**School of Economics**
**University of East Anglia**
**Norwich NR4 7TJ, United Kingdom**

University of East Anglia

# Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics

**Robert Sugden**
**University of East Anglia**

Talk for Economic Science Association International Meeting
Jerusalem, 7-11 July 2016

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 670103.

Gerardo Infante

Guilhem Lecouteux

Based on paper in *Journal of Economic Methodology* 2016, co-authored with Gerardo Infante (University of East Anglia) and Guilhem Lecouteux (École Polytechnique)

This talk is about a problem that has arisen in economics in the last 25 years, as a result of the inflow of ideas and research methods from psychology: **how to reconcile normative and behavioural economics.**

I review one of the main approaches that economists have followed in tackling this problem – **behavioural welfare economics**. The essential idea is that individuals have integrated **latent preferences**, but their psychology sometimes causes errors in decision-making. The task for normative economics is to recover individuals' latent preferences (by **purifying** revealed preferences) and to use the satisfaction of these preferences as the normative criterion.

I present a critique of this approach.

## The problem of reconciling normative and behavioural economics

From the early 20th century, the dominant form of economics has been *neoclassical*. A fundamental assumption of neoclassical theory is that individuals have *integrated preferences* over all relevant economic outcomes, and act according to those preferences (= 'maximise utility').

Integrated preferences are:

– stable (i.e. not subject to random or arbitrary variation);

– context-independent (i.e. not affected by arbitrary changes of 'framing' of decision problems);

– internally consistent (i.e. satisfying 'rationality' principles such as transitivity and menu independence).

Neoclassical welfare economics uses the satisfaction of preferences as its normative criterion.

However, a lot of evidence suggests that individuals *don't* act on integrated preferences. In particular: preferences revealed in choices are *context-dependent*. A typical example:


**Apples and Mars bars** (experiment by Read and van Leeuwen, 1998):

Office workers are approached either just after lunch, or in late afternoon; they are offered a choice of snacks that will be delivered at a fixed time (just after lunch, or in late afternoon) *a week later*; they have to choose now which snack to have. Some snacks are 'healthy' (e.g. apple), some are 'unhealthy' (e.g. Mars bar).

Result: irrespective of the *delivery time*, individuals are more likely to choose unhealthy snacks if the *choice* is made in late afternoon.


Notice:

– The difference in choice has a psychological cause (people are hungrier in late afternoon; thoughts about hunger-satisfying properties of food are more salient to people who are hungry).

– The psychological difference doesn't seem to be a *good reason* for the difference in choice (the options are familiar and daily fluctuations in hunger are very predictable);

– but there is no obvious way of determining whether it is more rational to choose an apple or a Mars bar (at either time of day).

So, a problem for normative economics:

If (revealed) preferences vary according to factors that are not relevant to individuals' interests or welfare, how can we justify using preference-satisfaction as a normative criterion?

And if we can't, what normative criterion *should* we use?

## Behavioural welfare economics

I'll focus on Sunstein and Thaler, as the most prominent advocates of behavioural welfare economics …

Sunstein and Thaler (2003). 'Libertarian paternalism is not an oxymoron', *University of Chicago Law Review*.

Thaler and Sunstein (2008). *Nudge*.

But similar 'preference purification' approaches have been advocated by many economists, e.g.

Bleichrodt, Pinto-Prades and Wakker (2001);
Camerer, Issacharoff, Loewenstein, O'Donaghue and Rabin (2003); Kőszegi and Rabin (2007);
Salant and Rubinstein (2008); Bershears, Choi, Laibson and Madrian (2008);
Bernheim and Rangel (2009);

… and this approach has been endorsed by a leading philsosopher of economics:

Hausman , *Preference, Value, Choice, and Welfare* (2012)

… and by a leading specialist in the economics of social policy:

Le Grand and New, *Government Paternalism* (2015)

… and in private communication, by Kahneman.

Richard Thaler

Cass Sunstein

Sunstein and Thaler claim that the findings of behavioural economics make the criterion of (revealed) preference-satisfaction 'incoherent', a 'non-starter'. They support this claim by using the *cafeteria* example…

Premise: customers' choices between food items depend on the positions in which they are displayed. Knowing this, how should the cafeteria director choose the positions?

'[the customers] lack well-formed preferences, in the sense of preferences that are firmly held and preexist the director's own choices about how to order the relevant items [along the counter]. If the arrangement of the alternatives has a significant effect on the selections of the customers make, then their true 'preferences' do not formally exist.'

So what criterion should the director use?

S&T say she should:

'make choosers better off, *as judged by themselves*' (2008).

The '*as judged by themselves*' clause is important for S&T. E.g. Thaler in *Mishbehaving* (2015):

'.. .a point that critics of our book [i.e. *Nudge*] seem incapable of getting. [We] have no interest in telling people what to do. We want to help them achieve their *own* goals'.

Pointing to the '*as judged by themselves*' clause in *Nudge:*

'The italics are in the original but perhaps we should also have used bold and a large font, given the number of times we have been accused of thinking that we know what is best for everyone. … We just want to reduce what people would themselves call errors.'

Implication: the director/ planner/ choice architect should respect individuals' *subjective judgements* about their own welfare (NB: S&T are not invoking an objective concept of welfare)….

… but should *not* presuppose that these judgements are revealed in individuals' choices.

Problem: When choices are context-dependent, how are we to understand these judgements? And how is the planner to reconstruct them?

The nearest S&T get to answering these questions is in discussing *decision-making errors*…

Immediately after the remark about making choosers better off, as judged by themselves, S&T say they will show that:

'in many cases, individuals make pretty bad decisions – decisions that they would not have made if they had **paid full attention** and **possessed complete information, unlimited cognitive abilities, and complete self-control'**.

Such decisions are 'inferior decisions in terms of their [i.e. the individuals'] own welfare'.

Implication: S&T's welfare criterion is the satisfaction of the **latent preferences** that an individual would have revealed *in the absence of reasoning imperfections*. Preference purification = reconstructing what the individual would have chosen in the absence of reasoning imperfections.

Implicit assumption: **latent preferences are context-independent** (since S&T are claiming to solve a problem created by the context-dependence of revealed preferences). But S&T never justify this assumption.

S&T's supporting rhetorical strategy is to characterise conventional economists as assuming that humans are 'Econs' who are immune to reasoning imperfections and can:

'think like Albert Einstein, store as much memory as IBM's Big [Deep?]Blue, and exercise the willpower of Mahatma Gandhi'.

S&T claim (uncontroversially) that 'the folks we know are not like that', with the suggestion that:

– this is *why* real human choices are context-dependent (i.e. context-dependent choices are the result of reasoning imperfections);

– their own approach represents human psychology as it really is.

I will challenge both suggestions.

## The model of the inner rational agent

S&T's implicit model of human agency: a faulty Econ.

**Inside the human being, there is a neoclassically rational agent, with coherent preferences**. These are the latent preferences that can be reconstructed by 'purification'.

But **the rational agent is trapped in a psychological shell**. Its interactions with the world are mediated by the shell.

Properties of the psychological shell cause **errors** in decisions:

– lack of memory capacity, so the inner agent does not have access to all the information it requires;

– lack of attention, so relevant information does not always reach the inner agent;

– lack of cognitive ability, so computations that the inner agent requires are not always accurate;

– lack of self-control, so the inner agent's decisions are not always executed.

Compare the Martians in *The War of the Worlds*

(the Rational Slug?)

**What is wrong with this model?**

*It doesn't take psychology seriously*.

The starting point for behavioural economics was recognising that the mental processes actually used in decision-making do not necessarily generate choices with the rationality properties assumed in economics.

An obvious corollary, noted by Kahneman (1996): **rational choice is not self-explanatory**, i.e. behaviour that is consistent with the standard economic theory is just as much in need of psychological explanation as are 'anomalies'.

But in the model of the inner rational agent, human psychology is represented as a set of forces which affect behaviour by *interfering with* rational choice. Rational choice itself is not given any psychological explanation.

There is no psychological explanation of why latent preferences exist at all. If *actual* choices are determined by context-dependent cues (e.g. the Mars bar case), what is the function of latent preferences?

Perhaps the idea is that our psychology endows us with *reasoning abilities* that are generally useful (a possible interpretation of Kahneman's arguments about 'System 2').

I.e. the implicit assumption is that there is some mode of *latent reasoning*, accessible to an agent who is not subject to reasoning imperfections, which generates context-independent preferences. **But we are never told what this mode of reasoning is.**

**Why we should be sceptical: the case of SuperReasoner**.

Consider an ordinary human being, Joe, in Sunstein and Thaler's cafeteria. He chooses whichever of fresh fruit or cream cake is displayed nearer the front of the counter. What is his latent preference?

S&T's test: imagine SuperReasoner, who is just the same as Joe except that he:

- has access to all relevant information;
- gives full attention to all relevant information;
- has no cognitive limitations;
- has perfect self-control.

SuperReasoner's choice reveals Joe's latent preferences.

The crucial question: **will SuperReasoner's choices between fruit and cake be context-independent?**

I claim the answer is: quite possibly not...

It's fundamental to the preference purification approach that latent preferences are *subjective.*

So we can't assume that the question 'In Joe's judgement, which is better for him, fruit or cake?' has an *objective* answer.

But then, we're not entitled to assume that the question has a *determinate* answer, accessible by applying unlimited cognitive ability to full information.

So suppose that SuperReasoner finds that valid reasoning, based on full information, can't determine which of fruit and cake is better for him. SuperReasoner has Joe's *feelings*, and so feels an inclination to choose whichever of fruit and cake is more prominently displayed. **Why shouldn't he act on that inclination?**

I can't see any good reason why he shouldn't.

But a possible contrary line of argument …

In my account, SuperReasoner's preferences/ judgements are *incomplete* – he doesn't maintain *either* 'All things considered, fruit is at least as good as cake' *or* 'All things considered, cake is at least as good as fruit'.

Could one make it axiomatic that rationality *requires* preferences to be complete? (Possible thought: if preferences are incomplete, some choices can't be made on the basis of rationality. So to be 'truly' rational, one must have complete preferences.)

I don't find this way of thinking about rationality persuasive, but in any case, it doesn't help…

All it would tell us is that SuperReasoner would construct *some* preference between fruit and cake, even though (in the case we are assuming) that preference would not be derived from valid reasoning.

But we still may not know what that preference would be (it's an arbitrary choice by an imaginary agent!); and it would be hard to make sense of the claim that the preference is latent *in Joe*.

**If the inner rational agent model is so hard to defend, why have so many economists used it?**

I suggest the use of this model is a by-product of a modelling strategy in behavioural economics that is defensible *in descriptive theorising* – the strategy of **behavioural optimisation**.

This strategy uses conventional rational-choice theory as a template, and models the individual as maximising a *behavioural utility* function. Many properties of conventional utility functions are retained. Additional variables are introduced to represent psychological influences on choice (e.g. regret, ambiguity aversion, probability weighting, loss aversion, attention biases…).

Often, conventional utility is a special case of behavioural utility.

Merits of this strategy (arguments used by Hausman (2012) and Rabin (2013)):

– It looks for incremental improvements to existing theories, rather than starting from scratch (a merit if existing theory is a reasonable first approximation to the truth).

– It allows re-use of a large body of abstract theoretical results about optimising behaviour.

– It makes it easier to identify and test novel features of behavioural theories.

Notice that these arguments do not depend on the *rationality* properties of conventional theory – only on the theory being *widely accepted* and *reasonably successful, descriptively*.

Perhaps conventional theory is successful because it takes account of some of the most important factors that influence choice, in a simple and tractable framework – *not* because it is a good representation of reasoning processes.

But because the conventional theory is framed in terms of rationality, it is tempting to interpret a behavioural optimisation model as representing the combined effects of *rational factors* (the special case of the model that corresponds with conventional theory) and *errors* (the added psychological factors).

In the absence of a theory of how people reason, this interpretation is unjustified. It may be fairly harmless in descriptive economic theorising – but that is because *the distinction between 'rationality' and 'error' has no explanatory role*.

But behavioural welfare economics requires a defensible and operational method of making that distinction – but hasn't provided one!

**Where do we go from here?**

We need some way of reconciling behavioural and normative economics. If not by preference purification, how?

My preferred approach: find a normative criterion (e.g. my 'opportunity criterion', which I've argued for in a series of papers) which respects individual choices without referring to the preferences that may (or may not) lie behind them.

Other approaches are possible…

But the first essential is that economists learn to live with the facts of human psychology. We need a normative economics that does not presuppose a kind of rational human agency for which there is no known psychological foundation.

Thank you for listening.