Studies in History and Philosophy of Science 117 (2026) 102131

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

The precision fallacy: On the futility of preference purification ☆

Johanna Thoma

Universität Bayreuth, Germany

ARTICLE INFO

Keywords:
Welfare economics
Expected utility theory
Behavioural economics
Precision
Vagueness
Preference purification

ABSTRACT

The standard theory of choice in economics involves modelling human agents as if they had precise attitudes when in fact they are often fuzzy. For the normative purposes of welfare economics, it might be thought that the imposition of a precise framework is nevertheless well justified: If we think the standard theory is normatively correct, and therefore that agents ought to be in this sense precise, then doesn't it follow that their true welfare can be measured precisely? I will argue that this thought, central to the preference purification project in behavioural welfare economics, commits a fallacy. The standard theory requires agents to adopt precise preferences; but neither the theory nor a fuzzy agent's initial attitudes may determine a particular way in which she ought to precisify them. So before actually having precisified her preferences, the welfare of fuzzy agents may remain indeterminate. I go on to consider the implications of this fallacy for welfare economics.

1. Introduction

Precision has traditionally been highly valued in science, and often been treated as a hallmark of scienticity. A lack of precision, and in particular a lack of precise quantification has sometimes been taken as a reason for thinking of the social and human sciences as *lesser* compared to the natural sciences.[1] In economics in particular, mathematical precision has been key in establishing its perceived scientific credentials. And there are some good reasons for thinking precision is important in science: More precise claims seem to be more easily testable, more informative, and hence also more useful in practical application.

But there are also compelling reasons for thinking that precision is sometimes a hindrance rather than a help in science. Levins (1966) argued that there is a tradeoff between precision, realism and generality in model-building in biology, so that sometimes, sacrificing precision for generality and realism could be called for. Elliott-Graves (2020) provides a case study highlighting that complexity and causal heterogeneity in particular contribute to clinching things in favour of more imprecise models. In anthropology, Sørensen (2016) argues that commitment to precision may keep us from even investigating those aspects of human life that cannot easily be studied precisely. And Neto (2020) stresses that sometimes, imprecise *concepts* can be useful in scientific inquiry.

In economics also, too strong a focus on precision has been criticised from two directions: From those who generally oppose its mathematisation as inappropriate for its subject matter (Lawson, 2003), but also from those who believe that quantitative models appropriate for the complexity of economics would be such that they do not allow for analytic solutions and definite results — the focus on easily tractable models in macroeconomics in particular has been blamed for some of the failings of economics in the run-up to the financial crisis (Col et al., 2009). As this latter criticism illustrates, while formalisation and quantification is often seen as a prerequisite for the kind of precision seen as desirable for science, quantitative or formal models can still be imprecise, for instance by not yielding definite results, or, as we will

Many of the authors criticising precision have stressed that imprecision is often called for not for epistemic reasons, but because of the nature of the subject matter being studied. Even in the case of physics, it has been argued that the real number precision of most measurement systems does not match the granularity of the physical facts (see Miller, 2020, Teller, 2018). More radically, in the human sciences, Larroulet Philippi (2024) discusses the long-standing 'quantity objection' according to which some of the attributes studied by human sciences, such as happiness, are such that they do not admit of quantification, and thus also not of the precision that quantification makes possible — for instance because of their multi-dimensional nature.

In this tradition, I will focus on how in economics, and more specifically in welfare economics, precision is at least sometimes a hindrance in modelling human agents because of what human agents are like. They are fuzzy in the sense that their attitudes are often vague or indeterminate. It is fairly straightforward to see how this could be a

☆ This article is part of a Special issue entitled: 'Measuring the Human' published in Studies in History and Philosophy of Science.

E-mail address: johanna.thoma@uni-bayreuth.de.

[1] For a useful overview of this position, see Elliott-Graves (2020), Orzack and Sober (1993) for a paradigmatic example in biology, and Rosenberg (1989) who argued that economics had been limited by making by and large only generic predictions. See Weintraub (2002) on the mathematisation of economics.

[2] Teller (2018) in fact argues that the excessive preciseness of measurement systems should more generally be understood as idealisation.

https://doi.org/10.1016/j.shpsa.2026.102131

Available online 10 March 2026
0039-3681/© 2026 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

problem for *descriptive* decision theory, which aims to capture how real agents make choices. Imposing a precise model might mean ascribing a determinate attitude to somebody who does not have one, and so the model will in fact end up *less* accurate than an imprecise one might be. This point appears to be acknowledged by many within economics (see, e.g. Woodford, 2020), though we might think the precision assumption is an appropriate idealisation for many descriptive and explanatory economic purposes.[2]

My main concern, however, is a more specific and less obvious problem, which occurs in the case of measurement of welfare for more normative purposes. In this kind of context, there appear to be good reasons to use a precise framework even for fuzzy agents. The notion of welfare that most economists work with is one that is deferential to people's own attitudes or choices — it captures their *subjective* interests. This is what I will take to be the intended measurand for the purposes of this paper. This notion of welfare expresses an anti-paternalist ideal, that economists or policy-makers should not be imposing their own conception of what is good for people on them. Traditionally, welfare is thus inferred from observed choices or expressed preferences. And it is done so assuming that agents abide by the axioms of expected utility theory (EUT), and furthermore exhibit context-independence and stability in their preferences.

Expected utility theory, as I will elaborate on in the following, assumes precision in agents' preferences: It requires preferences to be complete (agents are either indifferent between or have a strict preference between any two options), thereby ruling out indeterminacy, and transitive, which is often not the case for agents with vague desires. Context-independence rules out arbitrary shuffling over time. When agents have fuzzy – indeterminate or vague – attitudes manifested in their preferences, they may thus not be well described by the framework standardly used to estimate their welfare. This is, I argue, one (but not the only) reason we see many violations of EUT in practice.

For normative purposes, however, it might be thought that the imposition of a precise framework is nevertheless justified: If we think EUT is the correct normative theory, that is, if agents *ought to* abide by its axioms, and therefore ought to be in this sense precise, then imposing a precise framework may still be the best way to measure their welfare. We had merely be correcting for irrational mistakes that they are making, in fact measuring their *true* subjective interests, what is actually good for them by their own lights.

This line of argument, central to the research programme of preference purification in behavioural welfare economics, commits the fallacy that is my main concern here. From the supposition, which I will grant, that agents ought to have precise preferences in the way expected utility maximisers do, it does not follow that we can and should measure the welfare of real, fuzzy agents precisely. The argument fails because EUT just requires fuzzy agents to adopt precise preferences; but neither EUT nor the agents' initial fuzzy attitudes may determine a particular way in which they ought to precisify them. So before actually having precisified their preferences, the subjective interests of fuzzy agents may remain indeterminate. This makes the project of preference purification futile *in principle* whenever we have reason to think the agents we are studying are fuzzy, besides the more epistemic worries frequently voiced by critics. This fallacy can be avoided by those approaches in behavioural welfare economics that allow for indeterminate welfare rankings, such as Bernheim and Rangel's (2009). Alternatively, imposing precision on fuzzy agents may be justified on objective grounds. But this would require changing the measurand — that is, adopting a different, more objective concept of welfare and thereby giving up, at least in part, on the anti-paternalist ambitions of welfare economics.

After providing some background on behavioural welfare economics (Section 2), I will explore various ways in which violations of EUT may be due to fuzziness in agents' attitudes, and how in those contexts, we can think of the requirements of the standard framework as precisification requirements (Section 3). Section 4 then sets out the Precision Fallacy and what is problematic about it in greater detail. Section 5 looks at implications and concludes.

2. Background: Preference purification

Suppose you want to rent a flat. First consider the choice between two equally priced flats with the following main attributes:

Flat A: Small, short commute, far from nature, tasteful fixtures.

Flat B: Large, long commute, close to nature, tacky fixtures.

Size, length of commute, closeness to nature and tastefulness are all things that you care about, and so the decision is not trivial. If you choose Flat A one day, but then change your mind and go with Flat B the next, and your preferences are strict each day (you'd pay an extra euro to get your preferred choice), you would be displaying *unstable* preferences. If you choose Flat A in some contexts (e.g. if you have just had a really annoying commute), and Flat B in others (e.g. if the Estate Agent's is decorated with beautiful pictures of nature), and you at the same time judge these contexts to be substantively irrelevant for your choice,[3] you are displaying *context-dependent* preferences — indeed context-dependence might be what often explains instability in practice.

Now imagine you are comparing these flats (and you have no taste):

Flat C: Small, medium commute, close to nature.

Flat D: Medium size, short commute, far from nature.

Flat E: Large, long commute, medium distance to nature.

Comparing C and D, you (strictly) prefer D, because it is larger and has a shorter commute, and this makes up for the distance to nature. Comparing D and E, you (strictly) prefer E, because it is even larger, closer to nature, and this makes up for the long commute. And comparing E and C, you (strictly) prefer C, because it has a shorter commute, is closer to nature, and this makes up for its small size. These preferences are cyclical, and hence they violate transitivity, which makes them *inconsistent* according to EUT.

Orthodox welfare economics is built on a number of very neat results, showing that if agents' choice behaviours are stable, context-independent and consistent in a particular way, we can ascribe to them preferences that are themselves stable, context-independent and consistent with EUT.[4] Most importantly, preferences according to EUT are complete (agents are either indifferent between or have a strict preference between any two options) and transitive (weakly preferring a to b and b to c implies weakly preferring a to c). Under conditions of risk or uncertainty, preferences are such that we can model agents as maximising expected utility according to some utility function we ascribe to them together with an assumed or simultaneously derived probability function.[5]

The preferences and utilities ascribed to such stable and consistent agents are used in the orthodox framework as measures of people's welfare, which then feeds into the evaluation of policies or economic institutions. The idea is that this implements an anti-paternalist ideal, of deferring to people's own judgements and/or choices in matters concerning themselves – either because one thinks people's own attitudes constitute their wellbeing, or because one wants to respect their autonomy whether they are good judges of their wellbeing or not.

[3] This caveat is important. Sometimes, what we may call "context" are actually things that really do change your options in ways that you care about — the decor of a restaurant, for instance, may genuinely change your culinary experience. In that kind of case, your preferences are not context-dependent in the sense intended here. Rather, restaurant decor should be part of the description of your options.

[4] As shown by revealed preference theorems such as Houthakker's (1950) building on Samuelson (1938).

[5] As shown by EUT representation theorems such as von Neumann and Morgenstern's (1944), which works with assumed probabilities.

'Welfare' on the anti-paternalist picture is supposed to track what I will call 'subjective interest' – what people want or judge to be desirable insofar as they themselves are concerned. This is our measurand of concern.[6]

Important developments in recent years have been motivated by the fact that insights from behavioural economics have put this methodology under pressure: people often display instability or context-dependence in their choice behaviours, or violate the consistency axioms of expected utility theory — just like in our opening examples, which I take to be relatable. Behavioural welfare economics has grappled with how we should measure their welfare in such circumstances, and whether and how this can still be done in an anti-paternalist spirit.

The dominant approach, variously known as the "New Consensus" or "preference purification"[7] Sugden (2018) involves still ascribing to agents with inconsistent, unstable or context-dependent choice patterns preferences that are consistent, stable and context-independent.[8] Welfare is then taken to correspond to the consistent and stable preferences ascribed to the agents instead of the inconsistent preferences exhibited in their behaviour. Any divergence between actual choices and these preferences are treated as mistakes that could potentially be corrected through intervention.[9] Where interventions to correct these mistakes are supported, the idea is that these are merely "means paternalist" helping agents to better achieve their own goals (see also Thoma, 2024). The presupposition is that the consistent preferences ascribed to agents capture their *true* subjective interests, and that the welfare measure is thus still deferential to agents, in line with the original anti-paternalist ambition of welfare economics. The most sophisticated methods within this research programme involve explicitly modelling agents in a way that allows us to identify and correct for distorting factors that might make agents violate standard EUT (e.g. Bleichrodt et al., 2001), while others merely make plausible presuppositions about the agents' true subjective interests (e.g. Sunstein & Thaler, 2003).

Infante et al. (2016) criticise preference purification for making the implausible assumption that there is an "inner rational agent" within us, even when we act inconsistently, which is merely distorted by an outer psychological shell. The implausibility of the idea of an inner rational agent can be highlighted using our earlier examples. Preference purification would need to involve ascribing a determinate and stable preference (which could be indifference) between Flats A and B, and transitive preferences over Flats C, D, and E. On the inner rational agent model, that would mean that at least some of the context-dependent or cyclical preferences displayed by the real agent in our examples are cases of choice against the agent's own better judgement — that is, cases of weakness of will. This may perhaps sometimes be what is going on in such cases, but clearly not always.

The more plausible way to think of preference purification is to consider the purified preferences not as preferences somehow already within the agent, but as those that agents *should* have, and indeed *would* have if they were rational. But we need not see them as already instantiated in some inner rational agent. How would such preferences still count as subjective, if they are not actually the agent's? We must think of them as somehow connected in the right way to the inconsistent agent's actual initial attitudes.[10] The most plausible way to think of this – which is also independently attractive (see also Thoma, 2021) – is to imagine agents having an underlying 'subjective motivational set' in Williams's (1979) sense, including desires, commitments, wants, etc which constitute her subjective interests. We can then think of preferences as a relation that ideally tracks what the subjective motivational set as a whole supports. This way, we can make sense of preferences being mistaken by the agent's own lights, and of the idea that there are alternative preferences the agent ought to have, which correctly track her subjective interests. Those who accept the standard framework as normatively correct think that the EUT axioms, coupled with context-independence and stability are constraints on the preferences agents ought to have, in addition to them tracking the agent's underlying desires. If this is so, you might think, the correct way to measure the welfare even of inconsistent agents is by stable EUT preferences.

Critiques of this picture have largely been either normative or epistemic in kind. Normative criticism either comes in the form of the idea that even mere means paternalism is problematic (see, e.g. Camerer et al., 2003). Or it appeals to critiques of expected utility theory as the correct theory of rationality (such as, e.g. Buchak's 2013). Epistemic criticism points to inadequacy in the methods available to us for identifying purified preferences (see, e.g., Rizzo & Whitman, 2009). What I want to argue is that in many circumstances, and in particular when agents are fuzzy, the project of preference purification is in principle futile — before we even encounter epistemic obstacles, and even if we grant that preference purifiers are assuming the correct normative theory. Preference purifiers commit what I will call a "Precision Fallacy".

3. Fuzziness and violations of expected utility theory

Let me start by exploring how violations of EUT as well as of context-independence and stability might be linked to agent fuzziness. I will look at two general kinds of fuzzy sources of violations: Vagueness in individual desires that are part of an agent's relevant subjective interests; and indeterminacy in how desires are to be combined and weighed up when making complex choices. I will then consider indeterminacy in risk attitudes as a special case. In each case, we can think of the standard framework as imposing a precisification requirement on agents.

3.1. Vague desires

Consider, first, your desire for tastefulness in our flat choice example earlier. Note that "tasteful" is a vague predicate. The most uncontroversial criterion for whether a predicate is vague is whether it allows for borderline cases (Sorensen, 2023), that is, cases where it is unclear whether something counts as tasteful or not. And the existence of borderline cases is usually thought to make very plausible *tolerance*, meaning that small enough changes can never make a difference to whether the predicate applies or not. In our case, a plausible tolerance principle is that a small enough change in a flat's fixtures – like replacing one ornate golden door handle – can never turn it from non-tasteful to tasteful.

[6] Choices do not track subjective interest thus understood when people act altruistically. For simplicity, and also because behavioural welfare economics has focused on these, I will consider only examples of choices affecting only the agent.

[7] Not to be confused with purification in game theory (Morris, 2008).

[8] I do not count as part of the preference purification project approaches that make welfare ascriptions weakening the assumptions of the standard framework, for instance those that allow for incompleteness. As we will see, adopting such a weaker framework is one potential solution to the problem I will lay out.

[9] See, e.g. Beshears et al. (2008) for a defence of this idea, and Sugden (2018) and Rizzo and Whitman (2020) for critique.

[10] The picture I have in mind here is similar to idealised desire theories of wellbeing in philosophy, as defended, e.g., by Railton (1986). There is a related debate in this literature about whether idealised desire theories can still be appropriately subjective (see Griffin, 1986, Rosati, 1995 , Heathwood, 2006). Some of the problems discussed in this literature concern complications from idealising for complete information, which is an orthogonal issue to our concerns here. More generally, the means paternalist welfare economist aims to only correct for failures of instrumental rationality, that is, failures to take the best means to one's ends. If she really does correctly identify the agent's actual ends and the best means to pursue them – a big if, as we shall see – this is less problematically still a subjectivist project.

Now imagine that you are comparing flats that differ in their fixtures, but are perfectly equivalent in everything else you care about. And your desire for tastefulness is binary: You want your flat to be tasteful, but you make no further distinction between tasteful flats. Preferences that track your desire would then be as follows: Strictly prefer tasteful to non-tasteful flats, and be indifferent between flats that are either both tasteful or both non-tasteful. Since desires that strictly track vague predicates will inherit borderline cases of fulfilment from the vagueness of the predicates, I will refer to these desires as themselves vague.

In our case, if tolerance holds, preferences that track the vague predicate of tastefulness may end up violating transitivity. Imagine an ordered series of flats that have tiny differences in their fixtures, where the first is clearly tacky, and the last is clearly tasteful, but differences between them are so small that you would either always count two adjacent ones as both tasteful, or as both non-tasteful. You'd then be always indifferent between two adjacent ones, but also strictly prefer the last one to the first one.[11]

Granted, some accounts of vagueness end up denying tolerance for vague predicates. But they also all explain at least why tolerance *seems* very attractive, and why we cannot *detect* sharp boundaries for vague predicates.[12] This is enough to make the intransitive preferences just described very natural for agents with desires that track vague predicates. But this example of a binary desire is admittedly very artificial. Things we are choosing between usually differ along a number of different dimensions, and most of our desires admit of degrees of fulfilment. Still, even attitudes that come in degrees are often, in Andreou's (2022) terms "categorial", classifying things in ordered categories ranging, e.g., from very tacky to tacky to neutral to tasteful to very tasteful, where the borders between categories are again plausibly vague. And adding further dimensions to your choice does not do away with the possibility of intransitivity. For instance, small enough differences in fixtures might lead you to focus entirely on other dimensions to your choice. But then a series of comparisons could lead you to a clearly tacky option, with the tackiness in fact outweighing the other desirable features of the flat.[13]

Now EUT requires agents to have transitive preferences. In cases where intransitivity is due to vague desires in the sense I described, we can think of this as a precisification requirement. For instance, in the ordered series of tacky to tasteful flats, settling on a unique transitive preference relation requires agents to at some point have a strict preference between adjacent flats where the underlying attitudes the agent starts out with actually make no distinction — where both flats are judged non-tasteful, or both are judged tasteful, and all else is equal.

There are two different ways in which we can think of this precisification requirement: We could either say that there was something wrong with the agent's original underlying desires: Rationality requires people not to have vague desires, and thus to refine their tastes. Desires should, in particular, not track vague predicates like "tastefulness", but rather, for instance, some precise operationalisation of the concept that always yields clear verdicts. Or we could say that there is nothing wrong per se with having vague desires, but that rationality requires agents to have 'precise' preferences – ones that abide by all the assumptions of the standard framework – even when their underlying desires are vague. This option is open to us if we think of preferences, for instance, as mere behavioural dispositions (see Thoma, 2020), or as attitudes we form for the purposes of choice (see Andreou, 2007). In that case we would think of the required strict preference between at least one set of adjacent flats as an arbitrary pick. In both cases, we can think of preferences as still responsive to our desires, but also as potentially displaying regularity, precision or consistency beyond what those desires warrant.

My main point does not depend on either one of these interpretations of EUT's precisification requirement, and despite their differences, I will refer to them together: Precisification in either case involves settling on a unique transitive preference relation despite at least initially vague desires. But note that the second interpretation fits better with the merely means paternalist ambitions of welfare economics. For on the first interpretation, agents with intransitive preferences stemming from vague desires are irrational not in virtue of being bad at pursuing their own interests. They are irrational for having the wrong desires; and correcting that kind of mistake is ordinary paternalism, not merely means paternalism.

I will grant that EUT in fact is the correct theory of rationality, and thus that rationality requires the kind of precisification I described. The most persuasive case for this, in my view, are various pragmatic arguments that establish that agents who violate EUT's axioms are prone to making sure losses in dynamic choice situations, such as the money-pump argument (Davidson et al., 1955).[14] This kind of argument goes particularly well with the second interpretation of precisification. Being instrumentally rational seems to require the preferences that guide or constitute one's choices to be responsive to one's desires or ends, so that one acts so as to fulfil them. But beyond that it also seems to require preferences to be consistent if this is necessary in order to avoid sure losses in relation to those very desires or ends. In our flat choice examples, for instance, instrumental rationality requires that agents end up with a flat that on balance serves their interests, but also to avoid the kinds of sure losses that are involved in constant shuffling, or in being money-pumped by an estate agent who takes a fee every time you change your choice. If we accept this kind of argument, we accept that rationality requires precisification of preferences even for agents who have and continue to have vague desires.

3.2. Indeterminacy in aggregation

Looking at the other flat search-related desires I mentioned above, they might in fact well be precise if taken on their own: You may like a flat more to the precise extent to which it is close to the nearest bit of open landscape, or close to work, or in precise proportion to its square footage, assuming distance and area are all precise properties. Nevertheless, instability, context-dependence and intransitivity seemed like relatable phenomena in our examples. This may be due to a further kind of fuzziness in your attitudes, namely indeterminacy in how your various desires are to be combined in order to inform choice. How exactly do you trade off size against the daily annoyance of commuting, or against the experience of nature? You may not have an answer to this question precise enough to give you a unique and complete ordering of flats according to these characteristics. And in practice, of course, many more things are relevant to your decision.

Perhaps you will employ some simple rule for weighing up pros and cons to help you out. But in complex multi-attribute decision problems, simple rules of aggregation may lead you into intransitivity. This is illustrated by the choice between flats C, D, and E: If you just count

[11] This example mirrors one introduced by Aldred (2007) involving a desire for an ancient commons to be unspoilt, which he calls the 'tragedy of the disappearing commons'.

[12] On the epistemicist picture, for instance, vague predicates have sharp boundaries, but it is impossible to know where they are (see Williamson, 1994). And on the contextualist picture, there are sharp boundaries, but they vary with context such that they never are where we are actually looking (see Raffman, 1994).

[13] Quinn's (1990) self-torturer problem is arguably an especially stark example of how intransitivity can arise in a multidimensional case involving a vague desire.

[14] For a comprehensive account of these arguments, see Gustafsson (2022). Thoma (forthcoming) provides a critique but also argues a conditional version of these arguments is promising.

the number of advantages vs disadvantages when comparing any two flats and prefer the flat with more advantages, you end up with the cyclical preferences in our example. Whatever method of aggregation you use — if you want consistency with EUT, it can't be that. And indeed, most empirical evidence of intransitive preferences concerns multi-dimensional decision problems, showing that people really do find it hard to make tradeoffs in a way that is consistent with EUT.[15]

Note that it does not help to point out that, when an agent finds it hard to weigh up many different considerations when making a choice, they could simply be thought of as indifferent between the options. EUT does of course allow indifference. Suppose an agent is in fact indifferent between flats C and D. And suppose there are lots of flats available in the same building as D in a large variety of sizes. The largest one of these is clearly preferable to C. EUT, along with the assumptions we made about the relevant desires would require us to be able to identify the precise size at which a flat in D's building becomes preferable to flat C. The kind of indeterminacy I am interested in is one whereby this is simply not settled by the agent's underlying attitudes — neither by the internal strength of the relevant desires for size or commute length, nor by some meta-attitude that compares their importance. For an agent who has imprecise attitudes in this sense, there will be a range of flats in D's block for which the agent's attitudes do not determine whether the agent should strictly prefer them to C or should be indifferent.[16]

Such a situation can easily lead to a violation of EUT axioms or otherwise to instability or context-dependence. One way in which such indeterminacy about aggregation may express itself is in incomplete preferences — the agent could be neither indifferent nor have a strict preference between some pairs of options. This would be a straight-out violation of EUT axioms, since these axioms include completeness. Or, she could form complete preferences. But in that case, there would be some arbitrariness involved in the preferences she forms. And this arbitrariness could give rise to further inconsistency, context-dependence or instability. We could, for instance, think of the simple aggregation rule leading to intransitivity over C, D, and E as a way of settling on preferences in the context of arbitrariness — in which case we end up with a violation of EUT axioms. Or suppose that flats A and B above are such that the agent's attitudes leave it indeterminate whether she should strictly prefer A to B, B to A or be indifferent. Her settling on one of these preference relations will then be arbitrary. And then what she does settle on may well be triggered by the context, or simply just change over time.[17] As far as the pursuit of her subjective interests is concerned, none of these choices would individually be mistakes. But of course she would be violating the context-independence and stability requirements of the standard framework.

Again we can think of the EUT consistency requirements along with context-independence and stability as requirements of precisification in these kinds of circumstances. Jointly, they require agents to settle on a preference relation that abides by the EUT axioms and that is unique, not changing with context or over time (at least not without good reason). This would again require precision beyond what the agent's initial attitudes warrant — her initial underlying attitudes in fact do not determine such a unique preference relation. And again we can interpret this in one of two ways: Either, the imprecision in her initial attitudes is irrational, and she actually needs to make her underlying attitudes more precise. That is, she needs to adopt a meta-attitude where there was none, or she needs to change her meta-attitude into one that is not vague. Or, she does not change her underlying desires, and the requirement is to only precisify preferences beyond the precision of the underlying attitudes. Here, this would mean settling arbitrarily on a unique and consistent preference relation, and sticking with it across contexts and over time. Again, this is the interpretation that squares better with means paternalist ambitions, as it does not identify the agent's ends (as picked out by her underlying attitudes) as mistaken. And precisification in this sense could again be argued for by appeal to pragmatic arguments: Not only preferences inconsistent with EUT, but also instability and context-dependence make agents vulnerable to making a sure loss.

Table 1
Allais Paradox: First Choice.

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery A	€250 million	€1.25 billion	€0
Lottery B	€250 million	€250 million	€250 million

Table 2
Allais Paradox: Second Choice.

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery C	€0	€1.25 billion	€0
Lottery D	€0	€250 million	€250 million

3.3. Imprecision in risk attitude

Many of the most recalcitrant violations of EUT involve choice in the context of risk and uncertainty. These violations concern not transitivity, but the axioms crucial for the expectational form of the framework under risk and uncertainty — e.g. the independence axiom in the case of von Neumann and Morgenstern's (1944) framework. Take the famous Allais Paradox (Allais, 1953), keeping in mind there is also solid evidence for systematic violation of EUT in practice and in more common-place decision contexts (see Harrison & Swarthout, 2023). Adjusting the original monetary values for inflation, Allais asks us to consider our preferences between lotteries in two choice scenarios. In each case, our prize will be determined by a fair draw from 100 tickets. In the first choice scenario, we choose between a certain €250 million (Lottery B) and, effectively, an 89% chance of €250 million, a 10% chance of €1.25 billion, and a 1% chance of nothing (Lottery A). In the second choice scenario, we choose between a 10% chance of €1.25 billion, and an 11% chance of €250 million (see Tables 1 and 2).

Allais hypothesised, and it has since been supported experimentally,[18] that people often prefer Lottery B in the first choice scenario, and Lottery C in the second. Yet, this combination of preferences is incompatible with EUT. Provided money is all we care about in these choices, EUT requires us to prefer B in the first choice if and only if one prefers D in the second choice. Essentially this is because EUT implies that how two lotteries compare depends only on those states of the world where they lead to different outcomes. So all that matters, in both the first and second choice, is what happens when tickets 90-100 are drawn. But then the first and the second choice come down to the same question: How do a 10% chance of €1.25 billion, and an 11% chance of €250 million compare?

As with other requirements of EUT, there are pragmatic arguments for abiding by this kind of separability condition, showing that agents who violate it are prone to making a sure loss in certain dynamic choice situations.[19] To be consistent with EUT, agents who originally display the "paradoxical" preferences now have two options: They

[15] See Tversky (1969) for an early example.

[16] You may want to distinguish two kinds of sources of indeterminacy here: There could simply be an absence of a meta-attitude allowing us to make the relevant trade-offs. Or the meta-attitude could be a vague attitude in the sense we just described — one that allows for borderline cases, for instance by inheriting them from the property it is tracking.

[17] Indeed, in the decision sciences, preferences are often thought of as constructed on the spot, see, e.g. Lichtenstein and Slovic (2006), as also pointed out by critics of the dominant framework in behavioural welfare economics like Sugden (2018).

[18] See Blavatskyy et al. (2022) for a recent overview.

[19] See, e.g. Machina (1989) for a dynamic version of the Allais Paradox.

could keep their preference in the first choice scenario, and change their preference in the second, or they could keep their preference in the second scenario, and change their preference in the first. As with the previous types of cases, the point I wish to make is that it could be – though it need not be – that there is nothing in the attitudes the agent starts out with that determines which way she should go.

It is sometimes thought that EUT is in an important sense risk neutral, in which case we would not get the kind of indeterminacy I have in mind here. Or at least, we would not get it as long as the agent's desire for money is precise, in that it not only ranks all monetary outcomes, but also determines differences in value between different outcomes. The sense in which EUT is sometimes thought to be risk neutral is that it requires that agents should pursue their subjective interests – in this case degrees of monetary desire satisfaction – in a risk neutral way, by maximising expected degrees of desire satisfaction. In that case, knowing about how an agent evaluates outcomes together with the probabilities determines how lotteries should be ranked. For instance, if an agent values money linearly (her desire for each euro is exactly the same, no matter how much she already has), risk neutrality would require choosing Lotteries A and C in the Allais scenarios.

But in fact the standard EUT axioms are consistent with a non-neutral pursuit of subjective interest (see Dyer & Sarin, 1982). For instance, an agent who prefers B and D in the Allais choices while having a linear desire for money would not thereby be violating EUT axioms, and would at the same time display risk aversion in her pursuit of monetary desire satisfaction. In that kind of case, we speak of non-neutral 'pure' attitudes to risk: The choices are explained in terms of the agent's aversion to taking risks, not in terms of her valuing money less the more she already has of it.

In practice, agents often display non-neutral pure attitudes to risk in a way that does violate the EUT axioms – like the typical Allais preferences do – and thus cannot be accommodated within EUT. They are then often better captured by alternative theories of choice under risk and uncertainty, such as cumulative prospect theory (Tversky & Kahneman, 1992) or rank-dependent expected utility theory (Quiggin, 1982). One finding that motivates prospect theory, for instance, involves agents being risk seeking for losses while being risk averse for gains, against a context-dependent status quo point. This is something that EUT cannot accommodate, even though it can accommodate some non-neutral attitudes to risk.

Still, once we allow that non-neutral pure attitudes *can* be rational, we have created space for the kind of indeterminacy we are interested in again. If an agent has no settled pure attitude to risk of the type that fits neatly into an EUT framework to start with, there might be nothing in her existing attitudes that determines in which way she should make herself consistent with EUT. Agents who have the "paradoxical" preferences in the Allais example, for instance, have open to them a less and a more risk averse way of resolving their inconsistency. My hypothesis is that there is often no psychological fact about them that settles which one more truly reflects their attitude to risk.[20] EUT asks them to precisify — to settle, in one way or another, on a consistent, context-independent and stable preference relation despite the (initially) fuzzy underlying attitudes. But neither the theory, nor their existing attitudes settle how.

What I have argued, then, is that there are various ways in which fuzziness in agents' attitudes can lead them to violate EUT axioms or the further requirements of the standard framework of context-independence and stability. This is not to say that violations of the standard framework always stem from such fuzziness.[21] But when they do, we can understand these requirements as requirements of precisification. And, importantly for what I want to argue in what follows, in such cases, neither EUT, nor the agent's original (fuzzy) attitudes settle *how* she should precisify.

[20] There are again a variety of potential sources of this: A simple absence of a relevant attitude to risk, an existing attitude to risk, but one that is irrational in that it will inevitably lead to violations of independence, or the presence of an attitude to risk that is vague by allowing for borderline cases.

4. The precision fallacy

In Section 2 we saw that the research project of preference purification involves ascribing to agents and using as a welfare standard preferences that abide by EUT axioms as well as being context-independent and stable — despite agents violating these requirements in practice. One initially plausible way of justifying this is to point out that if EUT is the normatively correct theory, those are the kinds of preferences agents ought to have. And so it might seem apt to use such models for the normative purposes of welfare economics, even if less restrictive models may be needed for more descriptive purposes.

However, this commits what I want to call a "Precision Fallacy". The last section explored how violations of the standard framework may arise from fuzziness in people's attitudes, and how in those circumstances, we can think of the requirements of this framework as requirements of precisification. The Precision Fallacy involves concluding from the fact that agents ought to precisify their preferences that a precise framework is appropriate for measuring their welfare.[22] The reason why this is a fallacy comes down to the possibility of a kind of *non-uniqueness*: There may not be one uniquely permissible way for an agent to precisify her preferences, but several.[23]

The last section explored cases where precisely such non-uniqueness holds. When underlying desires are vague, there is indeterminacy in how to trade-off various different considerations, or when agents do not have a settled well-behaved attitude to risk, neither EUT, nor their existing attitudes determine exactly how they should precisify their preferences; EUT only settles *that* they should do so. In that sense, its requirements are *structural*.[24]

When non-uniqueness holds, the problem faced by preference purifiers is that before the agent has actually precisified her preferences, we have no basis for ascribing any particular stable EUT preference relation to her. Thus, it seems, there is also no basis for using such a relation as a measure of her welfare, where welfare is meant to capture

[21] Cases where violations do not stem from fuzziness are beyond the scope of this paper. The easier such cases for the means paternalist welfare economist to handle are ones where preferences are simply mistaken representations of the agent's true subjective interests. This seems to be what most preference purifiers have in mind, especially when they subscribe to the idea of an inner rational agent. The hard cases are ones where desires are not vague, and they can only be faithfully captured by an inconsistent preference relation. Such cases force the welfare economist to either adopt ends paternalism and declare the desires as irrational, or to give up on the consistency requirement.

[22] Note that if there were an "inner rational agent" a precise framework would thereby be appropriate. So those who believe in an inner rational agent for independent reasons need not commit the Precision Fallacy as I outline it here. But for one, inner rational agents are implausible for the reasons outlined above. And moreover, insofar as the existence of an inner rational agent is sometimes simply inferred from the idea that rationality requires consistency, belief in an inner rational agent is in fact a symptom of the Precision Fallacy rather than a way out from it.

[23] See also Thoma (2021) on this kind of non-uniqueness thesis. Also note that this uniqueness thesis is similar to discussions of non-uniqueness in epistemology. There, uniqueness is taken to refer to the idea that there is only one rationally permitted credal state in response to any particular body of evidence, and so non-uniqueness would involve there being several permitted responses to at least some bodies of evidence (see Titelbaum and Kopec, 2016). Non-uniqueness in our context involves there being more than one rationally permissible preference relation in response to at least some sets of underlying conative attitudes an agent might have. However, note that the plausibility of the uniqueness thesis in epistemology is at least partly due to he fact that its proponents typically allow for suspension of belief as one possible required credal state. The equivalent to that in our context would be incomplete preference, which EUT explicitly excludes. Uniqueness is a much less plausible thesis once we grant the requirements of EUT.

[24] See Beck (2023) for an appeal to structural rationality in the context of behavioural welfare economics.

her subjective interests. If you are a fuzzy agent, your welfare remains indeterminate within some ranges.[25] While the behavioural economist could simply ask agents to precisify, this would obviate the need for preference purification models, and be impractical for large-scale applications. And in fact, depending on what kind of precisification we think EUT demands, one could argue that indeterminacy in welfare remains even after an agent has rationally precisified. If precisification involves actually refining one's underlying attitudes, this does make precise what one's subjective interests are. But if it merely involves arbitrarily settling on a stable and consistent preference relation for the purposes of choice, while retaining desires and ends that remain fuzzy, the case for saying that your welfare can be indeterminate remains.

The welfare measures we get out of the preference purification project are thus in danger of being more precise than is warranted by the subjective attitudes of the agents that welfare economics aims to be deferential to. The problem with this is two-fold. For one, it means these measures are not faithful representations of what they want to represent. What they aim to represent is agents' true subjective interests. But in the context of fuzziness in attitudes and the resulting non-uniqueness, true subjective interests may remain indeterminate. This part of the problem resembles other scientific contexts in which philosophers of science have argued measures are more precise than the subject matter warrants.[26] But the other dimension of the problem is the moral and political implications of employing these measures in practice. The welfare measures we get from preference purification are intended to be used for policy evaluation or for means paternalist intervention. They may inform how policy-makers take into account the interests of inconsistent agents; or they may inform measures to nudge people or otherwise get them to correct their inconsistencies to make them better off 'by their own lights'.

In the case of means paternalist intervention, overly precise welfare measures lead to the danger of overcorrection, of intervening where agents are not actually acting against their subjective interests.[27] Given most means paternalist interventions come at least with a small cost in terms of interference with autonomy (aside from the cost of hiring a behavioural economist...), this is problematic. For instance, shuffling back and forth between flats A and B when it is in fact indeterminate which the agent should prefer, and the shuffling itself does not come with any costs to the agent, does not contravene their interests. Yet, a preference purifier may identify one of the flats as the one the agent should prefer, and may propose interventions to make them choose it and stick with it. Granted, if the standard framework is normatively correct, agents are irrational if they shuffle, and ought to precisify their preferences. But it is not clear that standard behavioural interventions can really help them to do so. In any case, in practice, interventions are usually geared at specific types of choices that are branded as 'mistakes' – like choosing sugary snacks when presented to you at eye-level at the checkout till. They are not targeted at helping individuals achieve structural rationality.

This last example also illustrates a second kind of more general problem with overly precise welfare measures. If a more faithful welfare measure would allow for indeterminacy, on what basis do economists, and then by extension policy-makers precisify, as it were, on behalf of the population? The choice is unlikely to be arbitrary, as there is an invitation here for inserting the values of the analyst or policy-maker. The following two examples are suggestive of this actually happening. First, in classic discussions of nudging in the context of consumer choice, like Thaler and Sunstein's (2008), there tends to be a presupposition that consumers really prefer the more healthy of two options, and that unhealthy choices are a mistake. No explicit preference purification model is proposed to support this claim. But this illustrates a willingness to err on the side of what, from a full, ends paternalist perspective might be taken to be the best option — even in cases where people's actual interests allow for both a more healthy and a less healthy kind of precisification. Second, as I explore in more detail in Thoma (2024), more formally sophisticated methods for preference purification in the context of risky choice, namely methods that employ prospect theory to 'debias' preferences (Bleichrodt et al., 2001), involve imposing risk neutrality on people, reflecting a bias towards this way of evaluating risk.

There may of course be good reasons for erring on the side of health or risk neutrality,[28] but what should be acknowledged is that when this is done, we are no longer just being deferential to people's subjective interests. That is, we are changing the measurand, and in so doing we are giving up, at least in part, on the anti-paternalist ambitions of welfare economics. Economists' or policy-makers' values are entering the picture to a greater extent. Committing the precision fallacy leads economists to not acknowledge this, to think that their precise measures of welfare are in fact uniquely accurate representations of what they set out to measure, namely subjective interest. All in all, then, this is an especially stark example of how choice of representation – in this case, the imposition of precision – has an important moral or political dimension.[29]

5. Implications and conclusion

Preference purification, the dominant research programme in behavioural welfare economics commits what I have called the Precision Fallacy. It concludes from the fact that, if EUT is the correct normative theory agents ought to have precise preferences, that we can measure their welfare precisely, and that this welfare measure would still be subjective in line with the means paternalist ambitions of welfare economics. But the argument fails for fuzzy agents: While EUT requires them to precisify their preferences, neither EUT nor their existing attitudes may settle how they should do so. Their subjective interests could thus be indeterminate. The welfare measures employed by preference purifiers are thus more precise than a faithful representation of subjective interest would be. Understood as the project of generating a unique and consistent welfare measure that faithfully represents subjective interest, preference purification is futile in principle for fuzzy agents. This is not only scientifically problematic, but also has moral and political implications.

I see two broad options for the measurement of welfare coming from this. One option is to adopt alternative frameworks that do accommodate fuzziness and yield indeterminate welfare rankings. A case in point in the economics literature is the framework proposed by Bernheim and Rangel (2009).[30] The framework yields indeterminate welfare rankings whenever observed preference is context-dependent or unstable without there having been a clearly identifiable mistake (like having a relevant false belief). It does yield determinate welfare rankings when agents consistently display the same preference over contexts. The advantage of this kind of framework is that, insofar as context-dependence is due to fuzziness, it yields welfare measures

[25] In that case, non-uniqueness of the normative type I described (there is not a unique preference relation the agent ought to adopt) results in non-uniqueness regarding measurement (as described, e.g. by Grégis, 2023) — there is no one unique true welfare value for the agent.

[26] See Grégis (2023), Miller (2020), Teller (2018).

[27] For similar reasons, Jackson (2020) argues that epistemic permissivism limits the range of cases where epistemic paternalism may be warranted.

[28] In fact, drawing on Hausman (2016), Beck (2023) appeals to such additional reasons as a response to worries about indeterminacy when requirements of rationality are structural in kind.

[29] See also Gould (1981) and Harvard and Winsberg (2022).

[30] In the philosophical formal epistemology literature, many authors represent credal states with families of probability functions, and sometimes extend their frameworks to families of utility functions (see Bradley, 2019). Both may yield families of preference relations that disagree on some judgements, which may be interpreted in terms of indeterminacy (see Levi, 1985).

that are more faithful to what subjective interest is actually like. It thus allows us to stick with the measurand welfare economics has traditionally been focused on. It is consequently also more true to the merely means paternalist ambitions of welfare economics than overly precise measures are. At the same time, Bernheim (2016) argues that the framework can still be used to inform policy. It will be more conservative in proposing interventions because it does not license intervening on every case of context-dependence. But it does still yield recommendations in cases where subjective interests are clear, as in the case of pensions plan defaults (Bernheim et al., 2015). Still, unfamiliarity by practitioners is an obstacle at least for now to its more wide-spread use.

The other option involves changing the measurand to a notion of welfare that is not purely subjective. We could then keep on using a precise framework, and appeal to something other than subjective interest to precisify on behalf of agents. In one way or another, this is inevitably already done whenever a precise framework is used to measure the welfare of fuzzy agents. But the rhetoric by behavioural welfare economists tends to remain one of mere means paternalism, of not imposing ends on agents, merely helping them pursue their own ends better. This obfuscates the way in which, for instance, welfare measures may in practice err on the side of health or risk neutrality. If we have less stark anti-paternalist leanings, we may in fact see agent fuzziness as an opportunity to bring in pro-social or objective values in the choice of precisification, and thereby to at least do so without directly contravening subjective interests. But if this is done, it should be done in a conscious and transparent way, and one that is subject to democratic control. In any case, this would change what it is welfare economists are measuring, and give up on at least some of the merely means paternalist ambitions of welfare economics. And so the choice between these two general options is at least in part a moral and political one.

CRediT authorship contribution statement

Johanna Thoma: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

No data was used for the research described in the article.

References

Aldred, J. (2007). Intransitivity and vague preferences. *The Journal of Ethics*, *11*(4), 377–403.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole americaine. *Econometrica*, *21*(4), 503–546.

Andreou, C. (2007). There are preferences and then there are preferences. In B. Montero, & M. D. White (Eds.), *Economics and the mind*. Routledge.

Andreou, C. (2022). *Choosing well: the good, the bad, and the trivial*. Oxford University Press.

Beck, L. (2023). The econ within or the econ above? On the plausibility of preference purification. *Economics and Philosophy*, *39*(3), 423–445.

Bernheim, D. (2016). The good, the bad, and the ugly: A unified approach to behavioural welfare economics. *Journal of Benefit-Cost Analysis*, *7*(1), 12–68. http://dx.doi.org/10.1017/bca.2016.5.

Bernheim, D., Fradkin, A., & Popov, I. (2015). The welfare economics of default options in 401(k) plans. *American Economic Review*, *105*(9), 2798–2837.

Bernheim, D., & Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, *124*, 51–104.

Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, *92*, 1787–1794.

Blavatskyy, P., Ortmann, A., & Panchenko, V. (2022). On the experimental robustness of the allais paradox. *American Economic Journal: Microeconomics*, *14*(1), 143–163.

Bleichrodt, H., Pinto-Prades, J.-L., & Wakker, P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility theory. *Management Science*, *47*, 1498–1514. http://dx.doi.org/10.1287/mnsc.47.11.1498.10248.

Bradley, S. (2019). Imprecise probabilities. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Buchak, L. (2013). *Risk and rationality*. Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780199672165.001.0001.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review*, *151*, 1211–1254. http://dx.doi.org/10.2307/3312889.

Colander, D., Goldberg, M., Haas, A., Juselius, K., Kirman, A., Lux, T., & Sloth, B. (2009). The financial crisis and the systemic failure of the economics profession. *Critical Review: A Journal of Politics and Society*, *21*(2–3), 249–267.

Davidson, D., McKinsey, J. C. C., & Suppes, P. (1955). Outlines of a formal theory of value, I. *Philosophy of Science*, *22*, 140–160.

Dyer, J. S., & Sarin, R. K. (1982). Relative risk aversion. *Management Science*, *28*(8), 875–886. http://dx.doi.org/10.1287/mnsc.28.8.875.

Elliott-Graves, A. (2020). The value of imprecise prediction. *Philosophy, Theory, and Practice in Biology*, *12*(4).

Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton & Company.

Grégis, F. (2023). Do quantities have unique true values? The problem of non-uniqueness in measurement. *Measurement*, *221*.

Griffin, J. (1986). *Well-being: its meaning, measurement, and moral importance*. Clarendon Press.

Gustafsson, J. E. (2022). *Money-pump arguments*. Cambridge University Press.

Harrison, G. W., & Swarthout, J. T. (2023). Cumulative prospect theory in the laboratory: A reconsideration. In G. W. Harrison, & D. Ross (Eds.), *Models of risk preferences: descriptive and normative challenges*. Emerald.

Harvard, S., & Winsberg, E. (2022). The epistemic risk in representation. *Kennedy Institute of Ethics Journal*, *32*(1), 1–31.

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

Heathwood, C. (2006). Desire satisfactionism and hedonism. *Philosophical Studies*, *128*(3), 539–563.

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economia*, *17*, 159–174.

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*(1), 1–25. http://dx.doi.org/10.1080/1350178X.2015.1070527.

Jackson, E. (2020). Epistemic paternalism, epistemic permissivism, and standpoint epistemology. In A. Bernal, & G. Axtell (Eds.), *Epistemic paternalism reconsidered: conceptions, justifications and implications*. Rowman and Littlefield.

Larroulet Philippi, C. (2024). Is quantitative measurement in the human sciences doomed? On the quantity objection. *British Journal for the Philosophy of Science*.

Lawson, T. (2003). *Reorienting economics*. Routledge.

Levi, I. (1985). Imprecision and indeterminacy in probability judgment. *Philosophy of Science*, *52*(3), 390–409.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431.

Lichtenstein, S., & Slovic, P. (2006). The construction of preference: An overview. In S. Lichtenstein, & P. Slovic (Eds.), *The construction of preference* (pp. 1–40). Cambridge University Press.

Machina, M. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, *27*(4), 1622–1668.

Miller, M. E. (2020). Worldly imprecision. *Philosophical Studies*, *178*, 2895–2911.

Morris, S. (2008). Purification. In S. N. Durlauf, & L. E. Blume (Eds.), *New Palgrave Dictionary of Economcis* (2 ed.). Palgrave.

Neto, C. (2020). When imprecision is a good thing, or how imprecise concepts facilitate integration in biology. *Biology and Philosophy*, *35*(6), 1–21.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Orzack, S., & Sober, E. (1993). A critical assessment of levins's the strategy of model building in population biology (1966). *The Quarterly Review of Biology*, *68*(4), 533–546.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, *3*(4), 323–343. http://dx.doi.org/10.1016/0167–2681(82)90008–7.

Quinn, W. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*(1).

Raffman, D. (1994). Vagueness without paradox. *Philosophical Review*, *103*(1), 41–74.

Railton, P. (1986). Facts and values. *Philosophical Topics*, *14*(2), 5–31.

Rizzo, M. J., & Whitman, D. G. (2009). The knowledge problem of new paternalism. *BYU Law Review*, *4*, 905–967.

Rizzo, M. J., & Whitman, G. (2020). *Escaping paternalism: rationality, behavioral economics, and public policy*. Cambridge University Press, http://dx.doi.org/10.1017/9781139061810.

Rosati, C. (1995). Persons, perspectives, and full-information accounts of the good. *Ethics*, *105*(2), 296–325.

Rosenberg, A. (1989). Are generic predictions enough? *Erkenntnis*, *30*, 43–68.

Samuelson, P. (1938). A note on the pure theory of consumer's behaviour. *Econometrica*, *5*(17), 61–71.

Sørensen, T. F. (2016). In praise of vagueness: Uncertainty, ambiguity and archaeological methodology. *Journal of Archeological Method and Theory*, *23*(2), 741–763.

Sorensen, R. (2023). Vagueness. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (winter 2023 edition)*. https://plato.stanford.edu/archives/win2023/entries/vagueness/.

Sugden, R. (2018). *The community of advantage: a behavioural economist's defence of the market*. Oxford University Press, http://dx.doi.org/10.1093/oso/9780198825142.001.0001.

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

Teller, P. (2018). Measurement accuracy realism. In I. F. Peschard, & B. C. van Fraassen (Eds.), *The experimental side of modeling*. University of Minnesota Press.

Thaler, R., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth and happiness*. Yale University Press.

Thoma, J. (2020). Folk psychology and the interpretation of decision theory. *Ergo*, *7*.

Thoma, J. (2021). On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology*, *28*(4), 350–363.

Thoma, J. (2024). Merely means paternalist? Prospect theory and 'debiased' welfare analysis. *Philosophy of Science*, *91*(1), 204–224.

Thoma, J. Money Pumps and the Instrumentalist Foundations of Decision Theory, In M. S. Sagdahl and A. Tanyi (Eds.), Problems of Choice: Normativity, Rationality, Value, and Morality, Routledge (Routledge Studies in Ethics and Moral Theory), forthcoming.

Titelbaum, M. G., & Kopec, M. (2016). The uniqueness thesis. *Philosophy Compass*, *11*(4), 189–200.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. http://dx.doi.org/10.1007/BF00122574.

Weintraub, E. R. (2002). *How economics became a mathematical science*. Duke University Press.

Williams, B. (1979). Internal and external reasons. In R. Harrison (Ed.), *Rational action* (pp. 101–113). Cambridge University Press.

Williamson, T. (1994). *Vagueness*. Routledge.

Woodford, M. (2020). Modeling imprecision in perception, valuation and choice. *Annual Review of Economics*, *12*, 579–601.