Economics & Philosophy (2023), **39**, 423–445
doi:10.1017/S0266267122000141

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

The Econ within or the Econ above? On the plausibility of preference purification

Lukas Beck

University of Cambridge, Cambridge, UK and Mercator Research Institute on Global Commons and Climate Change (MCC), Berlin, Germany
Email: beck@mcc-berlin.net; lb760@cam.ac.uk

(Received 18 April 2021; revised 1 April 2022; accepted 16 April 2022; first published online 05 September 2022)

Abstract
Scholars disagree about the plausibility of preference purification. Some see it as a familiar phenomenon. Others denounce it as conceptually incoherent, postulating that it relies on the psychologically implausible assumption of an inner rational agent. I argue that different notions of rationality can be leveraged to advance the debate: procedural rationality and structural rationality. I explicate how structural rationality, in contrast to procedural rationality, allows us to offer an account of the guiding idea behind preference purification that avoids inner rational agents. Afterward, I address two pressing challenges against preference purification that emerge under the structural rationality account.

Keywords: Preference purification; rationality; well-being; behavioural welfare economics

1. Introduction

Preference satisfaction continues to be the dominant welfare criterion in economics. Just consider the fields of cost-benefit analysis (e.g. Adler and Posner 2006) or the nudging movement in evidence-based policy (e.g. Thaler and Sunstein 2008). However, over the last couple of decades, behavioural economics has put a lot of pressure on the idea that people are rational. As a result, behavioural *welfare* economists have alluded to the idea of preference purification. Very roughly, the idea behind preference purification is that people's actual preferences are in some sense mistaken and that a set of 'laundered' or 'purer' preferences is needed to track people's welfare. This idea holds a lot of sway among behavioural economists. However, apart from the fact that purified preferences should be somehow viewed as a corrected version of people's actual preferences, there is no explicit consensus on what purified preferences are supposed to be. Yet, as will become apparent, clarifying the precise nature of purified preferences is crucial for assessing the plausibility of preference purification.

So far, the most elaborated attempt in this respect was offered by Infante, Lecouteux and Sugden (Infante *et al.* 2016*a*, henceforth ILS). They investigate

the implicit assumptions of preference purification. This investigation leads them to the verdict that preference purification rests on the psychologically implausible assumption of an inner rational agent (IRA), which is meant to undermine preference purification severely.

However, their critique has not persuaded the proponents of preference purification. Instead, those usually reply that preference purification is not committed to IRAs (e.g. Hausman 2016). Yet, as of now, nobody has explicated an alternative account that matches the comprehensiveness of ILS's account. As a result, it remains unclear how exactly preference purification is meant to avoid IRAs.

This paper offers the following two contributions to the ongoing debate. The first aim is to lift the present deadlock by explicating a comprehensive alternative to ILS's picture. After introducing the current status quo (sections 2 and 3), I leverage two different notions of rationality to show that there is conceptual space for an alternative understanding of preference purification. In particular, I argue that ILS's account aligns with a procedural notion of rationality (cf. Kacelnik 2006), while certain defenders of preference purification such as Hausman (2012, 2016) seem to closely align with a structural notion of rationality (cf. Broome 2013) (section 4). I then explicate how structural rationality allows us to offer an account of purified preferences that does not appeal to IRAs (section 5). The basic idea is to identify purified preferences with the preferences an agent would have if she were to satisfy the requirements of structural rationality, and not with the outcome of a particular psychological process (i.e. IRAs). I shall call this the SR-account of preference purification.[1]

The second aim of the paper is to begin with the appraisal of the SR-account (section 6). The account by itself cannot resolve the debate because it faces several challenges that preference purification would avoid under the IRA picture. Nevertheless, these challenges are less decisive than the charge of psychological implausibility that would follow if preference purification were committed to IRAs. To motivate this, I outline how we can make progress towards addressing several pressing challenges against the SR-account via combining it with the so-called evidential account of the relationship between preferences and welfare (Hausman and McPherson 2009).[2] All in all, I thereby hope to shift the debate into a more fruitful direction and to ultimately arrive at clearer conceptual foundations for preference purification.

2. Welfare economics and preference satisfaction

Economists defend preference satisfaction as a welfare criterion on various grounds, e.g. for utilitarian, or libertarian reasons, or based on its alleged neutrality (see

[1]Dietrich *et al.* (2021) have recently noted that a Broomean account of reasoning can help us make sense of two different types of error about preferences in general and Savages' response to the Allais's paradox in particular. In this context, they also ask whether Savage's reasoning could help us make sense of purified preferences. Yet, they dismiss this line of thought on the ground that it is allegedly unable to capture the notion of mistake used by behavioural welfare economists. Thus, one can see the first contribution of my paper as picking up this Broomean thread and asking whether it allows us to reconstruct the guiding idea behind preference purification rather than specific uses of the term mistake.

[2]Hausman (2016) already notes that preference purification could be seen as a method for extending the scope of the evidential account.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

Camerer *et al.* 2003; Thaler and Sunstein 2008; Hausman and Welch 2010; Glaeser 2011). Yet, no matter the precise justification for using preference satisfaction, traditional welfare economics holds that agents' preferences need to exhibit certain properties (i.e. completeness, transitivity, stability and context-independence) in order to be relevant for welfare decisions (see Bernheim and Rangel 2008; Grüne-Yanoff 2018; Sugden 2018). For the purpose of this paper, I will call preferences that satisfy all the relevant properties rational preferences.[3]

However, over the last couple of decades, behavioural economics has put pressure on the idea that our preferences usually satisfy these properties (see, for instance, Ariely 2010). The literature is full of examples that aim at demonstrating people's irrationality. For a canonical example of the violation of context-independence, consider Read and Van Leeuwen (1998), who demonstrate that the choice of workers between a chocolate bar and an apple (both received at noon a week from the day on which the question was posed) depends on the supposedly irrelevant fact of whether they were given the choice in the early or late afternoon. Classical examples of robust violations of transitivity can be found in Tversky (1969). For the case of decision-making under uncertainty, the famous Allais paradox is probably the most well-known example (see Savage 1954). It demonstrates a violation of the independence-of-irrelevant-alternatives assumption (for further examples, see Kahneman *et al.* 1990; Park *et al.* 2000; Johnson and Goldstein 2003).

After generating such results for several decades, behavioural economists started to advocate that economics needs a heavy injection of psychological theory and method in order to be able to account for people's irrationality (e.g. Rabin 2002). Importantly for the present purpose, these experimental results also undermine the appeal of preference satisfaction as a welfare criterion. After all, so the thought goes, if people's preferences are irrational, why should we pay attention to them? Consequently, one would expect behavioural economists to abandon preference satisfaction as a welfare criterion. Yet, quite surprisingly, many behavioural welfare economists try to retain it (Bernheim and Rangel 2008; Thaler and Sunstein 2008; see also Sugden 2018).

In this regard, it is argued that even though people's *actual* preferences should be viewed as mistaken (which is indicated by a failure to satisfy the properties mentioned above), we can construct a corrected set of *alternative* preferences out of people's *actual* preferences that can then be used to inform welfare judgements. The method in virtue of which we are supposed to arrive at the alternative set of preferences is sometimes called *preference purification*, and the alternative preferences are called *purified preferences* accordingly.

However, apart from the fact that they supposedly satisfy the above-mentioned properties, it is not clear at all what purified preferences are supposed to be. In the absence of an explicit consensus, various theoretical frameworks for operationalizing

[3]The term rational in economics has a technical meaning referring to preferences that are transitive and complete. While Sugden (2018) has coined the term *integrated preferences* for preferences that satisfy all the properties that make them relevant for welfare judgements, many others use the term *rational preferences* quite loosely. I will also not use the term in its narrow, technical sense. In the present context, this would threaten to obscure that there is a relation between different accounts of preference purification and more substantive accounts of rationality.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

preferences purification have been proposed. For example, Bernheim and Rangel (2008, 2009) propose a theoretical framework that relies on a so-called generalized choice situation. Generalized choice situations include objects from which the individual must choose and ancillary conditions, which influence behaviour but do not alter the option in any way. Bernheim and Rangel then propose to only rely on those preferences of individuals that do not depend on the ancillary conditions. Moreover, Salant and Rubinstein (2008) put forward a model that assumes that agents have a set of underlying preferences and aim at representing specific psychological factors that distort these preferences. Finally, Bleichrodt *et al.* (2001) develop an account for preference purification that is mainly aimed at (medical) professionals who are supposed to act in the interest of their clients. The first step of this method is to estimate a cumulative prospect theory model – which the authors take to be the descriptively adequate model of decision-making – that can capture the hypothetical choices made by the relevant agents. The second step is to derive an expected utility model – which the authors take to be the normatively adequate model – from the cumulative prospect theory model. This expected utility model is then supposed to capture the clients' purified preferences.

As this quick overview shows, there are several attempts at deriving purified preferences. However, given the absence of an account of what they are supposed to be, it is quite difficult to assess the plausibility and success of any of these attempts because we lack a clear standard against which they can be compared.

Before I turn to ILS's account, that can be understood as addressing this lacuna, note that they themselves see their account as an explication of the implicit assumptions of proponents of preference purification.[4] However, I take it to be an open question whether proponents of preference purification really hold these implicit assumptions or encode them in their practices (see, for instance, Bernheim 2016). In any case, even if ILS's account failed in its aims to capture the implicit assumptions of proponents of preference purification, we would still need an alternative story of what we are aiming at when attempting to purify preferences. Moreover, we can also question whether any of the current proposals for preference purification would be successful in uncovering purified preferences even if the (highly implausible) assumptions ILS attribute to proponents of preference purification were correct. This indicates that their account of preference purification is located at a relatively high level of idealization rather than a straightforward reconstruction of certain practices in behavioural welfare economics.

In light of these considerations, I will treat ILS's account *not primarily* as an attempt at reconstructing existing proposals for preference purification. Instead, to avoid getting bogged down by a debate about how much actual practice their account successfully captures or intends to capture, I will engage it at the more abstract level of reconstructing the guiding idea behind preference purification. In this regard, both accounts of preference purification discussed here can be viewed as attempts to uncover the implicit ontological and

[4]I thank Bob Sugden for highlighting the importance of emphasizing this.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

normative assumptions that we need to make in order to make sense of preference purification.[5]

3. Preference purification and the inner rational agent

I will now outline ILS's argument and explain how it is meant to undermine preference purification (3.1). I will then turn to Hausman's (2016) reply which exemplifies a typical response in that it simply denies an implicit commitment to IRAs. This will allow me to highlight that without an explicit alternative account of the guiding idea behind purified preferences, the debate is threatened to remain deadlocked (3.2).

3.1. ILS's critique

ILS (2016*a*) argue that we need the assumption of an IRA that is trapped in a psychological shell to make sense of preference purification. IRAs are supposed to be constituted by a psychological mechanism that can generate rational preferences. The psychological shell is assumed to consist of a bunch of other psychological mechanisms that distort the output of the IRA. Hence, ILS argue that preference purification commits us to a dualistic model of human agents, according to which the rational preferences generated by the IRA get distorted by the psychological shell.

ILS's account can then reconstruct the idea behind preference purification in the following way. First, we can account for what purified preferences are by defining them as the output of IRAs. Second, the distortion generated by the psychological shell can explain why we can treat the actual preferences of agents as mistaken. Third, attempts at preference purification can be understood as aiming at recovering the output of the IRA from the distortion of the psychological shell. To put it more vividly, under this view, preference purification amounts to peeling off the irrationality of the psychological shell with the aim of arriving at the rational preferences of the IRA.

In light of this account, ILS argue that preference purification has no psychological basis. The argument here is very straightforward: because there is no plausible psychological theory that can support the existence of IRAs, we have no reason for assuming that there are IRAs. Yet, if there are no IRAs, there are also no purified preferences. Hence, preference purification is a futile exercise. In other words, preference purification lacks solid psychological foundations.[6]

[5]By ontological assumptions, I simply mean assumptions of reference and existence, e.g. 'there are IRAs', or 'there exist certain modes of latent reasoning'. I do not wish to suggest that ILS are engaged in some abstract ontological project that can take place separate from science (for a detailed defence of the fruitfulness of explicating such ontological assumptions of various research programmes, as opposed to doing foundationalist ontology, see Lohse 2017).

[6]ILS (2016*a*) note that their argument could be challenged by advocates of dual process theory (DPT). In particular, proponents of preference purification could try to ground IRAs in System 2 processes. Yet, ILS convincingly dismiss this possibility. Even though I think there is more to be said about this – for instance, about the difference between default-interventionist vs. parallel-competitive interpretations of DPT (see Grayot 2020) – ILS provide an ultimately successful case against this option. I will, therefore, not explore this further here.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

ILS (2016*a*) add that preference purification could be reinterpreted as regularizing the data so that they fit a particular theoretical model. They hold that such regularization is almost always needed when a theoretical model encounters real data. ILS illustrate this with an economic model of the spatial distribution of unemployment. In this model, every job seeker and every job have a clear spatial location. However, in the real world, there are, for instance, people with multiple homes or none. So, economists need to make 'some more or less arbitrary classifications' to fit our data to the scheme of the model (ILS 2016*a*: 21). ILS suggest that something similar is going on when behavioural welfare economists are engaged in preference purification: As social planners we need to regularize the data so that they fit the model of decision-making we want to use. However, while data regularization may be a standard praxis when it comes to descriptive models (i.e. models that aim at explanation and prediction), imposing the standards of the social planner on agents appears to severely undermine the ability of purified preferences to provide us with genuine evidence about welfare.[7]

I, therefore, think that the regularization story is better viewed as an error theory: proponents of preference purification engage in data regularization because they are used to it from descriptive modelling. Yet, because it imposes the perspective of social planners on agents, it does not provide us with a satisfactory proposal for reconstructing preference purification.[8]

3.2. No IRAs needed?

Even though it makes a very convincing case against IRAs, ILS's argument has done little to convince the defenders of preference purification that their attempts are futile. For instance, in a typical reply to ILS, Hausman (2016) agrees that there are no IRAs. Yet, he denies that preference purification is committed to IRAs. His argument rests on the claim that there are uncontroversial 'truth conditions' for what an agent's purified preferences are that do not depend on the existence of IRAs. His idea seems to be that such uncontroversial truth conditions imply that there must be something to which purified preferences refer. Yet, given that there are no IRAs, the story ILS tell must be wrong.

To support this position, Hausman offers an example that is supposed to illustrate a case in which purified preferences can allegedly be determined in an uncontroversial manner (Hausman 2016: 27). If an agent's purified preferences can, at least sometimes, be determined in such a straightforward way, so the idea must be, there is no further need to expand on an account of purified preferences.

In Hausman's example, an agent is quite generally concerned with her health. Yet, she sometimes still acts on her intention to engage in sugar binges. Nevertheless, immediately after these binges, she expresses her regret about them. Moreover, we can assume that it is generally known that sugar binges are

[7]For a more in-depth discussion of the difference between models with descriptive and normative aims, see Beck and Jahn (2021).

[8]A cornerstone of the SR-account introduced below is that it assumes that agents themselves endorse the requirements of structural rationality.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

bad for one's health. Hausman now argues that, in this case, it seems uncontroversial to say that if the agent were fully rational, she would not engage in sugar binges. He tells us that even though some features of the agent's psychology might stand in the way of her getting rid of her habit of occasionally engaging in sugar binges, other features of her psychology, such as her general concern for her health, her regret and the knowledge that sugar binges are bad for one's health can help us to determine what her purified preferences are. For instance, Hausman asks us to imagine a case where the agent prefers sugary cake over fruit. In this instance, we could use our knowledge of the agent's other psychological features to determine that she *would* prefer fruit over cake if she were rational. According to Hausman, this shows that we can sometimes quite easily determine purified preferences. None of this, Hausman explains, requires the existence of an IRA or any assumption about the processes by which the agent's psychological features came about. Therefore, he holds that ILS miss their mark.

ILS (2016*b*) counter Hausman's reply by arguing that one can only make sense of his example by assuming IRAs and offer a battery of counterexamples. I do not think that continuing the debate like this is very fruitful. I take the main problem here to be the absence of an alternative account of purified preferences that avoids IRAs. Hence, I will now turn to outlining and appraising such an account.

4. Two conceptions of rationality

To develop this alternative account, I will leverage two distinct notions of rationality. I will start by outlining procedural rationality (see Kacelnik 2006). I will explain why it fits with the idea that we need IRAs in order to make sense of preference purification (4.1). I will then introduce structural rationality (4.2).[9]

4.1. Procedural rationality

Procedural rationality ascribes rationality to agents that have formed their attitudes by relying on a particular process.[10] This process is usually thought of as correct deliberation or reasoning. According to this notion of rationality, we cannot simply read off rationality by observing an agent's behaviour or getting information about her attitudes. Instead, we can only call the agent rational if she arrives at her attitudes correctly, i.e. by reasoning or deliberating correctly. Consequently, to be considered procedurally rational, one must exhibit a

[9]My argument is not that whether one settles for the IRA-account or the SR-account depends on whether one thinks of procedural or structural rationality as the correct notion of rationality. In fact, I am very suspicious of the claim that there is a capital R theory of rationality. Instead, I want to leverage these two conceptions to show that similar to how one can think of rationality in a procedural or structural way, we can also think of purified preferences in a procedural or structural way. Having said this, the SR-account is, of course, ultimately committed to the view that there is a certain (evidential) relationship between attitudes that exhibit a certain structure and welfare (see section 6.3).

[10]My use of procedural rationality differs from the way in which it is used in Max Weber's distinction between formal/procedural and substantive rationality (Elster 2000).

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

psychological process of the right kind. Agents that fail to execute such processes would be classified as irrational (Kacelnik 2006). For example, we would classify an agent that arrives at a certain belief by employing correct deliberation as rational. Yet, an agent that would form her belief arbitrarily would count as irrational (see Brown 1995). Of course, much of the details of any procedural account of rationality will depend on how we flesh out the idea of correct deliberation.

Nevertheless, this brief sketch of procedural rationality already allows me to explain why it fits nicely with the picture of IRAs. ILS hold that purified preferences are the outcome of IRAs, i.e. a particular psychological process. In other words, unless a preference was formed in a specific way, it cannot count as rational. Hence, if we think about rationality in procedural terms, it becomes easy to see why IRAs become necessary for preference purification. Consequently, I hold that the IRA picture is generally aligned with a procedural notion of rationality. ILS, of course, only focus on this proceduralist notion because they think it offers us the best reconstruction of the implicit assumptions of certain behavioural welfare economists.[11] Again, I do not want to get stuck in a debate about how much of actual practice their account indeed captures. Instead, I want to ask whether there is a non-procedural way of thinking about purified preferences.

What aids me in this attempt is that philosophers have highlighted the need for a different way of thinking about rationality. To illustrate this, consider the following example. An eccentric billionaire offers you a hefty sum for believing p, believing if p then q, and yet believing not-q. Normally, we would say that if you deliberate correctly while believing p and believing if p then q, you should come to believe q. However, let us also say that the price you receive from the billionaire would allow you to ensure that humanity will survive the destruction of the earth by our own hands. It is not too far-fetched to hold that in this case, correct deliberation would result in you believing not-q. Nevertheless, even then, there still seems to be something irrational about the resulting pattern of beliefs (cf. Reisner 2011).

4.2. Structural rationality

To account for such considerations, philosophers, foremost Broome (2010, 2013), have developed what I call a structural notion of rationality. Structural rationality is not concerned with the relation between a fact and an attitude or with how the attitude was formed (Kolodny 2005). Instead, structural rationality is concerned with the relations between an agent's different attitudes.

On the one hand, we can think of structural rationality as a class of requirements. The requirements it issues specify appropriate relations between our attitudes, e.g. between our beliefs, desires and intentions. For example, the instrumental

[11]ILS will argue that we need a procedural account and, ultimately, IRAs to make sense of the claim that individuals would judge their preferences as mistaken by their own lights (cf. Hausman 2016). This claim is intended to fend off charges of excessive paternalism. However, this reconstruction seems to presuppose that agents endorse the perspective of their IRAs. The SR-account introduced below can capture the idea that agents would judge their preferences as problematic by their own lights via presupposing that they endorse the requirements of structural rationality. Hence, I do not see how anti-paternalism boxes us into the proceduralist corner.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

requirement as formulated by Broome (2013) states that rationality requires of agent *N* that, if

(1) *N* intends at *t* that *e*, and if
(2) *N* believes at *t* that *m* is a means implied by *e*, and if
(3) *N* believes at *t* that *m* is up to her herself then, *then*
(4) *N* intends at *t* that *m*.

To illustrate, rationality requires of you that if you have the intention to buy an apple, and if you have the belief that going to the market is a means implied to buy an apple, and if you believe that it is up to you to go to the market, *then* you also have the intention to go to the market. As long as you have the first three attitudes but not the fourth, you violate the instrumental requirement. Simpler examples of requirements of rationality include a requirement for not having contradictory beliefs and a requirement for not having intransitive preferences. Much of the work on structural rationality is dedicated to making such requirements explicit (see Fink 2014).

On the other hand, structural rationality can also be viewed as a property of an agent's mind. In this sense, a person has the property of being (fully) rational iff she satisfies (all) the requirements of rationality she is under, i.e. if (all) her attitudes stand in appropriate relations to each other (Broome 2010, 2013). What is important for my argument is that even though structural rationality can be seen as a property of the mind, it is a property possessed in virtue of the relations between an agent's attitudes and not a property possessed in virtue of any (cognitive) processes an agent can execute. In other words, what matters for structural rationality is only whether an agent's attitudes stand in appropriate relations to each other. Therefore, structural rationality is a coherentist notion of rationality. In principle, an agent can have the property of structural rationality without executing any particular psychological process like reasoning or deliberating correctly. The agent could simply satisfy the requirements of structural rationality by default, or she could have had help from a friend who points out the mismatches between her attitudes and suggests how to resolve them.

5. Structural rationality and preference purification

In this section, I will outline how we can think of preference purification in terms of structural rationality. I call this the SR-account (5.1). I will then illustrate this account with a version of Hausman's example from section 2 (5.2). Finally, I outline how the debate about preference purification stands to benefit from introducing the SR-account (5.3).

5.1. The SR-account

The SR-account defines purified preferences in the following way:

> The purified preferences of an agent are those preferences contained in the attitudes of an alternative version of the agent who satisfies the

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

requirements of structural rationality but whose attitudes are otherwise as close to the attitudes of the actual agent as possible.

Two comments are in order here. First, what do I mean by 'attitudes being otherwise as close to the attitudes of the actual agent as possible'? The idea is that we should not make any unnecessary changes to an agent's attitudes to arrive at a structurally rational version of her that should be viewed as exhibiting the agent's purified preferences. If rationality can basically be viewed as having a well-ordered mind, we could always try to make things even tidier by decluttering. In general, the fewer attitudes there are, the less potential for them to stand in inappropriate relations to each other.

In light of this, the intuition here is that we should not drop, add or change more of the agent's attitudes when we can also make the agent structurally rational by doing less. In this sense, only the closest structurally rational version of the agent should be seen as determining her purified preferences. Of course, a lot of detailed work will have to be done on the appropriate notion of closeness here. Questions relevant in this regard include how we should individuate attitudes, whether different types of modifications are really on par (e.g. changing vs. dropping an attitude), and whether different types of attitudes should be all weighted equally. However, to get the SR-account off the ground, I rely on an intuitive understanding regarding what lesser and greater alterations of an agent's attitudes are.[12]

Second, I define purified preferences in terms of the requirements of structural rationality *simpliciter* rather than restricting them to the requirements pertaining only to the attitudes relevant to the choices for which we want to purify the agent's preferences. Given that all of us may virtually always fail to satisfy some of the requirements of structural rationality, not introducing such a restriction may appear overly demanding.

However, focusing just on some attitudes may allow us to arrive at a version of the agent that is locally structurally rational, i.e. structurally rational only with respect to the attitudes in question. Yet, it is possible that these local modifications can lead to the agent being globally, i.e. with respect to all her attitudes, less rational. It would be implausible to allow purified preferences to be grounded in globally less rational attitudes than the agent's actual attitudes. Nevertheless, for practical reasons, we will certainly have to restrict our focus to the subset of an agent's attitudes relevant to the choices in question when trying to derive purified preference. In this regard, proponents of preference purification will have to make the bet that modifying our attitudes locally will generally not lead to more irrationality globally.

With these remarks out of the way, let me expand on how the SR-account matches ILS's account. First, as stated above, the SR-account gives us a clear definition of what purified preferences are. In other words, it gives us an explication of Hausman's idea that there are 'truth conditions' for purified preferences. Importantly, *pace* ILS, we, therefore, get a definition of purified preferences that does not rest on the assumption of IRAs.

[12]I briefly return to this issue in section 6.5.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

Second, it also gives us a clear standard for when we can call a set of attitudes, including preferences, rational or irrational. The SR-account allows us to classify certain sets of preferences as 'mistaken' on the basis that they do not conform to the requirements of structural rationality (cf. Dietrich *et al.* 2021). Crucially, *pace* the IRA picture, making this classification does not require us to postulate a distorting psychological shell or to make any assumptions about how those attitudes were formed.

Third, the SR-account also provides us with an idea of what the procedure of preference purification amounts to. In a nutshell, preference purification is the attempt of asking how an agent's preferences would look like if she were to satisfy the requirements of structural rationality. In this regard, structural rationality offers us a standard for what it would mean to clean up people's preferences. Importantly, *pace* ILS, the external standard provided by structural rationality also makes clear why preference purification is not just an attempt at 'regularizing' our data so that it fits the favoured model of the social planner.

In sum, considering the SR-account, we do not need to assume a dualistic model of the human agents to make sense of preference purification.[13] To put things more vividly, under the SR-account, preference purification is more analogous to cleaning up one's room than to peeling onions, as the IRA picture would have it.

5.2. Returning to Hausman's example

In order to illustrate the SR-account and to show how it fits with what defenders of preference purification have argued, I will focus on a paraphrased version of Hausman's example from section 2. In Hausman's example, the agent is concerned with her health. We can plausibly restate this as having an intention to stay healthy. It also seems to be the case that she believes that fruit is the means implied to stay healthy. Supposedly, this is why she expresses regret over her choice for cake. Moreover, without controversy, we can assume that the agent in the example believes that she can either pick cake or fruit. Hence, it makes sense to ascribe the following attitudes to her: (i) an intention to stay healthy, (ii) the belief that choosing the fruit over cake is the means implied to stay healthy, and (iii) the belief that it is up to her to choose between fruit and cake.

If we now look at the Instrumental Requirement from section 3, we can see that structural rationality requires of you that if you have those three attitudes, you should also have the intention to choose fruit over cake. Consequently, if the agent lacks this intention and instead has the incompatible intention to choose cake over fruit, she would be structurally irrational. Yet, if we were to switch her intention to choose cake with an intention to choose fruit, the thus created

[13]In principle, the SR-account and the IRA-account could be extensionally equivalent. To see this, imagine Aby, who intends to save for her retirement but fails to sign up for her firm's retirement plan. According to the IRA-account, Aby's IRA would sign up for the plan, but the psychological shell prevents her from doing so. According to the SR-account, Aby violates the requirements of structural rationality, but the closest sets of structurally rational attitudes would have her sign up. Both accounts reach the same conclusion albeit for different reasons. In cases like this, the accounts provide us with two independent rationales for the same purification. However, I have already argued that ILS make a convincing case against IRAs. Hence, while the possibility of having two independent rationales is interesting, I will not further pursue it here.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

alternative version of the agent would satisfy the Instrumental Requirement and, therefore, count as more rational.

An ideal application of the SR-account would now demand that we go through all the requirements of structural rationality an agent is under and ask how we would have to change her attitudes to arrive at an alternative version of her that satisfies them while staying as close as possible to the agent's actual attitudes.[14] Whatever preferences are contained in the outcome of this procedure are the agent's purified preferences.

Of course, any plausible method for implementing preference purification in practice will likely be an imperfect approximation of the general idea behind it. For instance, it will certainly have to restrict itself to a subset of the agent's attitudes (i.e. those that matter for the choices at hand). Yet, given that there is a reasonable case to be made for viewing scientific methods as fallible heuristics (see Grüne-Yanoff 2021), this should not deter us from attempting to develop such methods if the general idea behind preference purification is indeed plausible. Here I do not wish to argue for any specific way of approximating the general idea behind preference purification under the SR-account.

Nevertheless, to illustrate how such an approach could look like, let us assume that decision-making in line with expected utility theory proceeds from attitudes that conform to the requirements of structural rationality (cf. Dietrich *et al.* 2021). Under this assumption (which I do not necessarily endorse), multiple choices that jointly violate the independence-of-irrelevant-alternatives assumption could be seen as indicating that the agent fails to satisfy some of the requirements of rationality. In an attempt to get to the closest set of attitudes that satisfies the requirements of rationality, we may then try to estimate an expected utility model that most closely fits the agent's actual choices (cf. Harrison and Ng 2016; Harrison and Ross 2018: footnote 12). This model may then be interpreted as representing a version of the agent that satisfies the requirements of rationality but is otherwise as close to the agent's actual attitudes as possible. One may, of course, challenge whether this is a good approximation of the SR-account and question how such an approach would look like in detail. However, for now, I only seek to indicate how the SR-account can motivate attempts for approximating it in practice.

5.3. Taking stock

The SR-account provides us with an alternative picture of preference purification that does not rely on IRAs. Consequently, a fruitful debate about preference purification will have to acknowledge that there are, at least, two distinct accounts of its conceptual foundations. Both camps stand to benefit from it. More obviously, proponents of preference purifications should be interested in an account that explicitly demonstrates how they avoid IRAs. However, also

[14]The literature on structural rationality has mostly explicated synchronic as opposed to diachronic requirements of rationality. Most of the explicated requirements also deal with flat-out beliefs as opposed to credences. Yet, in order to tackle interesting cases of preference purification, we will also need to explicate diachronic requirements and requirements dealing with credences. If I am right, further exploration of these issues might be an interesting opportunity for philosophers to contribute to economic methodology.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

those who oppose preference purification have good reasons to clearly distinguish the two accounts. The main reason for this is that proponents of preference purification such as Hausman may otherwise have an easy time sliding back and forth between different understandings of purified preferences to avoid criticisms.[15] This is especially troublesome as there are problems for the SR-account that preference purification would not face under the IRA picture.

In this regard, I now assess two pressing challenges for the SR-account. While the implausibility of IRAs would be decisive under a procedural picture, I hold that the prospects of preference purification under the SR-account are more promising. To motivate this, I will outline how we can make progress towards addressing the two challenges against the SR-account and highlight the commitments that proponents of preference purification will need to make along the way. In particular, I will argue that addressing these challenges will require combining the SR-account with the so-called evidential account of the relationship between preferences and welfare (see Hausman and McPherson 2009). If this is right, defenders of preference purification may also have to endorse certain substantive (albeit still minimal) commitments about welfare and its relationship to preferences that, at least, some of them may wish to avoid in the name of neutrality.[16]

6. Preference purification and the evidential account

I will now briefly introduce two challenges for the SR-account (6.1), and then sketch the evidential account (6.2). I will then show that the evidential account offers us resources for addressing these challenges (6.3 and 6.4). The last subsection (6.5) addresses two objections against my arguments.

6.1. Two challenges for the SR-account

The first challenge for the SR-account is to explain how a set of hypothetical, structurally rational preferences is related to an agent's welfare. To put it differently, why should we be concerned with making judgements about welfare based on *hypothetical* sets of purified preferences that are not actual subjective states of the relevant agents (cf. Thoma 2021)? Arguably, the IRA-account avoids this challenge if it conceives of purified preferences as the *actually generated* output of IRAs that then gets distorted by the psychological shell.[17]

The second challenge is a problem of underdetermination. If there are multiple ways in which we could change an agent's set of attitudes that result in structurally rational sets of attitudes that are all equally close to the agent's actual attitudes, which of them should we take to contain the agent's purified preferences? This

[15]While I think that there is overall more reason to understand Hausman as implicitly endorsing the SR-account, ILS (2016*b*) highlight that some of his writings may suggest that he endorses the IRA picture. This opens space for interpreting Hausman as endorsing different views at different times. I thank Bob Sugden for emphasizing this possibility.

[16]I thank an anonymous reviewer for highlighting this.

[17]However, if the psychological shell interferes with the IRA before it can generate its output, then also under the IRA-account purified preferences are hypothetical mental states, i.e. mental states the agent would have had if the psychological shell had not intervened.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

is not a problem under the IRA-account. If we assume IRAs, then there is a unique answer to what an agent's purified preferences are. They are the output of the IRA.

6.2. The evidential account

Hausman and McPherson's (2009) argument for the evidential account starts with a survey of the familiar objections against the idea that preference satisfaction *constitutes* welfare. Based on these objections, they convincingly argue that welfare cannot be identified with preference satisfaction (even after purification). Yet, they do not think that this makes preferences irrelevant for forming judgements about welfare.

The evidential account's core claim is that preferences are reliable indicators for welfare if certain conditions are met. These conditions are as follows. Agents need to (i) have preferences that concern their own welfare, (ii) these preferences need to be sufficiently based on the relevant facts, and (iii) the preferences need to be free of all rational flaws, e.g. they need to have transitive preferences (see also Hausman 2012, 2016).

How are these conditions supported? The evidential account does not rest on a theory of welfare. In fact, Hausman (2015) seems sceptical that any substantive theory can be articulated here. Instead, the evidential account supports its conditions with platitudes about welfare, i.e. commonly agreed-upon propositions about what is conducive to people's welfare (Hersch 2015). Those platitudes are meant to support that preferences can be good guides to welfare if the relevant conditions are met. Importantly, even preferences that meet all these conditions should only be viewed as good but fallible guides to welfare (Hausman 2016).

Two points need emphasis here. First, the evidential account is a theory-free approach to welfare (see Fumagalli 2021). Hausman (2015) argues that welfare is a dynamic structure in which various goods such as friendship, happiness, health or a sense of purpose are integrated. Yet, he also holds that 'our evaluative abilities are limited with respect to our own lives' (Hausman 2015: 141). So, we are somewhat clueless about how exactly these various goods integrate into welfare. Yet, Hausman thinks that we do not need to have a full-fledged theory of welfare to know what *good sources of evidence* for welfare are. In fact, based on platitudes about welfare, we can converge on those sources independent of theory.

This brings us to the second point: the nature of platitudes. Here the idea is that while it may be futile to attempt the articulation of a full-fledged theory of welfare, there are nonetheless widely shared propositions about welfare because we all live human lives and try to do so well. Those platitudes allow us to identify plausible sources of evidence for welfare without the need to appeal to an 'exhaustive list of intrinsic goods and [the evidential account] depends on no philosophical theory that specifies what things are intrinsically good for people' (Hausman and McPherson 2009: 19). Platitudes basically refer to widely shared institutions about welfare. While I share Hausman's scepticism about the ability of such platitudes to guide us towards a full-fledged theory of welfare, this does not prevent them from guiding us towards plausible sources of evidence for welfare.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

Considering this, a final verdict on the evidential account will require, on the one hand, evidence that the intuitions that support it are indeed widely shared and, on the other hand, an extensive probing of those platitudes via intuition pumps. One important difference between such an assessment and the game that is played in debates about theories of welfare is that the relevant intuitions do not concern what 'welfare really is', but rather whether a particular piece of information is good evidence for welfare. A full defence of the evidential account is more than I can offer here. However, it suffices to highlight its initial plausibility to demonstrate how the evidential account can be combined with the SR-account to address the challenges outlined above.[18]

6.3. Not my preferences

The reason for thinking that preferences need to be *actual* mental states to be relevant for welfare is that unless purified preferences are actual subjective states, satisfying purified preferences does not constitute an improvement in agents' welfare (cf. Thoma 2021). While satisfying hypothetical preferences would bear on the agent's welfare if she were to exhibit them, satisfying those preferences will not impact the agent's welfare in the actual world.

The evidential account offers a quick reply here. Even if an agent does not actually possess structurally rational attitudes, knowing what she would prefer if she would have the relevant attitudes could still be considered valuable evidence for welfare. If what makes something good for her is *not* the satisfaction of her preferences itself, then not actually possessing those preferences does not need to influence what is good for the agent.

On top of that, combining the evidential account with the SR-account allows us to build a more positive reply. The first step is to note that structural rationality can offer us an analysis of what it means to satisfy condition (iii) of the evidential account (i.e. 'being free of rational flaws'). We can spell this out via the requirements of structural rationality. To support the plausibility of this reading, consider that there is reason to doubt whether preferences that are part of a 'messy' mind can be reliable guides to welfare. In contrast, in the case of a well-structured mind, no such worry occurs. To further motivate this, consider that if preferences are supposed to represent (or be indicative of) 'betterness' along the dimension of welfare, the logic of comparative adjectives such as 'better than' requires them to be transitive (cf. Broome 1991). So, if they are not transitive, we have reasons to doubt that they track welfare (cf. Hausman 2012: 19). Note that nothing in this argument requires us to assume that intransitive preferences should be disregarded because they result from unsound reasoning or the

[18]Several objections against the evidential account have been raised. For instance, Hersch (2015) argues that the evidential account cannot discriminate between measures of welfare that frequently deliver contradictory results. Moreover, Fumagalli (2021) argues that the domain of application of the evidential account is quite limited and that without theory we cannot determine when condition (i) is met. However, Hersch's argument seems to miss the fallible character of the evidential account. Moreover, as will become clear below, preference purification under the SR-account may allow us to address Fumagalli's worry about the limited domain. Finally, I do not take the potentially fuzzy boundaries of condition (i) to be a decisive reason against relying on the evidential account.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

distortion of the psychological shell. Instead, we doubt that they reliably track welfare in virtue of their structural features. In fact, excluding clear-cut cases of cognitive error that ILS are also fine with correcting, I take this image to offer a far better explanation of why certain preferences do not track welfare than an appeal to specific cognitive processes or procedural errors.[19] Hence, it seems plausible that satisfying the requirements of structural rationality is one of the necessary prerequisites for preferences tracking welfare.

The next step is then to show that certain sets of hypothetical, structurally rational preferences are also good guides to an agent's welfare. In particular, we have to establish that the preferences contained in hypothetical sets of attitudes that satisfy the requirements of structural rationality but are otherwise as close to the attitudes of the actual agent as possible can also be good guides to an agent's welfare.

How could we establish this? In the absence of an explicit theory, it can be tricky to offer a direct argument here. However, the evidential account can, of course, avail itself to platitudes about welfare. What will become crucial, then, is how we assess the following principle:

> Evidential Extension: If the only change between two states of the world A and B is that an agent, who does not satisfy the requirements of structural rationality in A, satisfies the requirements of structural rationality in B, while her attitudes in B are otherwise as close as possible to the attitudes she has in A, and the preferences contained in her attitudes in B are good guides to the agent's welfare in B, then those preferences in B are also good guides to her welfare in A.

I take it that Evidential Extension has initial plausibility in the light of what I take to be commonly agreed-upon propositions about welfare. If we accept this principle, we grant that certain hypothetical preferences can deliver evidence that can be as useful as the evidence gained from certain actual preferences.

Of course, one may challenge the plausibility of Evidential Extension. It could, for instance, be argued that evidence that something is good for an agent in B will not be evidence that it is also good for her in A precisely because of the different attitudes she has in A. To illustrate, let us assume that the agent's preferences in B give us good evidence that eating an apple would be good for her. Evidential Extension would then tell us that eating an apple is also good for her in A. However, one might object that if we give her the apple in A, she will likely use the apple in ways that are not conducive to her welfare (e.g. trade the apple for a chocolate bar and perhaps even pay extra to be able to do so).[20] However, this objection overlooks that what we are after is evidence for what is good for an agent's welfare and not an overall recommendation of what to do. Nothing in Evidential Extension tells us how we should make use of the evidence that eating an apple will improve the agent's welfare. If there is further evidence that simply providing certain goods, which would be conducive to agents' welfare if consumed, creates

[19]The evidential account would tackle cases of obvious cognitive error via condition (ii).

[20]I thank an anonymous reviewer for raising this objection.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

incentives for acting in ways that reduce welfare, simply providing those goods is not necessarily what we should do.[21] Instead, it would be more prudent to rationally persuade the agent of the apple. The outlined objection does, therefore, not provide us with a successful intuition pump against Evidential Extension.

A more complete evaluation of Evidential Extension would, of course, require us to subject it to further intuition pumps. However, I maintain that it has initial plausibility. If this is correct, my discussion suggests that the evidential account is highly promising when it comes to addressing the worry that purified preferences become mere hypothetical preferences under the SR-account.[22] There are good reasons for holding that, under the evidential account, this does not prevent them from informing judgements about welfare. In this regard, Evidential Extension turns out to be a crucial principle that defenders of preference purification should accept. Preference purification under the SR-account can, thus, be viewed as extending the scope of the evidential account.

6.4. Too many purified preferences

The second challenge is motivated by the fact that the requirements of rationality advanced by the most prominent account of structural rationality (i.e. Broome 2013) frequently allow for several modifications to an agent's attitudes. These different modifications can result in different sets of structurally rational attitudes. However, prima facie, none of these modifications can be viewed as a greater departure from the agent's actual attitudes. Therefore, none of these sets of attitudes can be privileged on the grounds that it is closer to the agent's actual attitudes. If this happens, the question becomes: which of these different sets of attitudes we should use to make judgements about the agent's welfare?

To motivate why this can be a severe problem, consider the example of Hausman again. I have argued that the agent initially violates the Instrumental Requirement and shown that this violation can be overcome by switching the intention to choose cake with the intention to choose fruit, and that this matches Hausman's claims about purified preferences. However, according to a Broomean framework, another way to satisfy the Instrumental Requirement, in this case, would be to simply drop the intention to stay healthy.[23]

[21]Under the evidential account preference satisfaction and choice do not directly constitute welfare. Yet, under the right conditions the results of our choices are usually such that they promote welfare. Hence, while we can learn from agents' preferences, we do not need to assume that choice directly generates welfare. To give a simple example, buying a home near a nice park is evidence that living near a park is good for the agent. Yet, it does not mean that the act of buying the house itself generates substantial welfare effects.

[22]There is a debate about idealizations and subjectivism about welfare. Some authors, such as Railton (1986) and Sobel (2009), hold that only the satisfaction of desires we would have if we were idealized in certain ways constitute welfare. Others, such as Enoch (2005), argue that subjectivism lacks a rationale for such idealizations. Given that idealized desires can be seen as mere hypothetical attitudes, one may think that this debate is highly relevant in the present context. However, subjectivism is an account of the constituents of welfare. Yet, the evidential account holds that welfare cannot be reduced to (idealized) preference satisfaction. It only claims that, under the right conditions, preferences reliably track what is good for us.

[23]There is another requirement of rationality, the basing prohibition, that prevents one from dropping the two beliefs. Very roughly, you would be irrational for basing a belief on you not intending something. Yet, Broome (2013: 155) holds that there is no corresponding requirement for intentions.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

Dropping the intention to stay healthy would also make the agent rational because the Instrumental Requirement only states that rationality requires of us that *if* we have a particular set of attitudes, we also have another attitude (e.g. the intention to choose fruit).[24] Hence, by removing some of the initial attitudes, we satisfy the requirement trivially. Moreover, (in contrast to the two beliefs) there are no further requirements of rationality, or so Broome argues, that can prevent us from dropping the intention to stay healthy.[25] As a result, we can end up with at least two structurally rational versions of the agent that *both* result from just one modification of the agent's attitudes. Prima facie, none of these two alternative versions can be seen as closer to the actual agent. Yet, they can entail contradictory advice. One set of attitudes, implies that cake is good for the agent, while the other implies that we should nudge her towards fruit. So, which of these sets of attitudes should be viewed as containing the agent's purified preferences?

Of course, one can plausibly argue that dropping the intention to stay healthy, unlike the intention to choose cake, will give rise to structural irrationality in a large set of other cases. Therefore, we should focus on the intention to choose cake. Nevertheless, I wish to grant that there can frequently be multiple structurally rational versions of the agent that are all equally close to the agent's actual attitudes and could, therefore, all be viewed as exhibiting the agent's purified preferences.

I now outline that the evidential account allows us to deal with this underdetermination. In particular, in Hausman's example, there seem to be good additional reasons – that do not stem from concerns about structural rationality – for not dropping the intention to stay healthy. Further evidence about welfare and health could serve as tiebreaker for selecting between several structurally rational sets that are all equally close to the agent's actual attitudes. Crucially, appealing to such further sources of evidence does not make preference purification obsolete.

One may argue that if we bring in further sources of evidence, it becomes questionable why we rely on preference purification in the first place. If other sources already enable us to make judgements about welfare, why do we not apply them more directly? If we already know that health is generally conducive for an agent's welfare, why do we need to go through the trouble of purifying preferences? Why not just directly nudge her to take fruit? Hence, one could worry that introducing further sources of evidence threatens to undermine the relevance of preference purification.

My reply to this is that, under plausible weightings of different sources of evidence, preference purification can make use of other sources of evidence. To illustrate, let us assume that based on our general experience with our own and

[24]As Broome (2013) argues, plausible formulations of the requirements of rationality need to be wide as opposed to narrow scope.

[25]One could argue that the attitude in Hausman's example is more appropriately construed as *valuing* health. It could then be claimed that there is an equivalent to the basing prohibition for beliefs that applies to valuings and prevents us from dropping the relevant attitude. Yet, as I will argue below, this move is not necessary.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

other people's welfare, we agree that living healthily is, in general, conducive for an agent's welfare. Yet, accepting this platitude about welfare does not tell us how we should judge the welfare of someone who consistently prefers an unhealthy diet containing a lot of cake. There is no problem with holding that despite the correctness of the platitudes, eating a lot of cake is indeed conducive to the welfare of the agent in question. To emphasize this, imagine that we are in a context in which the conditions of the evidential account are met. The agent is, therefore, not violating any requirements of structural rationality. Hence, the agent's preference for cake would be good evidence for the fact that cake is indeed good for her. This can override the platitude that living healthily is in general conducive for an agent's welfare.

Importantly, that we give more weight to the agent's preferences if the conditions of the evidential account are met does, of course, not mean that we discard the evidence encoded in the relevant platitude. However, one can plausibly hold that this evidence is not strong enough to convince us that the cake is bad for the agent's welfare in light of preferences that meet the conditions of the evidential account. Note that holding that preferences which meet these conditions are stronger evidence than other types of evidence does not require one to make assumptions about preference formation (e.g. IRAs). The SR-account supports the strong link between preferences and welfare via widely shared intuitions and remains open to different mechanisms attuning our preferences to welfare. We can have knowledge of expertise without knowing exactly how experts work. Similarly, we can have knowledge that preferences under the right conditions are good guides to welfare without knowing exactly how they were formed.[26]

Now, if several sets of structurally rational preferences are equally close to the agent's actual attitudes, Evidential Extension suggests that each bears the same evidential punch concerning what is conducive for an agent's welfare. Hence, we seem to find ourselves in an evidential deadlock if we only consider the evidence we get from preferences. Luckily, if we also accept the platitude about health and welfare, there is an easy way to tip the scales. While we may not consider the relevant platitude to be strong enough to override the evidence that we can get from preferences if the conditions of the evidential account are met, this does not prevent it from functioning as a tiebreaker for selecting between multiple, structurally rational sets of attitudes that are all equally close to the agent's actual attitudes.

Due to this tiebreaker, there can be overall more evidence that an apple is good for you than that cake is good for you. This is entirely compatible with the idea that evidence from preferences (at least under the right conditions) is the best kind of evidence and usually outweighs other types of evidence. Yet, the tiebreaker strategy would not be open to us if we opted for a constitutive link between preferences and welfare because then information about preference would ultimately be all the information about welfare there is. Hence, the evidential account introduces the possibility of tiebreakers, which adds to its attractiveness for proponents of preference purification under the SR-account.

[26]More paternalistically inclined people will believe that other types of evidence weigh usually more than agents' preferences. However, arguing against such sceptics falls outside the scope of the present project.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

6.5. Objections

Before I conclude, let me address two worries about my assessment of the SR-account. The first worry concerns the fact that we are aiming at a moving target. For instance, an agent's next choice for cake may cement her preference for cake and, thereby, eliminate her structural irrationality. Now, if we would attempt to purify her preferences before this pivotal choice, a tiebreaker may lead us to the wrong conclusion that fruit is good for her. This indicates the need to be mindful of our target's moving nature before jumping to any policy conclusions based on the evidence we get from purified preferences. Yet, the SR-account itself would remain untouched by this worry. After all, after the next choice for cake, the closest structurally rational version of the agent would be the actual agent. Therefore, after the next choice of cake, we are in a position in which we have most evidence that cake is good for the agent. Yet, before she made that choice, the most justified conclusion could still have been that fruit is good for her. Whether we should already have acted on that information is ultimately a question that cannot be settled merely by an analysis of preference purification.[27]

The second worry relates to the notion of closeness in the SR-account. I have argued that several questions still need to be answered in order to develop an appropriate notion of closeness. If one is sceptical about our ability to answer those questions, one will doubt that we can really determine a meaningful ranking of sets of attitudes in terms of closeness to the agents' actual attitudes. In the following, I will address two concerns on which such scepticism can be founded.

To illustrate the first concern, consider the following example: A person has the actual attitude set {a,b,c,d,e}, and {a,b,c} and {d,e} are both maximal sets or rational attitudes. A naïve reading of my proposal would be that {a,b,c} is a smaller departure from the agent's actual attitudes than {d,e}. However, one may now worry that d or e are attitudes that are more important to the agent and should therefore receive a greater weight (e.g. simple desires vs. deep values). One could object that respecting such weightings makes developing an appropriate notion of closeness excessively difficult.[28] However, if we are really dealing with qualitatively different attitudes, structural rationality will likely not permit us to pick and choose which of them to modify (cf. footnote 25). The reason for this is that the leeway we get from structural rationality usually concerns attitudes of the same type (e.g. one intention vs. another intention). Hence, that different attitudes should be weighted differently does not appear too troublesome for the SR-account.[29]

The second concern about closeness is that there will always remain an ambiguity regarding how many attitudes an agent has. Consider, for example, the belief that

[27]I thank an anonymous reviewer for introducing this concern.

[28]I am grateful to Bob Sugden for providing this example.

[29]I assume that we can distinguish these attitudes in a non-normative way. This is because whether something is just a garden-variety belief or a deeply held commitment will certainly have an impact on an agent's thinking, feeling, and acting and is, therefore, a descriptive matter. Moreover, determining the weight of an attitude does not necessarily involve making assumptions about preference formation. I may correctly determine that Donald likes ice cream and loves Olaf without having any information about how these two attitudes were formed. Yet, I am certain that the second attitude should weigh heavier than the first.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

there are fewer than 9 planets in the solar system. Does having this belief also mean that one believes that there are fewer than 11 planets? Or fewer than 198? In short, if we were to intervene on this belief, how many attitudes would be affected? Due to such considerations, one may be sceptical that sets of attitudes can be meaningfully compared in terms of closeness to an agent's actual attitudes. However, while these considerations may imply that it will be hard to give a definite answer to the question of how many attitudes an agent has overall, this does not prevent us from giving meaningful rankings of closeness for sets of attitudes for particular cases in which we can restrict ourselves to the attitudes relevant for the case at hand. Moreover, we could also spell out the notion of closeness in terms of the number of hypothetical interventions that we need to make to arrive at a structurally rational set of attitudes. Consider an agent who believes that there are fewer than 9 planets, that there are more than the square root of 81 planets, and that the square root of 81 is 9. To make this agent structurally rational, one of these beliefs must go. There may be reasonable disagreement about how many attitudes we affect by dropping either of these beliefs. However, it seems less controversial that all we need is one intervention on her attitudes. In sum, while answering questions about the correct notion of closeness is important for fleshing out the SR-account, it does not pose insurmountable challenges.

7. Conclusion

In this paper, I have explicated the SR-account that promises to unravel the guiding idea behind preference purification. The account evades the problem of IRAs that preference purification would face under ILS's picture. I have also started the appraisal of the SR-account by showing that it fits well with the evidential account. Combining these two accounts allows us to make progress with respect to two pressing challenges against the SR-account. If all of this is correct, we have good reasons to think that the SR-account offers a more defensible outlook on preference purification than the procedural picture that would commit us to IRAs.

However, it would also mean that proponents of preference purification still have much work to do. For instance, I have mentioned that the SR-account stands to benefit from a more elaborated notion of closeness. Moreover, the evidential account and the Evidential Extension principle should be subjected to further discussion and a battery of intuition pumps. Nevertheless, framing the problem in terms of structural rationality would shift the debate into a more fruitful direction. It would allow us to see that instead of being a misguided attempt at finding the Econ within us, preference purification can aim at a more coherently structured version of ourselves: the Econ above.

Acknowledgements. I thank Måns Abrahamson, Anna Alexandrova, Marcel Jahn, Bob Sugden, Bele Wollesen and two anonymous referees for very valuable feedback on earlier drafts of this paper. I also thank Ranjani Parthasarathy for checking and improving my English.

Financial support. The research that formed the foundation of this article was supported by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

References

Adler M.D. and E. Posner 2006. *New Foundations of Cost-Benefit Analysis*. Cambridge, MA: Harvard University Press.

Ariely D. 2010. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York, NY: Harper Perennial.

Beck L. and M. Jahn 2021. Normative models and their success. *Philosophy of the Social Sciences* **51**, 123–150.

Bernheim B.D. 2016. The good, the bad, and the ugly: a unified approach to behavioral welfare economics. *Journal of Benefit-Cost Analysis* **7**, 12–68.

Bernheim B.D. and A. Rangel 2008. Choice-theoretic foundations for behavioral welfare economics. In *The Foundations of Positive and Normative Economics: A Handbook*, ed. A. Caplin and A. Schotter, 155–192. Oxford: Oxford University Press.

Bernheim B.D. and A. Rangel 2009. Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics* **124**, 51–104.

Bleichrodt H., J.L. Pinto and P. Wakker 2001. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* **47**, 1498–1514.

Broome J. 1991. Utility. *Economics & Philosophy* **7**, 1–12.

Broome J. 2010. Rationality. In *A Companion to the Philosophy of Action*, ed. T. O'Connor and C. Sandis, 283–292. Chichester: Wiley-Blackwell.

Broome J. 2013. *Rationality Through Reasoning*. Chichester: Wiley-Blackwell.

Brown H.I. 1995. *Rationality*. In *The Oxford Companion to Philosophy*, ed. T. Honderich, 744–745. Oxford: Oxford University Press.

Camerer C., S. Issacharoff, G. Loewenstein, T. O'Donoghue and M. Rabin 2003. Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review* **151**, 1211–1254.

Dietrich F., A. Staras and R. Sugden 2021. Savage's response to Allais as Broomean reasoning. *Journal of Economic Methodology* **28**, 143–164.

Elster J. 2000. Rationality, economy, and society. In *The Cambridge Companion to Weber*, ed. S. Turner, 21–41. Cambridge: Cambridge University Press.

Enoch D. 2005. Why idealize? *Ethics* **115**, 759–787.

Fink J. 2014. A constitutive account of 'rationality requires'. *Erkenntnis* **79**, 909–941.

Fumagalli R. 2021. Theories of well-being and well-being policy: a view from methodology. *Journal of Economic Methodology* **28**, 124–133.

Glaeser E.L. 2011. The moral heart of economics. New York Times, <https://economix.blogs.nytimes.com/2011/01/25/the-moral-heart-of-economics>

Grayot J.D. 2020. Dual process theories in behavioral economics and neuroeconomics: a critical review. *Review of Philosophy and Psychology* **11**, 105–136.

Grüne-Yanoff T. 2018. Boosts vs. nudges from a welfarist perspective. *Revue d'économie politique* 128, 209–224.

Grüne-Yanoff T. 2021. Justifying method choice: a heuristic-instrumentalist account of scientific methodology. *Synthese* **199**, 3903–3921.

Harrison G.W. and J.M. Ng 2016. Evaluating the expected welfare gain from insurance. *Journal of Risk and Insurance* **83**, 91–120.

Harrison G.W. and D. Ross 2018. Varieties of paternalism and the heterogeneity of utility structures. *Journal of Economic Methodology* **25**, 42–67.

Hausman D.M. 2012. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hausman D.M. 2015. *Valuing Health: Well-being, Freedom, and Suffering*. Oxford: Oxford University Press.

Hausman D.M. 2016. On the econ within. *Journal of Economic Methodology* **23**, 26–32.

Hausman D.M. and M.S. McPherson 2009. Preference satisfaction and welfare economics. *Economics & Philosophy* **25**, 1–25.

Hausman D.M. and B. Welch 2010. Debate: to nudge or not to nudge. *Journal of Political Philosophy* **18**, 123–136.

Hersch G. 2015. Can an evidential account justify relying on preferences for well-being policy? *Journal of Economic Methodology* **22**, 280–291.

https://doi.org/10.1017/S0266267122000141 Published online by Cambridge University Press

Infante G., G. Lecouteux and R. Sugden 2016*a*. Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology* **23**, 1–25.

Infante G., G. Lecouteux and R. Sugden 2016*b*. 'On the Econ within': a reply to Daniel Hausman. *Journal of Economic Methodology* **23**, 33–37.

Johnson E.J. and D. Goldstein 2003. Do defaults save lives? *Science* **302**, 1338–1339.

Kacelnik A. 2006. Meanings of rationality. In *Rational Animals?*, ed. S. Hurley and M. Nudds, 87–106. Oxford: Oxford University Press.

Kahneman D., J.L. Knetsch and R.H. Thaler 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* **98**, 1325–1348.

Kolodny N. 2005. Why be rational? *Mind* **114**, 509–563.

Lohse S. 2017. Pragmatism, ontology, and philosophy of the social sciences in practice. *Philosophy of the Social Sciences* **47**, 3–27.

Park C.W., S.Y. Jun and D.J. MacInnis 2000. Choosing what I want versus rejecting what I do not want: an application of decision framing to product option choice decisions. *Journal of Marketing Research* **37**, 187–202.

Rabin M. 2002. A perspective on psychology and economics. *European Economic Review* **46**, 657–685.

Railton P. 1986. Moral realism. *The Philosophical Review* **95**, 163–207.

Read D. and B. Van Leeuwen 1998. Predicting hunger: the effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes* **76**, 189–205.

Reisner A. 2011. Is there reason to be theoretically rational? In *Reasons for Belief*, ed. A. Reisner and A. Steglich-Petersen, 34–53. Cambridge: Cambridge University Press.

Salant Y. and A. Rubinstein 2008. (A, f): choice with frames. *Review of Economic Studies* **75**, 1287–1296.

Savage L. 1954. *The Foundations of Statistics*. Chichester: John Wiley & Sons.

Sobel D. 2009. Subjectivism and idealization. *Ethics* **119**, 336–352.

Sugden R. 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

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

Thoma J. 2021. On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology* **28**, 350–363.

Tversky A. 1969. Intransitivity of preferences. *Psychological Review* **76**, 31.

Lukas Beck is a member of the Scientific Assessments, Ethics, and Public Policy working group at the Mercator Research Institute on Global Commons and Climate Change (MCC) in Berlin, where he works on the FORMAS-funded Rivet project (with Lund University) on 'Risk, values, and decision-making in the economics of climate change'. He is also a PhD candidate in Philosophy of Science at the University of Cambridge and currently preparing to defend his thesis.

Cite this article: Beck L (2023). The Econ within or the Econ above? On the plausibility of preference purification. *Economics & Philosophy* **39**, 423–445. https://doi.org/10.1017/S0266267122000141