achieved. Its lack of success in the realm of what is, was transformed into an acceptable standard for what ought to be.

It should also be noted that Thaler was far from alone in this interpretation of the relationship between descriptive and normative behavioral science. It was a core methodological commitment of the behavioral decision research school that influenced Kahneman, Tversky, and many others who contributed to the development of behavioral economics, particularly the influential heuristics and biases program:

> The study of decisions addresses both normative and descriptive questions. The normative analysis is concerned with the nature of rationality and the logic of decision making. The descriptive analysis, in contrast, is concerned with people's beliefs and preferences as they are, not as they should be. (Kahneman and Tversky 2000, 1)

This interpretation of the descriptive-normative relationship as well as the characterization of utility-maximizing *homo economicus* as descriptively inadequate, but setting the normative standard for rational decision-making, became the dominant view within behavioral economics and for much of BWE. Acceptance of this set of ideas also provided the backdrop for the development of specific approaches to BWE: in particular the influential libertarian paternalism (hereafter LP) approach to behavioral interventions of Thaler and Sunstein.[16]

The LP literature is massive, continually expanding, and has many different branches – some methodologically focused and some quite theoretical – but the largest body of literature by far is concerned with practical application and public policy. Although it is important to stay on task about the history of preference purification, given the fact that LP lurks in the background of so much of the preference purification literature, some discussion of LP seems to be in order.

Neoclassical economists have traditionally discussed paternalism in the context of utility-maximizing behavior. Those who might need paternalistic help are those

[16] See note 6 for the primary references.

who have problematic preferences in the sense that their own utility maximization causes them harm: various addictions being obvious examples. But the anomalies of behavioral economics have opened the door to another way of thinking about paternalism. Behavioral anomalies drive a wedge between what individuals actually choose and what they would choose if they were fully informed, free of biases, and maximizing their true preferences; LP makes the case that various changes in the choice context (choice architecture) can, and should, be used to nudge individuals into more rational decision-making. Since standard welfare economics equates well-being with individual preference satisfaction, more rational behavior means higher preference satisfaction, and this in turn means increased well-being. Nudging people by changing the choice architecture in ways that help them correct mistakes and maximize their true (or latent) preferences, makes individuals better off as judged by themselves. As Robert Sugden explains:

> The implication is that what makes an individual better off 'as judged by himself' is defined by the preferences he would have revealed, had his decision-making not been affected by limitations of attention, information, cognitive ability or self-control. So Sunstein and Thaler's approach to normative economics treats context-dependent choices as the result of errors of reasoning. It requires the reconstruction of individual's *latent preferences* by simulating what they would have chosen, had their reasoning not been subject to these errors. This is preference purification. (Sugden 2015, 583)

Thaler and Sunstein use the terms *Humans* for those who make mistakes and do not act optimally on their true (i.e. purified) preferences and *Econs* for those who are mistake-free and act rationally on their true preferences.

> Whether or not they have ever studied economics, many people seem at least implicitly committed to the idea of *homo economicus*, or economic man – the notion that each of us thinks and chooses unfailingly well, and thus fits within the textbook picture of human beings offered by economists ... But the folks that we know are not like that ... To keep our Latin usage to a minimum we will hereafter refer to ... Econs and Humans. (Thaler and Sunstein 2009, 7)

The argument is that both Econs and Humans have "an ideally rational agent skulking within" (Hausman 2016, 26), but because of various heuristics and biases Humans make mistakes and fail to satisfy their true preferences; Humans are thus "faulty

Econs" (Infante et al. 2016a, 23).[17] LP can be used to nudge them back into rational behavior – making them: i) behave like Econs rather than Humans, ii) satisfy their true preferences, iii) have higher levels of individual preference satisfaction and thus higher welfare, and iv) be better off as judged by themselves.

Till Grüne-Yanoff provides a nice summary many of the different arguments discussed thus far in this section:

> Thaler early on (1980) proposed to distinguish descriptive models of consumer choice from normative ones. The former predict what consumers actually do, while the latter describe what rational consumers should do. Thaler's explicit aim in 1980 was to improve the descriptive models by revising them in the light of the recent experimental evidence ... Importantly, he left the normative model untouched, in effect asserting that the standard economic models of choice were normatively valid. The thus-opened chasm between descriptive and normative models led behavioral scientists to think about ways to lead people back from how they actually behave to how they should behave, and hence provided both motivation and justification for behavioral interventions. This was a new role for decision theory—as long as its models were considered both descriptively and normatively adequate at the same time, this question simply did not arise. (Grüne-Yanoff 2017, 69)

Finally, let us return to the McQuillin and Sugden quote and consider the last few sentences which note the impact that behavioral anomalies have on traditional Paretian welfare economics, in particular the *First Fundamental Theorem of welfare economics* which is often considered to be the most important theoretical result in modern economics.

During the period immediately following the neoclassical revolution in the 1870s, most – particularly British – neoclassicals where committed to hedonistic utilitarianism with respect to both positive economics (the pursuit of hedonistic utility motivated individual economic behavior) and ethically normative welfare economics (hedonistic utility determined what the society ought to do to bring about the most good). However, for a variety of reasons – some practical, some epistemic, some political, and a host

[17] Since there are two parts to rational choice – having well-behaved preferences and acting rationally/optimally on those preferences – it seems that purification might be needed to correct errors in optimizing rather than, or at least in addition to, modifying the agent's problematic preferences. It has been argued (Hands 2020) that if one is taking Econs and the Econ-Human relationship seriously, purification should be focused on errors in optimization/computation rather than problems with the agent's preferences. This said, the vast majority of the literature is about preference purification, not optimization purification, so this paper will be exclusively concerned with purification of preferences/utility functions.

of other factors – hedonistic cardinal utility gave way to ordinal utility in positive economics early in the twentieth century, and while ordinal utility theory was viewed as a significant improvement in the realm of positive consumer choice theory, it lost the straightforward link to welfare economics since it was no longer possible to add up pleasures and pains in the way that hedonistic utility had provided.

The solution which eventually emerged and became (and still is) the standard characterization of welfare efficiency in mainstream theory was the concept of *Pareto efficiency*: an allocation where it is impossible to make a re-allocation that would make one individual better off without making someone else worse off.[18] If there exists a Pareto improving re-allocation – one that can make at least one person better off without making anyone else worse off – then the reallocation (it seemed obvious) should be made. When all Pareto improvements have been exhausted the allocation is Pareto efficient – and Pareto efficiency has remained the normative baseline for mainstream welfare economics since the middle of the twentieth century.

The *first fundamental theorem* links welfare efficiency directly with Walrasian general equilibrium theory. It says that *every competitive equilibrium allocation is Pareto efficient*. The first modern mathematical presentations of the theorem came from the independent work of Arrow (1951) and Debreu (1951) and they were both connected to, and reinforced by, the proof of the existence of Walrasian competitive equilibrium in Arrow and Debreu (1954).[19]

So to see how exactly behavioral anomalies pose, to use McQuillin and Sugden's words, "severe problems" for the first fundamental theorem, it is useful to review why there did not seem to be any such problems in mid-twentieth century neoclassical theory. In traditional theory, each individual maximized preferences subject to the relevant constraints, so choices always accurately reflected preferences. But since welfare was simply preference satisfaction, choices also accurately reflected welfare. Under the assumptions of standard theory, preferences provided a tight linkage between choice and welfare.

But now enter behavioral economics where what people choose may not reflect their true preferences and thus will not reflect the level of welfare they would have if they had acted in a fully rational way on true preferences. But this means that the

[18] The Pareto efficiency condition was originally called *Pareto Optimality*, but as emphasized by Pareto himself (Tarascio 1969) as well as by Bergson (1938; 1954), Samuelson (1947; 1981), and many others, there is nothing "optimal" about Pareto Optimality. Since even in an extremely simple model of pure exchange with only two individuals and two goods there are an infinite number of Pareto allocations, it was argued that the term "optimality" should be reserved for the *best* allocation. This was a major motivation for the change to the term "Pareto Efficiency."

[19] See Düppe and Weintraub (2016) and Weintraub (1983) for a detailed historical discussion.

demands they have for various goods, and thus their consumption level at equilibrium prices, will be different from what they would have been by acting rationally on true preferences. And this in turn means that their choices might not be Pareto efficient; it may be possible that LP nudging or other interventions could increase the preference satisfaction of individuals who are not acting rationally without reducing the welfare of the individuals who are acting rationally. Or to use Thaler and Sunstein's terminology, it may be possible to make Humans better off without making the Econs worse off. Thus LP nudging could bring about a Pareto improvement that competitive equilibrium driven by unpurified preferences would not: hence the severe problem for welfare economics.

But this argument is not restricted to LP or the post-behavioral economics literature. It can be found – minus the language about Humans, Econs, or preference purification of course – in discussions about the choice-preference and preference-welfare relationship in the general equilibrium literature of the 1970s. For example, Amartya Sen noted the following in an inaugural lecture at the London School of Economics in 1973:[20]

> ... this problem has an important bearing on normative problems of resource allocation formulated in terms of the dual link between choice and preference and between preference and welfare. The type of behaviour in question drives a wedge between

[20] Sen's lecture focuses on revealed preference theory and much of the contemporary literature on BWE frames the concerns about behavioral anomalies as a tension between revealed preferences – the preferences revealed by the empirical revealed preference techniques that developed out of the important work of Afriat (1967) – and true, or welfare-relevant, preferences. The problem is that the practice seems to suggest that empirical revealed preference techniques that begin with price-quantity data and use versions of the generalized axiom of revealed preference (GARP) to rationalize it, necessarily "reveal" the utility function/preferences behind the choice data. However, that is not what empirical revealed preference theory does (it might, but it need not). GARP-based empirical techniques *rationalize* price-quantity data – they find a utility function that is *consistent* with the data – one that if maximized would generate the relevant price-quantity data, not necessarily the utility function that *did* generate the data (if any did). Of course, providing a utility function that *could have been* behind the choices can often provide very useful information in the context of positive economics: particularly in applications to business and institutional decision-making. But for the purposes of welfare economics mere rationalization doesn't seem to provide any normative bite – either for rationality or ethics. As Paul Samuelson put it in the last paragraph of the original paper on revealed preference theory: "In closing I should like to state my personal opinion that nothing said here in the field of consumer's behaviour affects in any way or touches upon at any point the problem of welfare economics ..." (Samuelson 1938, 71). Given this, I will use the term "manifest preferences" (following Harsanyi 1977), rather than "revealed preferences," for what seems to be behind the price-quantity evidence. This means that the door is left open for the application of empirical revealed preference techniques, but it also leaves the door open for other sources of information about what people prefer. So for the remainder of this paper the preferences Sen is referring to as "defined in such a way as to preserve its correspondence with choice" will be *manifest preferences* and those "defined so as to keep it in line with welfare" are purified or true preferences. See Hands (2013) for a more detailed discussion of revealed preference and the issues of concern here.

> choice and welfare, and this is of relevance to general equilibrium theory as well as to other aspects of normative economics. Preference can be defined in such a way as to preserve its correspondence with choice, or defined so as to keep it in line with welfare ... but it is not in general possible to guarantee both simultaneously. Something has to give at one place or the other. (Sen 1973, 259)

Of course, Sen was not thinking in terms of the heuristics and biases anomalies associated with behavioral economics. He seemed to be mostly concerned about altruistic feelings that drive a wedge between choice and welfare – the ideas about commitment and sympathy that he examined in a number of works, particularly Sen (1977) – but the impact on choice and welfare, as well as the impact on the first fundamental theorem, is of the same nature as that associated with behavioral anomalies.

So with all this we can move beyond background, definitions, and organizational issues and begin to discuss some of the economic literature from the first half of the twentieth century that proposed versions of consumer choice theory that seem to address some of the same issues as preference purification.

3. The Forgotten History of Preference Purification

Perhaps it is useful to think about preference purification like one thinks about other forms of purification, say water purification. Why does one purify drinking water? It is typically to eliminate, or at least reduce, some of the impurities that reduce the well-being one gets from drinking it. So too with preference purification. If various preference impurities – context-dependence, instability, etc. – are reduced, or eliminated, by preference purification, then the preference satisfaction and thus well-being of the individual who acts on (or acts as-if they are acting on, or is nudged into acting on) such preferences will increase. But this motivation for purification is, at least in principle, what was also behind the imposition of the standard preference restrictions used in mid-twentieth century neoclassical consumer choice theory (although it is doubtful at the time that anyone thought about it in this way). If preferences are not complete, there are bundles of commodities that the consumer will not be able to value; if preferences are not transitive, the individual could end up in a cycle which would prevent a choice from being made, or be subject to a money pump, and so forth. Such impurities can distort demand functions, equilibrium prices, and the Pareto efficiency of the equilibrium allocation, thus reducing well-being. In general, preference purification is the elimination, or at least reduction, of welfare reducing impurities by imposing restrictions on preferences so that choices are more welfare-focused. Thinking about preference purification in this way reveals a symmetry

between contemporary preference purification and the tacit preference purification in neoclassical demand theory associated with the restrictions that lead to well-behaved individual utility functions. This symmetry will be historically examined in sections 4 & 5 below.

However, it is also useful to note that the idea of preference purification has been a concern in fields other than economics and decision theory. A good example is ethics, hedonistic utilitarian ethics in particular. There the issue is primarily about what kinds of preferences should "count" toward the greatest happiness for the greatest number. It seems that the happiness that someone might get from torturing others should not be added into the overall happiness of the society (even if the sum of the pain imposed on others is less than the happiness the torturer receives). But this is not just an issue with hedonistic utilitarianism; the ethical issues seem to be just as relevant in *any* ethics where the good is based on individual preference satisfaction. Daniel Hausman and Michael McPherson (2006, 125–27) point out a number of problems associated with individual preference satisfaction-based theories of social welfare. These preference impurities "demand that one discriminate among preferences" (125), that is, to allow some things into the social welfare function and not others. Some that should be eliminated are: "idiosyncratic or obnoxious" preferences (125), "expensive tastes" (126), "racist, sadistic, and other antisocial preferences," and preferences that were formed by "previous coercion or manipulation" (127), but there are many others. The fact that one must "discriminate among preferences" in moral assessments necessarily involves some variation of preference purification. It is preference purification in the interest of morality and social welfare, rather than preference purification in the interest of rationality and individual welfare, but a version of preference purification nonetheless.

So we see there is the possibility of preference purification in the interest of rationality as in LP and certain other versions of BWE, and there is also preference purification in the interest of morality as in various arguments to prune/launder social preferences. But why not both? One example of this is in the work of John C. Harsanyi. He explained that preference-based social welfare requires two stages of preference purification: first to eliminate things that interfere with rational decision-making and second to eliminate things that interfere with morality. He explains this in Harsanyi (1977) and even uses the term "true preferences":

> All we have to do is to distinguish between a person's manifest preferences and his true preferences. His manifest preferences are his actual preferences as manifested by his observed behavior, including preferences possibly based on erroneous factual

> beliefs, or on careless logical analysis, or on strong emotions that at the moment greatly hinder rational choice. In contrast, a person's true preferences are the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice. Given this distinction, a person's rational wants are those consistent with his true preferences ... whereas irrational wants are those that fail this test.
>
> In my opinion social utility must be defined in terms of people's true preference rather than in terms of their manifest preferences." (Harsanyi 1977, 646)

But on the next page he adds another layer of purification:

> I have argued that, in defining the concept of social utility, people's irrational preferences must be replaced by what I have called their true preferences. But I think we have to go even further than this: some preferences, which may very well be their 'true' preferences under my definition, must be altogether excluded from our social-utility function. In particular, we must exclude all clearly antisocial preferences, such as sadism, envy, resentment, and malice ... The part of ... personality that harbors these hostile antisocial feeling must be excluded from membership, and has no claim for a hearing when it comes to defining our concept of social utility. (ibid., 647)

While this demonstrates that ideas about true preferences and preference purification were being discussed in the scholarly literature before the development of behavioral economics, it also makes it clear that such purification can be layered, at least in theory, with one aimed at rationality and the other at morality.[21]

Although explicit recognition of the concept of purifying preferences in the interest of rationality and welfare was extremely rare in mainstream economic theorizing prior to behavioral economics, it was not totally absent. For example, Harold Hotelling noted that individual preferences can provide poor welfare guidance when compared to that provided by various experts, which seems like – without using the LP terminology of course – a justification for nudging people away from their actual/Human preferences and toward more welfare-efficient purified preferences. As Hotelling explained:

> Preference and demand functions and consumers' surpluses are commonly understood to refer to people's actual preferences and choices. Sometimes people do not

[21] See Hédoin (2015) for a discussion of how this, and related concerns, undermine the LP approach to paternalism and other aspects of BWE.

> make their choices rationally and consistently. There are situations in which the state, or a consumers' cooperative, or a consumers' information, research, and testing bureau can tell us what to consume better than we can judge independently ...
>
> If we are to consider such systems of planning and allocation, or the consumers' choices that would result from improved information on their part, then it is appropriate to take as utility or preference functions something based not on what consumers have been observed to do, nor yet on what they say when asked what they will, do, but rather on what they ought to do if they were entirely rational and well-informed. (Wold et al. 1949, 188)[22]

Such arguments of course suggest replacing individual decision-making with more socially guided decision-making – and thus are not literally preference purification – but as was the case with more philosophical arguments like Harsanyi's, they certainly emphasize the complexity associated with reconciling preference, choice and welfare.

These arguments are certainly interesting and help clarify the issues involved, but for the remainder of this paper, preference purification will be restricted, as it is in most of the literature that uses the term, to that associated with the reconciliation problem and BWE (e.g. to make Human preferences closer to Econ preferences). The next two sections will focus on efforts to tweak the ordinal utility-based theory of consumer choice in directions that bear a strong family resemblance to the preference purification currently being debated in the BWE literature.

4. Pareto on the Order of Consumption and Routine

Vilfredo Pareto (1843–1923) was a key figure in igniting the ordinal revolution in the early twentieth century and produced an extraordinary amount of research; he was "perhaps the most prolific economist who has ever lived" (Chipman 1976, 66).[23] Perhaps the best place to begin a discussion of how Pareto's characterization of utility relates to preference purification is with his discussion of the consumer's *order of consumption* or *consumption path.* This idea seems to be consistent with casual empiricism (then and now). For the majority of people their total utility is higher if they consume their main course (M) before their dessert (D), and lower if they consume D before M. As Pareto put it in his "earliest contribution to utility theory in 1892" (Chipman 1976, 67):

---

[22] I would like to thank Spencer Banzhaf for drawing my attention to this reference.

[23] Given that Pareto's work was so early in the ordinal revolution it is not surprising that it contained various inconsistencies with respect to cardinal and ordinal utility. See for example: Chipman (1976), Giocoli (2003), Mandler (1999), McLure (2005), Samuelson (2005), Stigler (1950), and Weber (2001).

> It is indeed evident that the pleasure afforded by a meal is not the same if one eats it in the order to which one is used, or if one started instead with the coffee and finished with the soup. (Pareto 1892–93 [2007], 104).

And much later in the *Manual*:

> Obviously, one does not experience the same enjoyment if one eats the soup at the beginning of the meal and the dessert at the end as if one begins with the dessert and ends with the soup. The order of consumption would thus have to be taken into account, … (Pareto 1909 [2014], 126)

However straightforward it may seem, the order of consumption is problematic for any utility-based theory of consumer choice since having different levels of utility associated with different orders of consumption for the same quantities of two (or more) goods is inconsistent with the concept of a utility function; a functional relationship requires that each independent variable (quantities of goods) be associated with one and only one dependent variable (utility). If the consumer has a utility function then the final level of utility will be the same if the final quantities of the goods are the same, but if different consumption orders produce different final levels of utility even though the final quantities of the goods are identical, then the different paths create a context-dependency that affects the consumer's level of individual preference satisfaction. This is a conflict between Pareto's theory of ophelimity/utility[24] (i.e. his characterization of Econ behavior) and what he considered to be the obvious fact that consumers experience different levels of pleasure for different orders of consumption. In the same way that behavioral economists associate unstable, inconsistent, and context-dependent preferences with less-than-fully-rational Human behavior and thus a reduction in individual preference satisfaction and welfare, those experiencing different levels of utility for different consumption paths have a similar break in the preference → choice → preference-satisfaction → welfare relationship.[25] This means that the order

[24] Pareto used the term "ophelimity" in his discussion of individual choice and demand rather than "utility" (Pareto 1909 [2014], 77–79). The distinction is generally characterized as the difference between purely economic utility and social utility: "ophelimity refers to a system of strictly economic forces, a system which constitutes a subsystem of the total social system. Utility refers to the total social system" (Tarascio 1969, 1). Since the discussion here involves individual preference satisfaction and individual choice, the term utility will be sufficient. The main reason the word "ophelimity" is introduced here is so it will not come as a complete surprise when it appears in quotes.

[25] Technically, when discussing Pareto's choice theory, one should stop at various context dependencies decreasing individual preference satisfaction and not take the final step to a reduction in welfare. For Pareto welfare was social, not individual, and social welfare for Pareto necessarily involved non-economic considerations. See for example Tarascio (1969, 4): "He argued that before economists can speak of a theory of policy, they must either expand the scope of their positive researches to include non-economic phenomena, or they must supplement economic theory with the theories

of consumption problem, like many anomalies identified by behavioral economists, drives a wedge between choice and individual satisfaction in ways that cause serious difficulties for traditional welfare analysis and open the door to arguments for nudging the consumer in the direction of their true, i.e. purified preferences and the associated choices.[26]

Pareto's order of consumption problem can be seen as a subspecies of the broad family of context dependencies identified by behavioral economists, but it is also useful to draw attention to one particular member of that family and discuss the close similarity between Pareto's discussion of the order of consumption and the important concept of a *reference point* in behavioral economics. If a reference point matters – there is reference dependence – then where the consumer starts (or takes as a reference point) will have an impact on the value of the final consumption bundle.[27] Different reference points generate different final values for the same final states.

To see the similarity between reference points and consumption paths, consider the example of *loss aversion*, one particular example of reference dependence that has been discussed extensively in the behavioral economics literature: "One of the most powerful findings of behavioral economics is 'loss aversion,' the psychological tendency to feel losses more acutely than gains" (Thaler 2017, 1801).

Assume the consumer starts with the quantities (x_0, y_0) of two goods x and y and increase the quantity of both goods by a small amount (Δ) so the consumer has (x_1, y_1) where $x_1 = x_0 + \Delta$ and $y_1 = y_0 + \Delta$. Assuming monotonic preferences they will have a

---

of other social science disciplines which deal with non-economic phenomena … Pareto's views on the interdependency of social phenomena find their most important illustration in his 'welfare' theory.". This said, preference purification works in the same way as a response to the order of consumption problem as it does in response to context-dependency. It is just that, for Pareto, the problem only concerns less than optimal preference satisfaction and not what he considered "welfare." There is a debate about the relationship between Pareto's conception of social welfare and the later twentieth century literature on social welfare – see for example Bergson (1983) and Chipman (1976) – but our concern here is only about individual welfare.

[26] Pareto actually discussed two separate order-like issues. One was the order of consumption problem: the M before or after D problem. But he also discussed path of consumption problems, where the path through the choice space by which the optimal bundle is reached may change the level of satisfaction at the optimal bundle. Pareto's primary discussion of these issues was in the "celebrated but mysterious" (Hicks and Allen 1934, Part I, 53) paper on "nonclosed cycles" (Pareto, 1906 [1971]) where he tried to work around both problems simultaneously. Keeping the two problems separate is not really necessary here. See Hands (2006, 161–62) for a more detailed discussion of the path/order issue.

[27] The concept of a reference point is ubiquitous within the behavioral economics literature and it is the foundation for a number of important behavioral anomalies. For example, it is key to the argument for prospect theory in the paper often considered to be the initial stimulus for behavioral economics (Kahneman and Tversky 1979); it is a major theme in Kahneman's Nobel prize lecture (Kahneman 2003) and it is peppered throughout most behavioral economics textbooks, for example Dhami (2016). Also, since the discussion here is concerned with risk-free consumer choice, it is important to note that there is a significant literature on reference effects in that context as well: including Kahneman, Knetsch, and Thaler (1991), Knetsch (1989; 1992), Munro and Sugden (2003), Tversky and Kahneman (1991), and others.

higher level of utility at $U(x_1, y_1)$ than at $U(x_0, y_0)$, i.e. $U(x_1, y_1) > U(x_0. y_0)$. Now suppose the consumer goes back to the original consumption level (x_0, y_0), but exhibits loss aversion. With loss aversion the utility gain associated with the initial increase of (Δ, Δ) will be less than the utility loss from the following decrease of (Δ, Δ) and the consumer will end up back at the original point with a level of utility that is less than the utility at the initial level. So the quantity (x_0, y_0) will give two different levels of utility for the same quantities of the goods because two different reference points are involved. This means, like in the case of different consumption paths, that the utility function does not exist and context – either consumption path or reference point – has an impact on individual preference satisfaction. Pareto focused more on order of consumption than reference points, but the difficulty that utility theory faces is fundamentally the same.

The bottom line is that maximization of true utility is purely consequentialist in the sense that only the consequences – i.e. outcomes, final states, etc. – of choice are relevant. Econs do not veer off the target of true utility maximization by distractions like the order of consumption any more than they veer off their maxU target because of the various context-dependencies identified by behavioral economics. Humans, on the other hand, often exhibit context-dependencies of various sorts which drives a wedge between Human decision-making and what a fully rational Econ would do. This sets up the reconciliation problem and makes the task of welfare analysis extremely difficult whether the "analysis" involves relatively abstract questions (like the first fundamental theorem), more everyday practical applications of Paretian welfare economics to public policy, or the application of specific approaches to BWE such as LP. Given the results of most behavioral economic research – which has been "to focus on a few important ways in which humans diverge from *homo economicus*" (Thaler 2017, 1800) – welfare analysis will require the reconstruction of "the preferences that the individual would have acted on, had her reasoning not been distorted by whatever psychological mechanisms were responsible for the mistakes, and to use the satisfaction of these reconstructed preferences as a normative criterion" (epigraph). In other words, these tasks (assuming one stays in the welfare = preference satisfaction tradition) require some version of preference purification.

Since Pareto found the wedge between what observation/introspection told him consumers do and his characterization of Econ choice to be problematic, he faced a reconciliation problem that was similar to BWE. This means that finding a way to circumvent the order of consumption problem – recall it was a problem grounded in what Pareto saw as the way real individuals behave – is a version of preference purification. So what did Pareto offer as a solution to the problem? The fact is that Pareto suggested a

number of different solutions – some in substantive detail and some more as passing remarks – but I will only discuss the one that had the greatest impact on the later literature as well as being the one where the similarity to preference purification is most clear.

Pareto's solution, and the solution for many later economists, was to restrict the consumer to *routine* or *repeated actions.* Pareto presumed that the consumer would correct for mistakes and converge to the path with the highest level of preference satisfaction; once this routine is established, choice will be mistake-free and thus context-independent. This makes the establishment of a routine as a form of preference purification (or at least context purification).

> We shall study logical actions repeated a great number of times, by which men procure things that satisfy their tastes ... this allows us to presume the link between these actions is a logical one. A man who purchases a certain type of food for the first time may buy more of it than is necessary to satisfy his tastes ... But, when making a second purchase, he will correct his mistake, at least in part, and so on, until little by little, he obtains exactly the quantity he wants. We shall consider him when he has reached this situation. (Pareto 1909 [2014], 72)[28]

Pareto's routine restriction was adopted by many economists in the 1920s and 1930s. A good example is Henry Schultz who took it as fundamental for the theoretical foundations of his statistical work on demand theory. Here it provides clear explanation of how it eliminates the order of consumption problem:

> What, then, can be done about the difficulty presented by the order of consumption which appears to undermine the very basis of our theory? It seems to me that the answer to this question is essentially at hand in the fact that the economic theory can approximate the facts of economic experience only if there is a routine in economic affairs ... when there is no routine, there can be no economic law. But if it is reasonable to assume a routine, is it not also reasonable to assume that the order in which the various courses of a dinner are consumed is known? It appears therefore, that too much attention has been attached in utility analysis to the problem of the order of consumption. Although it was ... discussed at length by Pareto, it has little or no significance in an economy dominated by routine. (Schultz 1938, 17)

In closing this section, perhaps it is appropriate to move beyond simply talking about what various early twentieth century economists argued and make a few more general

[28] See the discussion in Bee and Desmarais-Tremblay (2023, 26).

remarks. Schultz, like Pareto, considered routine as a guarantee that preferences and thus the associated utility function would be "stable" – that is, with repeated choice there would be "no significant changes in tastes and desires of the consumers" (Schultz 1938, 65) – which suggests they believed that if such assumptions were not made, preferences would frequently be changing. But as discussed early in section 2, the ordinal utility theory that became standard during the middle of the twentieth century would focus on the stronger assumptions that were needed for the derivation of consumer demand functions (such as completeness, transitivity, and continuity) and did not need to explicitly assume stability because the mathematical structure of the later theory guaranteed the stability of preferences and utility functions. So why did Schultz need to specify routine explicitly while later neoclassical theorists did not? This is a complex question, but the short answer is because Schultz was still living with the Pareto-eye view of what preferences had to be for scientific economics. Schulz, like Pareto, was more explicit about starting with the observational facts of choice (which are troubled by instability and context-dependency), while the later economists had much less concern about such issues and much more about deriving the various implications of ordinal utility theory. Once it became acceptable to simply start with a continuous utility function, budget-constrained ordinal utility maximization became standard theory and destabilizing effects like the order of consumption – and the need for explicit assumptions like routine to get around them – slowly faded out of the mainstream economics literature. But of course, in the last few decades this has all changed. Individual choice experiments matter again and we are back in the heat of the same types of debates that were occurring early in the twentieth century.

5. Integrability, Order of Consumption, and Preference Purification

This section, while still concerned with the relationship between early modern consumer choice theory and recent debates about preference purification, will take a different approach than the previous section. Rather than focusing on an example of early consumer choice theory that, with hindsight, can be seen as a version of preference purification, I will discuss the literature on integrability and non-integrability because it is *intertwined* with ideas that are related to preference purification. The distinction between "is a version of" and "is intertwined with" is subtle, but important.

Although the first serious discussion of integrability in economics was Antonelli (1886), Pareto was certainly the one who ignited the flurry of interest in integrability early in the twentieth century; it was a problem of great importance to him and "he dedicated a good portion of his energies to it in his last works on pure economics" (Bruni 2002, 26). The attention that Pareto and other mathematically-oriented economists of

the period gave to the integrability problem seems very strange from the viewpoint of contemporary economics. Even though the ordinal utility theory that stabilized to become "the" theory of consumer choice during the middle of the twentieth century was conceptually similar to that of the earlier period, the topic of integrability essentially disappeared from the economics literature. While it is not unusual for particular ideas to disappear from economic theory, it is very unusual for it to happen in neoclassical theory when the eliminated idea was consistent with both individual maximization and competitive equilibrium, and also employed the same set of mathematical tools as the standard theory. Nevertheless, that is case for integrability in consumer choice theory. While there are many interesting facets to its disappearance, the connection here is that integrability (or more accurately non-integrability) was believed by Pareto and many others to be deeply "intertwined with" the order of consumption problem. So how exactly are these two sets of ideas intertwined?

Integrability in demand theory is a mathematical property of a well-behaved functional relationship between the consumed/purchased quantities that guarantees the existence of an associated utility function. For Pareto, and most of the economists writing about integrability during the next few decades after the *Manual*, it was necessary that the relevant well-behaved functional relationship be observable, and for Pareto the indifference line (indifference curve) was such a well-behaved functional relationship: "a concept that is given directly by experience" (Pareto1909 [2014, 309). The problem is that, while the relationship between integrability and indifference curves was widely discussed in Pareto's economic works, he never settled on a single explanation of exactly how it is that indifference curves were experientially observable. Even in the *Manual*, Pareto's work where these issues were given the most attention, it is unclear whether "experience" reveals an entire indifference curve, parts of an indifference curve, or only the little local tangents associated with the marginal rate of substitution.[29] Since the indifference curve is involved in the integration process to derive the utility function, confusion about the former created confusion about the latter. This of course makes a detailed discussion of integrability and indifference in the work of Pareto and his immediate successors rather messy.

However, a detailed discussion of Pareto's conception of integrability is not necessary for our purposes here since it is not the mathematics of integrability that is of interest, but rather how integrability was "intertwined with" the order of consumption. The bottom line is that Pareto equated the order of consumption with the order of

[29] See Bruni (2010, 99–100), Montesano (2006), or Montesano's "Notes to the French Appendix," pp. 621–659 of Pareto (1909 [2014]) for discussion of these issues.

integration. If the order of consumption is independent of path, then any consumption path starting at, say, (x_0, y_0) and ending at (x_1, y_1) will end up with the same final level of utility. In his discussion of integrability Pareto characterized indifference curves in total differential form, and if the total differential is exact (has an integrating factor of 1) then a solution will always exist and the associated differential equation can be integrated to recover the underlying utility function. In the case of only two goods there is no integrability problem and a utility function always exists (although it need not be unique unless the differential is exact). If three or more goods are involved, then the differential may not be integrable and the utility function need not exist.[30]

The final point that brings integrability and order of consumption together is that integrability has implications for the associated line integral. In particular, if the differential equation is exact, then the line integral has the property of being independent of path; it only depends on the starting and end points. But Pareto equated the order of consumption with the path of integration, and thus equated integrability with the value of a consumption bundle being independent of the order of consumption. As Aldo Montesano explained in his "Notes to the French Appendix" in the 2014 edition of the *Manual*:

> Pareto calls the integration path "order of consumption." When ... he writes "the order of consumption does not affect the consumption choice," he means that the line integral does not depend on the path, but only on the starting and ending points; the field is therefore conservative and the differential equation ... can be integrated and its integral is the scalar field represented by the ophelimity index function. When Pareto writes "the order of consumption affects the consumption choice," he means that the equation ... cannot be integrated. (Pareto 1909 [2014], 622)

This shows why integrability – or more accurately non-integrability – was so problematic for Pareto and many of the other mathematical economists who followed in his footsteps. Given Pareto's identification of the integration path with the order of consumption, non-integrability implies that the order of consumption "was not a matter of indifference"[31] and that the utility function need not exist. So while no such condition was necessary in the case of only two goods, in the case of three or more goods such an integrability condition would be necessary for the viability of consumer choice analysis. Consequently, for the next few decades, any economist who wanted to work with the utility-maximizing consumer choice model would need to either

[30] This is the point raised by Vito Volterra's review (Volterra 1906 [1971]) and initiated Pareto's response (Pareto 1906 [1971]).

[31] Pareto (1909 [2014], 326), but this expression was used frequently in the *Manual*.

i) impose integrability conditions directly – which was generally quite complex (particularly in higher dimensions) and never economically intuitive – or ii) employ a stronger assumption that circumvents the integrability problem, but in that case it is also necessary to provide an explanation why the additional assumption is warranted.[32] So this is why the order of consumption problem "is intertwined with" the issue of integrability; nonintegrability means that the order of consumption matters to the consumer's choices and the utility function may not exist. At the same time, we can see why integrability is not "a version of" the order of consumption problem; the order of consumption has to do with the behavior of the consumer in the real world, while integrability is a mathematical property that is only a concern for economic analysis.

Finally, it is important to note that while Pareto tied non-integrability to the problem of the order of consumption, it was by no means the only way that non-integrability could emerge. It seems clear that any context-effect, path-dependency, intransitivity, incompleteness, instability, etc. – in other words, any preference impurity – would prevent the relevant differential equation from being integrable and thus raise the possibility of the utility function not existing. The general problem is that integrability is associated with well-behaved preferences and that means that non-integrability is, whether the order of consumption is a legitimate problem or not, an extremely wide-ranging problem associated with all but the most purified of preferences.

Returning to specific concerns about welfare, it has been argued throughout this paper that ideas like limiting consumer choice to repeated/routine behavior moves in the direction of true preferences and away from context-dependent manifest preferences, and as such, are versions of preference purification. While the similarity between these two literatures does not appear to be recognized in recent research, the connection was quite clear to certain mid-twentieth century economic theorists.

A good example is Samuelson, who had a very clear vision about the connection between non-integrability and its impact on welfare economics:

> A last argument might be built up against non-integrability: if people lack the consistency of behaviour that integrability implies, then the attractive branch of

[32] Of course, the easiest way to circumvent the integrability problem is simply *to assume the consumer has a well-behaved ordinal utility function* (or well-behaved preferences that represent it). If one starts with a utility function there is no integrability problem to solve and one can proceed to set up the consumer's optimization problem, write down the first order conditions, and go forward with the economic analysis. However, that would have been epistemically inappropriate for Pareto or the economists who followed in his tradition because the empirical – "directly by experience" – part of Pareto's approach is entirely missing if one just assumes a well-behaved utility function. This was not an option for many of the first-generation ordinal utility theorists, but, as we will see in the next section, the following generation of economic theorists was more mathematically sophisticated, but less epistemically squeamish, about such matters.

> individualistic welfare economics which says people's tastes should count loses most of its content; hence, we should rule out non-integrability. (Samuelson 1950, 375)

Samuelson clearly saw the conflict between non-integrability, impure preferences, and welfare economics. Non-integrability means behavior "inconsistent with defensible assumptions about rational choice" (epigraph) which reduces individual preference satisfaction and in turn undermines welfare economics. So how did Samuelson respond to this early version of the reconciliation problem? It was basically: Okay, let's just assume integrability, forget about context-dependence, and keep the fundamental theorems of welfare economics. Which seems to be pretty much what mainstream economics did until quite recently.

Finally, in closing this section it is useful to note that our focus on preference purification has led us to discuss non-integrability solely as a problem to be eliminated so that consumer choices can again be directly linked to individual preference satisfaction and increased welfare. But there was an important literature – particularly in the 1930s, although appearing here and there later in the twentieth century – that saw demand theory without integrability as a good, rather than a bad, thing. There was a literature on non-integrable demand theory that tried to develop consumer choice theories that would still be based on utility/preference, but grounded in more psychologically realistic foundations. In particular, the literature strove to have a broadly neoclassical and ordinal conception of the consumer's goals, but to do so without assuming the existence of well-ordered preferences, or an ordinal utility function, defined over the entire consumer choice space.

The details of the various approaches to non-integrable demand theory differ fairly widely, but one important example of such theorizing was a 1932 paper by R. D. G. Allen. One feature of the paper is how clearly he connected the theory to the mathematics of physical mechanics. He made the seemingly obvious assumption that the consumer can only "make a choice between very small changes (in the limit infinitesimal changes) from any particular combination" and, unlike standard consumer choice theory, will not be able to judge "preference for widely separated combinations" (Allen 1932, 297). This is not preference purification, but it is, for want of a better expression, *local preference purification*. Within a small area of a particular consumption bundle – generally an epsilon neighborhood – the individual has well-behaved (and integrable) preferences, but outside of that neighborhood, all kinds of choice anomalies are possible. Translating this into Thaler and Sunstein's language of Econs and Humans, one might say that consumers will have Econ behavior within a neighborhood of any

particular point in the choice space, but Human behavior – or at least the possibility for Human behavior – for any other bundle outside that neighborhood.

There was an extensive literature on non-integrable demand, including Allen (1932; 1936; 1938 [1950]), Evans (1930), Georgescu-Roegen (1936; 1950; 1958; 1968), Hicks and Allen (1934),[33] and later Katzner (1970; 1971). Even Samuelson's original 1938 paper on revealed preference theory was inspired by the non-integrable demand literature; it was demand theory that satisfied consistency conditions but did not require integrability or an underlying utility function.[34]

The discussion of non-integrable demand may have pulled us somewhat off the main path of examining the (non)history of preference purification, but in the final section I will link the fate of non-integrability in the middle of the twentieth century to one of the questions posed earlier. Why is it that preference purification-like concerns were not raised before the development of behavioral economics? Of course, one of the main points of the paper has been that preference purification-like concerns *were raised*, they just were not recognized as such by the community of economic theorists at the time and faded into history. But these early versions of preference purification were not recognized in the later behavioral literature either. The conclusion will briefly suggest some of the reasons why.

6. Conclusion

From a historical perspective, the most important idea in this concluding section is that *context matters,* or at least that it may matter, to the decision-making of economic theorists just as it matters to consumer decision-making in stores or the decision-making of subjects in laboratory experiments. As noted when this was briefly introduced in section 2, this is a simple historical idea, but one that is seldom recognized in discussions about behavioral economics or BWE. In particular, I will note some of the differences between the professional context of mathematical economists working on early ordinal utility theory and the professional context of the economists (and some psychologists) working in behavioral economics and/or BWE during the last few decades. Of course, the context of these literatures undoubtedly matters in many different ways, but I will only discuss two points that seem to be the most relevant to the issues in this paper: the *context of motivation* and the *epistemic context.*

---

[33] Actually Allen always considered Hicks and Allen (1934) non-integrable demand theory whereas Hicks did not, calling it "chasing a will-o'-the-wisp" (Hicks 1946, 19, n. 1). See Chipman and Lenfant (2002), Fernandez-Grela (2006), Hands (2006) and Samuelson (1950) on Hicks' and Allen's conflicting views on this topic.

[34] Of course, Samuelson's interest in non-integrable demand theory didn't last. As the previous quote about non-integrability and welfare from Samuelson (1950) makes clear, his early sympathy for non-integrability faded rather quickly.

The *context of motivation* concerns the primary motivations of the economists working in these two approaches to individual decision-making. For those working on early ordinal utility theory, the main focus was to move forward beyond the work of the first generation of neoclassicals by fully developing and drawing out the complete implications for the ordinal utility-based theory of consumer choice and demand. Of course, the early mathematical economists that have been discussed in this paper also wanted to improve the empirical foundations of consumer choice theory, but there was never a stable consensus about how exactly that would, or could, be achieved. On the other hand, behavioral economists are also interested in individual decision-making but have quite different motivations than the early ordinal utility theorists. Behavioral economists often show little interest in competitive markets and prices, and even less in mathematical derivation from axioms, and they take an explicitly experimental approach to predicting and explaining individual behavior. Behavioral experiments often generate anomalous results and behavioral researchers concluded early on that rational choice theory was descriptively inadequate. They employed *homo economicus*, but only as a normative standard against which to identify mistakes, not as a descriptive theory of individual decision-making. The primary motivation for the development of behavioral economics was to gain a better understanding of individual decision-making by experimental means and, based on that understanding, design various interventions that would help individuals make better decisions and have higher welfare. It is clear that the early ordinal utility theorists and contemporary behavioral economists have two quite different motivational contexts.

But there was also a related, but separable, difference regarding *epistemic context*. Early ordinal utility theory, while empiricist in principle, hung all of its epistemic weight on a fairly thin thread of observability; this is true of either Pareto's various ways of characterizing indifference curves as experiential, Allen's (and Hicks and Allen's) observable marginal rates of substitution at consumption bundles, or other approaches. And yet almost no information was provided about exactly how these things could be observed, and in particular observed as stable and consistent over time, even if, once this thin empirical thread was accepted as sufficient, the implications of the theory would be derived from a few basic assumptions using calculus techniques.[35]

On the other hand, the research of behavioral economists is based on laboratory and field experiments, where data is discrete, and cleaned and processed in ways similar to how data is handled in other experimental sciences. Thus the epistemic context of contemporary behavioral economics seems to be as far removed from that of the

[35] See Hands (2017) for a more detailed discussion of this issue.

early demand theorists as was the case for the context of motivation. Although there was no opportunity for the early ordinal utility theorists to draw on the resources of behavioral economics, resources flowing the other way was certainly a possibility and yet has attracted almost no attention. There are undoubtedly many other reasons why behavioral economists did not recognize preference purification-like concerns within ordinal utility theory besides the context-dependency discussed in this paper – and the question is more poignant in recent years when neoclassical and behavioral economists seem to be moving toward a synthesis – but answers to these questions will need to wait for another time.

Author note

Earlier versions of this paper were presented at the 6th "Transforming Homo Economicus" Workshop, December 8th, 2023, at the London School of Economics and at the History of Economics Society Meeting, July 14–17, 2024 in Santiago, Chile. I would like to thank several of the participants in these two conferences as well as John Davis and a number of anonymous readers for helpful comments on earlier versions.

Competing Interests

The author has no competing interests to declare.

References

Afriat, S. N. 1967. "The Construction of Utility Functions From Expenditure Data." *International Economic Review* 8 (1): 67–77.

Allen, R. G. D. 1932. "The Foundations of a Mathematical Theory of Exchange." *Economica* 36: 197–226.

Allen, R. G. D. 1936. "Professor Slutsky's Theory of Consumer's Choice." *Review of Economic Studies* 3 (2): 120–129.

Allen, R. G. D. 1938 [1950]. *Mathematical Analysis for Economists*. London: Macmillan and Co.

Antonelli, Giovanni Battista. 1886 [1971]. "On the Mathematical Theory of Political Economy." In *Preferences, Utility, and Demand*, edited by J. S. Chipman, L. Hurwicz, M. K. Richter, and H. F. Sonnenschein, 333–363., New York: Harcourt Brace Jovanovich.

Arrow, Kenneth J. 1951. "An Extension of the Basic Theorems of Classical Welfare Economics." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, Vol. I, 365–90. Berkeley, CA: University of California Press.

Arrow, Kenneth J. and Debreu, Gerard. 1954. "Existence of an Equilibrium for a Competitive Economy." *Econometrica* 22 (3): 265–290.

Arrow, Kenneth J. and Hahn, Frank H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Ashraf, Nava, Camerer, Colin F., and Loewenstein, George. 2005. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives* 19 (3): 131–145.

Aydinonat, Emrah. 2018. "The Diversity of Models as a Means to Better Explanations in Economics." *Journal of Economic Methodology* 25 (3): 237–251.

Backhouse, Roger E. and Cherrier, Béatrice. 2017a. "The Age of the Applied Economist: The Transformation of Economics Since the 1970s." In *The Age of the Applied Economist: The Transformation of Economics Since the 1970s*, edited by R. E. Backhouse and B. Cherrier, 1–33. Durham: Duke University Press [Annual Supplement to Volume 49 of *History of Political Economy*].

Backhouse, Roger E. and Cherrier, Béatrice. 2017b. "'It's Computers, Stupid!' The Spread of Computers and the Changing Roles of Theoretical and Applied Economics." In *The Age of the*

Applied Economist: The Transformation of Economics Since the 1970s, edited by R. E. Backhouse and B. Cherrier, 103–126. Durham: Duke University Press [Annual Supplement to Volume 49 of *History of Political Economy*].

Beck, Lukas. 2023. "The Econ Within or the Econ Above? On the Plausibility of Preference Purification." *Economics & Philosophy* 39 (3): 423–445.

Bee, Michele and Desmarais-Tremblay, Maxime. 2023. "The Birth of *Homo Oeconomicus:* The Methodological Debate on the Economic Agent from J. S. Mill to V. Pareto." *Journal of the History of Economic Thought* 45 (1): 1–26.

Bergson, Abram. 1938. "A Reformulation of Certain Aspects of Welfare Economics." *Quarterly Journal of Economics* 52 (2): 310–334.

Bergson, Abram. 1954. "On the Concept of Social Welfare." *Quarterly Journal of Economics* 68 (2): 233–252.

Bergson, Abram. 1983. "Pareto on Social Welfare." *Journal of Economic Literature* 21 (1): 40–46.

Bernheim, B. Douglas. 2009. "Behavioral Welfare Economics." *Journal of the European Economic Association* 7 (2–3): 267–319.

Bernheim, B. Douglas. 2016. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Benefit Cost Analysis* 7 (1): 12–68.

Bernheim, B. Douglas. 2021. "In Defense of Behavioral Welfare Economics." *Journal of Economic Methodology* 28 (4): 385–400.

Bernheim, B. Douglas and Taubinsky, Dmitry. 2018. "Behavioral Public Economics." In *Handbook of Behavioral Economics – Foundations and Applications* 1, 1st edition, edited by B. D. Bernheim, S. DellaVigna, and D. Laibson, 381–516. Amsterdam: North-Holland.

Biddle, Jeff E. and Hamermesh, Daniel S. 2017. "Theory and Measurement: Emergence, Consolidation, and Erosion of a Consensus." In *The Age of the Applied Economist: The Transformation of Economics Since the 1970s*, edited by R. E. Backhouse and B. Cherrier, 34–57. Durham: Duke University Press [Annual Supplement to Volume 49 of *History of Political Economy*].

Binmore, Ken. 2009. *Rational Decisions*. Princeton, NJ: Princeton University Press.

Blaug, Mark. 2002. "Is There Really Progress in Economics?" In *Is There Progress in Economics? Knowledge, Truth and the History of Economic Thought*, edited by S. Boehm, C. Gehrke, H. D. Hurz, and R. Sturn, 21–41. Cheltenham, UK: Edward Elgar.

Bruni, Luigino. 2002. *Vilfredo Pareto and the Birth of Modern Microeconomics*. Cheltenham, UK: Edward Elgar.

Bruni, Luigino. 2010. "Pareto's Legacy in Modern Economics the Case of Psychology." *European Journal of Social Sciences* 48 (146): 93–111.

Bruni, Luigino and Sugden, Robert. 2007. "The Road Not Taken: How Psychology was Removed From Economics, and How It Might be Brought Back." *Economic Journal* 117 (516): 146–173.

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 (3): 1211–1254.

Camerer, Colin F. and Loewenstein, George. 2004. “Behavioral Economics: Past, Present, Future.” In *Advances in Behavioral Economics*, edited by C. F. Camerer, G. Loewenstein and M. Rabin, 3–51. New York: Princeton University Press.

Chipman, John S. 1976. “The Paretian Heritage.” *Cahiers Vilfredo Pareto*, 14 (37): 65–171.

Chipman, John and Lenfant, Jean Sebastien. 2002. “Slutsky’s 1915 Article: How it Came to be Found and Interpreted.” *History of Political Economy* 34 (3): 553–597.

Davidson, Donald and Suppes, Patrick, with Sidney Siegel. 1957. *Decision Making: An Experimental Approach*. Chicago: University of Chicago Press.

Davis, John B. 2007. “The Turn in Economics and the Turn in Economic Methodology.” *Journal of Economic Methodology* 14 (3): 275–290.

Davis, John B. 2011. *Individual and Identity In Economics*. Cambridge: Cambridge University Press.

Davis, John B. 2024. *Identity, Capabilities, and Changing Economics: Reflexive, Adaptive, Socially Embedded Individuals*. Cambridge: Cambridge University Press.

Debreu, Gerard. 1951. “The Coefficient of Resource Utilization.” *Econometrica* 19 (3): 273- 92.

Debreu, Gerard. 1954. “Representation of a Preference Ordering by a Numerical Function.” In *Decision Processes*, edited by M. Thrall, R. C. Davis and C. H. Coombs, 159–165. New York: John Wiley and Sons.

Debreu, Gerard. 1959. *Theory of Value: An Axiomatic Analysis Of Economic Equilibrium*. New Haven, CT: Yale University Press.

Dhami, Sanjit. 2016. *The Foundations of Behavioral Economic Analysis*. Oxford: Oxford University Press.

Dold, Malte F. 2018. “Back to Buchanan? Explorations of Welfare and Subjectivism in Behavioral Economics.” *Journal of Economic Methodology* 25 (2): 160–178.

Dold, Malte F. and Schubert, Christian. 2018. “Toward A Behavioral Foundation of Normative Economics.” *Review of Behavioral Economics* 5 (3–4): 221–241.

Dold, Malte and Stanton, Alexa. 2021. “I Choose for Myself, Therefore I Am: The Contours of Existentialist Behavioral Economics.” *Erasmus Journal for Philosophy and Economics* 14 (1): 1–29.

Düppe, Till and Weintraub, E. Roy. 2016. “Losing Equilibrium: On the Existence of Abraham Wald’s Fixed-Point Proof of 1935.” *History of Political Economy* 48 (4): 635–655.

Evans, Griffith. 1930. *Mathematical Introduction to Economics*. New York: McGraw Hill.

Fernandez-Grela, Manuel. 2006. “Disaggregating the Components of the Hicks-Allen Composite Commodity.” In *Agreement on Demand: Consumer Choice Theory in the 20th Century*, edited by P. Mirowski and D. W. Hands, 32–47. Durham, NC: Duke University Press [Annual Supplement to Volume 38 of *History of Political Economy*].

Georgescu-Roegen, Nicholas. 1936. “The Pure Theory of Consumer’s Behaviour.” *Quarterly Journal of Economics* 50 (4): 545–593.

Georgescu-Roegen, Nicholas. 1950. “The Theory of Choice and the Constancy of Economic Laws.” *Quarterly Journal of Economics* 64 (1): 125–138.

Georgescu-Roegen, Nicholas. 1958. “Threshold in Choice and the Theory of Demand.” *Econometrica* 26 (1): 157–168.

Georgescu-Roegen, Nicholas. 1968. “Utility.” In *International Encyclopedia of the Social Sciences*, edited by D. L. Sills, 236–267. New York: Macmillan.

Gigerenzer, Gerd. 2015. “On the Supposed Evidence for Libertarian Paternalism.” *Review of Philosophy and Psychology* 6: 361–383.

Giocoli, Nicola. 2003. *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*. Cheltenham, UK: Edward Elgar.

Grether, David M. and Plott, Charles R. 1979. “Economic Theory of Choice and the Preference Reversal Phenomenon.” *American Economic Review* 69 (4): 623–638.

Grether, David M. and Plott, Charles R. 1982. “Economic Theory of Choice and the Preference Reversal Phenomenon: Reply.” *American Economic Review* 72 (3): 575.

Grill, Kalle. 2015. “Respect for What? Choices, Actual Preferences, and True Preferences.’ *Social Theory and Practice* 41 (4): 692–715.

Grüne-Yanoff, Till. 2016. “Why Behavioural Policy Needs Mechanistic Evidence.” *Economics and Philosophy* 32 (3): 463–483.

Grüne-Yanoff, Till. 2017. “Reflections on the 2017 Nobel Memorial Prize Awarded to Richard Thaler.” *Erasmus Journal for Philosophy and Economics* 10 (2): 61–75.

Grüne-Yanoff, Till. 2022. “What Preferences for Behavioral Welfare Economics?” *Journal of Economic Methodology* 29 (2): 153–165.

Grüne-Yanoff, Till and Hertwig, R. 2016. “Nudge versus Boost: How Coherent are Policy and Theory?” *Minds and Machines* 26 (1–2): 149–183.

Gul, Faruk and Pesendorfer, Wolfgang. 2007. “Welfare without Happiness.” *American Economic Review* 97 (2): 471–476.

Hands, D. Wade. 2006. “Integrability, Rationalizability, and Path-Dependency in the History of Demand Theory.” In *Agreement on Demand: Consumer Theory in the Twentieth Century*, edited by P. Mirowski and D. W. Hands, 153–185. Durham, NC: Duke University Press [*History of Political Economy* 38, Annual Supplement].

Hands, D. Wade. 2013. “Foundations of Contemporary Revealed Preference Theory.” *Erkenntnis* 78 (5): 1081–1108.

Hands, D. Wade. 2017. “The Road to Rationalization: A History of ‘Where the Empirical Lives’ (or has lived) in Consumer Choice Theory.” *The European Journal of the History of Economic Thought* 24 (3): 555–588.

Hands, D. Wade. 2020. “Libertarian Paternalism: Taking Econs Seriously.” *International Review of Economics* 67 (4): 419–441.

Hands, D. Wade. 2023. “Frank Knight and Behavioral Economics.” *The European Journal of the History of Economic Thought* 30 (3): 341–368.

Hargreaves Heap, Shaun P. 2013. “What is the Meaning of Behavioural Economics.” *Cambridge Journal of Economics* 37 (5): 985–1000.

Harrison, Glenn W. and Ross, Don. 2023. “Behavioral Welfare Economics and the Quantitative Intentional Stance.” In *Models of Risk Preferences: Descriptive and Normative Challenges*, edited by G. Harrison and D. Ross. Bingley, UK: Emerald, Research in Experimental Economics, forthcoming.

Harsanyi, John C. 1977. “Morality and the Theory of Rational Behavior.” *Social Research* 44 (4): 623–656.

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 (1): 26–32.

Hausman, Daniel M. 2018. “Philosophy of Economics: A Retrospective Reflection.” *Review of Economic Philosophy* 18 (2): 183–201.

Hausman, Daniel M. 2022. “Enhancing Welfare Without a Theory of Welfare.” *Behavioural Public Policy* 6 (3): 342–357.

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

Hédoin, Cyril. 2015. “From Utilitarianism to Paternalism: When Behavioral Economics Meets Moral Philosophy.” *Revue de Philosophie Économique* 16 (2): 73–106.

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

Hicks, J. R. 1946. *Value and Capital*, 2nd edition. London: Oxford University Press.

Hicks, John R. and Allen, R. G. D. 1934. “A Reconstruction of the Theory of Value.” *Economica*, Part I by J. R. Hicks, 1 (1): 52–76; Part II by R. D. G. Allen, 1 (2): 196–219.

Hoover, Kevin. 2023. “Models, Truth, and Analytic Inference in Economics.” In *Methodology and History of Economics: Reflections with and without Rules*, edited by B. Caldwell, J. Davis, U. Mäki, and E-M. Sent, 119–144. London: Routledge.

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): 1–25.

Infante, Gerardo, Lecouteux, Guilhem, and Sugden, Robert. 2016b. “‘On the Econ Within’: A Reply to Daniel Hausman.” *Journal of Economic Methodology* 23 (1): 33–37.

Jeffrey, Richard C. 1965. *The Logic of Decision*. New York: McGraw-Hill.

Kahneman, Daniel. 2003. “Maps of Bounded Rationality: A Perspective on Intuitive Judgment.” *American Economic Review* 93 (5): 1449-–1475.

Kahneman, Daniel, Knetsch, Jack L., and Thaler, Richard. 1991. “Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias.” *The Journal of Economic Perspectives* 5 (1): 193–206.

Kahneman, Daniel and Tversky, Amos. 1979. “Prospect Theory: An Analysis of Decisions Under Risk.” *Econometrica* 47 (2): 263–91.

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

Katzner, Donald W. 1970. *Static Demand Theory.* London: Macmillan.

Katzner, Donald W. 1971. "Demand and Exchange Analysis in the Absence of Integrability Conditions." In *Preferences, Utility, and Demand*, edited by J. S. Chipman, L. Hurwicz, M. K. Richter, and H. F. Sonnenschein, 254–70. New York: Harcourt Brace Jovanovich.

Knetsch, Jack L. 1989. "The Endowment Effect and Evidence of Nonreversible Indifference Curves." *American Economic Review* 79 (5): 1277–1284.

Knetsch, Jack L. 1992. "Reference and Nonreversibility of Indifference Curves." *Journal of Economic Behavior and Organization* 17 (1): 131–139.

Knuuttila, Tarja and Morgan, Mary S. 2019. "De-Idealization: No Easy Reversals." *Philosophy of Science* 86 (4): 641–661.

Lecouteux, Guilhem. 2021a. "Reconciling Normative and Behavioural Economics: The Problem That Cannot Be Solved." In *The Positive and the Normative in Economic Thought*, edited by S. Badiei and A. Grivaux, 148–166. London: Routledge.

Lecouteux, Guilhem. 2021b. "Behavioral Welfare Economics and Consumer Sovereignty." In *The Routledge Handbook of the Philosophy of Economics*, edited by C. Heilmann and J. Reiss, 56–65. London: Routledge.

Lecouteux, Guilhem. 2023. "The *Homer economicus* Narrative: from Cognitive Psychology to Individual Public Policies." *Journal of Economic Methodology* 30 (2): 176–187.

Luce, R. Duncan and Raiffa, Howard. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: John Wiley and Sons.

Mäki, Uskali. 1994. "Isolation, Idealization and Truth in Economics." In *Idealization VI: Idealization in Economics*, edited by B. Hamminga and N. De Marchi, 147–68. Amsterdam: Rodopi.

Mandler, Michael. 1999. *Dilemmas in Economic Theory: Persisting Foundational Problems of Microeconomics*. New York: Oxford University Press.

Marchionni, Caterina. 2017. "The Problem with Model-Based Explanation in Economics." *Disputatio* 9 (47): 603–630.

Mas-Colell, Andreu, Whinston, Michael D., and Green, Jerry R. 1995. *Microeconomic Theory*. New York: Oxford University Press.

McLure, Michael. 2005. "A Note on Pareto's 'Sunto'." *Journal of the History of Economic Thought* 27 (4): 399–403.

McQuillin and Sugden, Robert. 2012. "Reconciling Normative and Behavioral Economics: the Problems to be Solved." *Social Choice and Welfare* 38 (4): 553–567.

Montesano, Aldo. 2006. "The Paretian Theory of Ophelimity in Closed and Open Cycles." *History of Economic Ideas* 14 (3): 77–100.

Moscati, Ivan. 2019. *Measuring Utility*. New York: Oxford University Press.

Moscati, Ivan. 2023. *The History and Methodology of Expected Utility*. Cambridge: Cambridge University Press.

Moscati, Ivan. 2024. "Behavioural and Heuristic Models Are As-If Models Too – and That's OK." *Economics & Philosophy* 40 (2): 279–309.

Munro, Alistair and Sugden, Robert. 2003. "On the Theory of Reference-Dependence Preferences." *Journal of Economic Behavior and Organization* 50 (4): 407–428.

Nussbaum, Martha. 2011. *Creating Capabilities*. Cambridge, MA: Harvard University Press.

Nussbaum, Martha and Sen, Amartya, eds. 1993. *The Quality of Life*. Oxford: Clarendon Press.

Pareto, Vilfredo. 1892–93 [2007]. *Considerations on the Fundamental Principles of Pure Political Economy*, edited by R. Marchionatti and F. Mornati. London: Routledge. [originally published in Italian as "Considerazioni sui principi fondamentali dell'economia pura," published in *Giornale degli Economisti*, in five parts between May 1892 and October 1893].

Pareto, Vilfredo. 1906 [1971]. "Ophelimity in Nonclosed Cycles." translated by A. P. Kirman and annotated by J. S. Chipman. In *Preferences, Utility, and Demand*, edited by J. S. Chipman, L. Hurwicz, M. Richter, and H. F. Sonnenschein, 370–385. New York: Harcourt Brace Jovanovich.

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

Reiss, Julian. 2012. "The Explanatory Paradox." *Journal of Economic Methodology* 19 (1): 43–62.

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

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

Samuelson, Paul A. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5 (17): 61–71.

Samuelson, Paul A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, Paul A. 1950. "The Problem of Integrability in Utility Theory." *Economica* 17 (68): 355–385.

Samuelson, Paul A. 1981. "Bergsonian Welfare Economics." In *Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson*, edited by S. Rosefielde, 223–266. Cambridge: Cambridge University Press.

Samuelson, Paul A. 2005. "An Interview with Paul Samuelson on the New and Old Welfare Economics," interview by Kotaro Suzumura. *Social Choice and Welfare* 25 (2): 327–356.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: John Wiley and Sons.

Schultz, Henry. 1938. *The Theory and Measurement of Demand.* Chicago: University of Chicago Press.

Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica* 40 (159): 241–259.

Sen, Amartya. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6 (4): 317–344.

Sen, Amartya. 1979. "Equality of What?" In *Tanner Lectures on Human Values*, edited by S. M. McMurrin, 197–220. Cambridge: Cambridge University Press.

Sen, Amartya. 1999. *Development as Freedom*. New York: Anchor.

Sen, Amartya. 2002. *Rationality and Freedom*. Cambridge, MA: Belknap Press of Harvard University.

Sent, Esther-Mirjam. 2004. "Behavioral Economics: How Psychology Made Its (Limited) Way Back Into Economics." *History of Political Economy* 36 (4): 735–60.

Slutsky, Eugen E. 1915 [1952]. "On the Theory of the Budget of the Consumer," translated by Olga Ragusa. In *Readings in Price Theory*, edited by G. J. Stigler and K. E. Boulding, 27–56. Chicago: Irwin [originally published in *Giornale degli Economisti* 51: 1–26].

Stigler, George A. 1950. "The Development of Utility Theory II." *Journal of Political Economy* 58 (5): 373–396.

Sugden, Robert. 2008. "The Changing Relationship Between Theory and Experiment in Economics." *Philosophy of Science* 75 (5): 621–632.

Sugden, Robert. 2009. "Credible Worlds, Capacities and Mechanisms." *Erkenntnis* 70 (1): 3–27.

Sugden, Robert. 2010. "Opportunity as Mutual Advantage." *Economics & Philosophy* 26 (1): 47–68.

Sugden, Robert. 2015. "Looking for a Psychology for the Inner Rational Agent." *Social Theory and Practice* 41 (4): 579–598.

Sugden, Robert. 2019, *The Community of Advantage: A Behavioural Economist' Defense of the Market*. Oxford: Oxford University Press.

Sugden, Robert. 2021. "Hume's Experimental Psychology and the Idea of Erroneous Preferences." *Journal of Economic Behavior and Organization* 183: 836–848.

Sunstein, Cass R. and Thaler, Richard H. 2003. "Libertarian Paternalism Is Not an Oxymoron." *The University of Chicago Law Review* 70 (4): 1159–1202.

Suppes, Patrick. 1961. "The Philosophical Relevance of Decision Theory." *The Journal of Philosophy* 58 (21): 605–614.

Tarascio, Vincent J. 1969. "Paretian Welfare Theory: Some Neglected Aspects." *Journal of Political Economy* 77 (1): 1–20.

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

Thaler, Richard H. 2017. "Behavioral Economics." *Journal of Political Economy* 125 (6): 1799–1805.

Thaler, Richard H. and Sunstein, Cass R. 2003. "Behavioral Economics, Public Policy, and Paternalism." *The American Economic Review* 93 (2): 175–179.

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

Thoma, Johanna. 2021. "On the Possibility of an Anti-Paternalist Behavioural Welfare Economics." *Journal of Economic Methodology* 28 (4): 350–363.

Tversky, Amos and Kahneman, Daniel. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics* 106 (4): 1039–61.

Varian, Hal R. 2014. *Intermediate Microeconomics*, 9th edition. New York: W. W. Norton.

Volterra, Vito. 1906 [1971]. "Mathematical Economics and Professor Pareto's New Manuel," translated by A. P. Kirman and edited by J. S. Chipman. In *Preferences, Utility, and Demand*, edited by J. S. Chipman, L. Hurwicz, M. Richter, and H. F. Sonnenschein, 365–369. New York: Harcourt Brace Jovanovich.

Weber, Christian. 2001. "Pareto and the 53 Percent Ordinal Theory of Utility." *History of Political Economy* 33 (3): 541–76.

Weintraub, E. Roy. 1983. "On the Existence of a Competitive Equilibrium: 1930–1954." *Journal of Economic Literature* 21 (1): 1–39.

Whitman, Douglas Glenn and Rizzo, Mario J. 2015. "The Problematic Welfare Standards of Behavioral Paternalism." *Review of Philosophy and Psychology 6* (3): 409–425.

Wold, H. O. A., Hotelling, Harold, Koopmans, Tjalling C., and Roy, Rene. 1949. "*De la theorie des choix aux budgets de familles*: Discussion." *Econometrica*, 17 (Supplement) : 187–191.

Ylikoski, Petri and Aydinonat, Emrah. 2014. "Understanding with Theoretical Models." *Journal of Economic Methodology* 21 (1): 19–36.