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 at Department of Economics, University of Copenhagen, 27 April 2016

Based on joint work with Gerardo Infante (University of East Anglia) and Guilhem Lecouteux (École Polytechnique)

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.

This paper 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.**

Our paper (*Journal of Economic Methodology* 2016) reviews one of the main approaches that economists have followed in tackling this problem – **behavioural welfare economics** – and presents a critique of this approach.

This talk summarises that critique and adds a discussion of a specific model which featured in Edward Webb's dissertation: Bordalo, Gannaioli and Shleifer's model of 'salience and consumer choice'.

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).

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.

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.

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

… and this approach has been endorsed by a leading philsosopher of economics:
Hausman (2012). *Preference, Value, Choice, and Welfare*.

S&T argue for libertarian paternalism.

(Note: not all advocates of preference purification favour paternalistic policies. Our critique is of S&T's method of assessing individual welfare, not of how they think governments should use these assessments.)

S&T claim that the findings of behavioural economics make paternalism unavoidable: the anti-paternalist position is '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? She can't use (revealed) preference-satisfaction as her criterion, because:

'[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].'

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, 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.)

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.

We 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.

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.

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 are 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?**

We 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 ‘Which is better for Joe, 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.

(And since the advocates of preference purification haven’t identified any mode of reasoning which produces a determinate answer, we’re entitled to be sceptical.)

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?**

Implication: no reason to presuppose that purified preferences are context-independent.

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

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

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.

As a descriptive modelling strategy, this has pragmatic merits (as argued by Hausman (2012) and Rabin (2013)) -- it looks for incremental improvements to existing theories, and allows sharp tests of how far behavioural theories improve on neoclassical ones. But …

… it can lead us astray in normative analysis.

Formally, this modelling strategy allows a separation between ‘neoclassical’ and ‘behavioural’ arguments in the utility function.

It is easy to interpret the neoclassical special case as the individual’s latent preferences and deviations from this as errors.

But this interpretation is not justified.

An example: Bordalo, Gannaioli and Shleifer, 'Salience and consumer choice', *Journal of Political Economy*, 2013

The BGS model is motivated by experimental findings from psychology, economics and marketing. Core idea is from psychology:

'salience refers to the phenomenon that when one's attention is differentially directed to one portion of the environment rather than to others, the information contained in that portion will receive disproportionate weighting in subsequent judgements' (Taylor and Thompson, 1982).

Basic model: consumer chooses between two goods, each good k described by two attributes: quality (q_k) and price (p_k).

BGS's example: good 1 = French wine, good 2 = Australian wine; $q_1 > q_2$, $p_1 > p_2$.

Situation 1 ('supermarket'): French wine \$20, Australian wine \$10.

Situation 2 ('restaurant'): French wine \$50, Australian wine \$40.

Qualities are the same in both cases.

(Supposed) observation: consumer chooses Australian in situation 1, French in situation 2, contrary to consumer theory (assuming fixed preferences and non-perverse income effects).

BGS's explanation: the \$10 price difference is more salient in situation 1, i.e. it is larger in relation to the average price in the choice set (\$15 in situation 1, \$45 in situation 2). Salience hypothesis: the individual gives more attention to more salient attributes, and so acts as if more salient attributes had greater weight in the utility function.

My interest here is not in whether this effect occurs in reality, but in how BGS model it.

BGS's model:

The consumer has a 'rational' utility function

$u_k = \alpha_Q q_k - \alpha_P p_k$

i.e. α_Q and α_P are attribute weights in the utility of a rational consumer. BGS normalise these to 1 (i.e. the unit of quality is defined so that it has the same rational utility value as a unit of money).

These weights are 'distorted' by salience. In any given choice set, the weight of the more [less] salient attribute is greater than [less than] 1. The amount of distortion is determined by a person-specific parameter, with a limiting case representing no distortion.

This is an 'inner rational agent' model. The consumer has true ('rational') preferences, but makes mistakes because of a psychological mechanism (biased attention).

Applied to the wine example, this model implies that if the prices of both wines increase by the same amount:

1. The rational consumer’s preference is unaffected;
2. Depending on the parameters, a non-rational consumer may switch preference, but the switch is always from the lower-price wine to the higher-price wine.

So the model offers an explanation for the effect. But…

… ‘rational’ preferences play no part in this (or in BGS’s explanations of other effects). Empirically, all we need is that attribute weights are positively related to salience. BGS have defined a particular set of weights as ‘rational’, but this is extraneous to the empirical model.

How do BGS define ‘rational’ weights?

The structure of the BGS model implies that ‘rational’ preferences are revealed in choices between one good (q, p) and (0, 0), i.e. ‘willingness to pay’ tasks.

Rationale: BGS say that rational preferences can be elicited in settings in which ‘a good is evaluated in isolation and without price expectations’; such settings can be created in ‘lab experiments’.

The logic seems to be:

(1) A rational consumer must have a true utility valuation of each good which is independent of which other goods are in the choice set.

(2) If we find a pattern of choice contrary to this principle (as in the wine example), this must be because the utility valuation of some good is being distorted by comparisons with other goods.

(3) So the way to recover true utility valuations is to construct choice sets in which no such comparisons are possible.

But how is (1) justified? This is the inner rational agent model!

The idea that context-independent utility valuations can be elicited in willingness-to-pay tasks in lab experiments is contrary to the findings of decades of research.

Responses to WTP questions are known to be influenced by many 'irrelevant' factors, e.g.:

-- Response mode effects (e.g. choice or valuation -- preference reversal)

-- Anchoring effects (valuations influenced by irrelevant but salient cues)

-- Range/frequency effects (valuations influenced by the scale on which valuations are reported).

These effects are particularly strong when the good being valued has no known price to act as a reference point (e.g. valuations of health, safety, environmental public goods, unpleasant tastes and noises).

Most obvious inference: use of comparators to arrive at valuations is <u>how human psychology works</u>.   The idea that we can elicit true valuations by removing all comparators is a mistake.

Where do we go from here?

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

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.

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.

Thank you for listening.