Synthese
https://doi.org/10.1007/s11229-021-03078-8

Radical interpretation and decision theory

Anandi Hattiangadi[1] · **H. Orri Stefánsson**[1,2]

Received: 22 September 2020 / Accepted: 6 February 2021

Abstract
This paper takes issue with an influential interpretationist argument for physicalism about intentionality based on the possibility of radical interpretation. The interpretationist defends the physicalist thesis that the intentional truths supervene on the physical truths by arguing that it is possible for a radical interpreter, who knows all of the physical truths, to work out the intentional truths about what an arbitrary agent believes, desires, and means without recourse to any further empirical information. One of the most compelling arguments for the possibility of radical interpretation, associated most closely with David Lewis and Donald Davidson, gives a central role to decision theoretic representation theorems, which demonstrate that if an agent's preferences satisfy certain constraints, it is possible to deduce probability and utility functions that represent her beliefs and desires. We argue that an interpretationist who wants to rely on existing representation theorems in defence of the possibility of radical interpretation faces a trilemma, each horn of which is incompatible with the possibility of radical interpretation.

Keywords Radical interpretation · Physicalism · Decision theory · David Lewis · Donald Davidson

✉ H. Orri Stefánsson
orri.stefansson@philosophy.su.se

Anandi Hattiangadi
anandi.hattiangadi@philosophy.su.se

[1] Department of Philosophy, Stockholm University, Stockholm, Sweden

[2] Swedish Collegium for Advanced Study, Uppsala, Sweden

Published online: 03 March 2021

1 Introduction

This paper takes issue with an influential interpretationist[1] argument for physicalism about intentionality. A core commitment of physicalism is that the intentional truths—about what an arbitrary agent believes, desires, and means—supervene on the physical truths—those truths that are stateable in the terms of an ideal and complete physical theory. Roughly, to say that the intentional truths supervene on the physical truths is to say that the physical truths metaphysically necessitate the intentional truths.

The interpretationist claims that we can get a handle on the supervenience of the intentional on the physical by considering the predicament of a radical interpreter, an ideally rational being who knows all the physical truths and who sets out to deduce the intentional truths about an arbitrary agent—let's say, Karla—without recourse to any semantic or intentional information.[2] The radical interpreter is aided by a set of a priori principles or constraints on interpretation, such as most famously some kind of Principle of Charity. Radical interpretation is possible only if the physical truths, together with the postulated constraints, entail the correct interpretation of Karla. If radical interpretation is possible, intentionality can be reductively explained, and the intentional supervenes on the physical.

Is radical interpretation possible? We focus here on what we take to be the most compelling argument for its possibility, associated with Donald Davidson and David Lewis. This argument gives a central role to decision theory—the theory of an agent's choices and how they are influenced by her beliefs and desires. In Sect. 2, we explain the fundamental challenge that the interpretationist faces. In Sect. 3, we explain how decision theory seems to hold the promise of an elegant solution to this challenge, and how the whole interpretationist edifice rests on decision theoretic foundations.

In Sect. 4, we investigate the cracks in these foundations. We argue that champions of the interpretationist strategy face a trilemma. If the interpretationist relies on existing decision theories and theorems, the procedure of radical interpretation either (1) *underdetermines* the correct interpretation; (2) is *inapplicable* to ordinary agents; or (3) appeals to information that is in principle *inaccessible* to the radical interpreter. One way or the other, radical interpretation based on existing decision theories fails. In the concluding Sect. 5, we discuss the bleak prospect of a

[1] Classic articulations of interpretationism can be found in Davidson (1973), Dennett (1987), and Lewis (1974). For recent defences of interpretationism, see McCarthy (2002), Pautz (2013) and Williams (2018, 2020); for recent criticism, see Hattiangadi (2020), Simchen (2017), and Williams (2007, 2016). Note that Williams (2016) specifically objects to decision-theoretic radical interpretation, though his objection is different from those we raise here: he argues that the facts available to the decision theoretic radical interpreter fail to rule out deviant 'bubble interpretations' in which an agent has a very low credence in and is indifferent towards events that occur outside of the agent's immediate spatiotemporal surroundings. To keep our discussion as focused as possible, we will have to leave a more detailed comparison of our worries about radical interpretation with Williams' criticism to another occasion.

[2] This characterization of radical interpretation as figuring in an argument for physicalism has often been left implicit in interpretationist writings, where the inference seems to be from physicalism to the possibility of radical interpretation (cf. Lewis 1974). Chalmers (2012) makes it explicit.

yet-to-be-formulated decision theory that would allow the interpretationist to avoid the trilemma.

2 The reductionist's challenge

To begin with, some preliminary points are in order. First, we assume realism about intentionality, that there is a fact of the matter what Karla believes, desires, and means. In this setting, the radical interpreter sets out to *discover* Karla's intentional states, and he succeeds only if he arrives at the correct interpretation of her. Though the interpreter's task is framed in epistemological terms—going from knowledge of the physical to knowledge of the semantic and intentional—this serves merely to dramatize what is a fundamentally metaphysical question, as Lewis puts it: 'how do *the facts* determine these [semantic and intentional] facts?' (Lewis 1974, p. 334, emphasis in original)[3]. In contrast, deflationists, anti-representationalists, eliminativists, and the like, hold that there are no substantive semantic or intentional properties 'out there' in the world, awaiting interpretation. In an anti-realist setting, the radical interpreter might 'project' intentional concepts onto a purely physical reality, or 'construct' the intentional and the semantic through interpretation. Though anti-realist interpretationism raises interesting issues in its own right, they are orthogonal to the issues we wish to focus on here, so we will set them aside.[4]

Second, we will concentrate on the interpretationist argument for the view that the intentional supervenes on the *physical*.[5] We set aside here reductive accounts of intentionality that appeal to properties that are assumed not to supervene on the physical. For instance, some hold that the intentional supervenes at least partly on phenomenal properties, the 'what it is like' of experience or cognition (cf. Chalmers 2006; Horgan and Tienson 2002; Mendelovici 2018; Pautz 2013). Again, these alternatives to physicalism raise issues that are orthogonal to those that will be explored here.

[3] Thus, our characterization of the role of the interpreter follows Lewis more closely than Davidson, who suggests in places a more constructivist reading, for instance in his remarks on interpretation and the publicity of meaning and belief (cf. Davidson 1983, p. 315).

[4] For instance, Dennett (1987) holds that intentional categories are projected onto physical reality when we adopt the 'intentional stance' towards what is fundamentally a physical system. Davidson's remarks on the indeterminacy of meaning and the inscrutability of reference are suggestive of a sympathy for anti-realism (cf. Davidson 1991). However, since our aim here is not exegesis, we focus on realist interpretationism.

[5] We assume here that the properties investigated by the biological and chemical sciences supervene on the physical properties, so it is unnecessary to add these properties to the base. If it turns out that these properties do not supervene on the physical, they may be added to the base without affecting any of the arguments of this paper.

Third, the physicalist's supervenience thesis can be stated more precisely as follows.[6] Let *P* be a statement of all of the positive physical truths about the world, where it is a positive truth that there is an electron at such and such a position in space–time, and a negative truth that there are no unicorns. Let *T* be a 'that's all' statement to the effect that nothing more exists than is needed to satisfy whatever precedes it, so *PT* states that nothing more exists than is needed to satisfy *P*. And let *S* be a statement of all of the contingent semantic and intentional truths about the world, including all the truths about what Karla believes, desires, and means.[7] Note that *S* could either ascribe propositions understood as sets of possible worlds, as Lewis proposes, or consist of T-sentences specifying truth conditions, as Davidson does. With this in place, physicalism entails the following supervenience thesis:

Supervenience. Necessarily, if *PT* then *S*.

The interpretationist aims to defend Supervenience by giving a reductive explanation of meaning and intentionality, an account stateable in physical terms of what constitutes belief, desire and meaning. Any such account must meet the following two conditions of adequacy. First, there is a *deviance condition*. Supervenience states that in all metaphysically possible worlds, *w*, if *PT* is true at *w*, then *S* is true at *w*, which entails the impossibility of *deviant scenarios* in which *PT* is true and *S* is false. Any reductive account of the intentional must therefore rule out the possibility of all deviant scenarios. For instance, suppose that Karla believes X. If Supervenience is true, there is no possible world that is just like our world in physical respects (and that's all), at which Karla does not believe X, but believes Y, something indeterminate between X and Y, or nothing at all. If *PT* together with the constraints underdetermines *S*, then the deviance condition is not met, since a scenario in which *PT* is true but *S* is not remains an open possibility. Second, there is a *circularity condition*. Since the physicalist's challenge is to specify the *physical* truths on which the semantic truths supervene, the overall account of the intentional must ultimately bottom out in some intentional phenomena that supervene directly on the physical. For instance, this constraint would be violated if one were to give a reductive account of conventional meanings in terms of speaker intentions or beliefs, without a further reductive account of intentions and beliefs.

[6] For an overview of many different ways to formulate supervenience theses, see McLaughlin and Bennett (2018). The formulation we give follows Chalmers (2009), and is intended to capture the kind of minimal materialism endorsed by Lewis (1983, pp. 361–364). Many physicalists would endorse Semantic Strong Global Supervenience (SSGS), according to which for any two worlds, w_1 and w_2, any isomorphism between the domain of w_1 (i.e., set of objects existing at w_1) and the domain of w_2 that preserves the physical properties preserves the semantic properties, where any two worlds w_1 and w_2 are Ψ-indiscernible if there is a one-to-one function *f* from the domain of w_1 onto the domain of w_2, and for any Ψ-property, *P*, and for any object *a* in w_1, $P(a)$ iff $P(f(a))$ (cf. Shagrir 2009). For objections to SSGS, see Magidor and Kearns (2012). Since SSGS entails Supervenience, if the interpretationist argument for Supervenience fails, so too does the interpretationist argument for SSGS.

[7] Lewis (1975) distinguishes between languages understood as abstract objects, mappings from strings to meanings, and a language understood as a social phenomenon, involving speaking and understanding utterances. *S* states the contingent truths about language in the second sense, rather than the necessary truths about languages in the first sense.

The interpretationist's strategy is to articulate a set of a priori constraints on interpretation, which, together with the physical truths, entail *S*. If there is an a priori entailment from *PT* to *S*, then the possibility of *PT* & ~ *S* can be ruled out a priori, and the deviance condition can be met. If the constraints can be specified in purely physical terms, the circularity condition can be met. As we shall see, the appeal to decision theory is essential to what we take to be the strongest argument for the possibility of radical interpretation.

3 The decision-theoretic argument for the possibility of radical interpretation

Lewis' (1970, 1972, 1994) argument for the possibility of radical interpretation is best viewed against the background of his overall analytic functionalist approach to the reduction of mind.[8] According to Lewis, folk psychology can be regarded as a term-introducing scientific theory, containing a set of platitudes that serve to implicitly define theoretical terms such as 'pain', 'belief', 'desire', and 'meaning'. The implicit definitions of these terms specify the functional roles of the states they ascribe by way of their characteristic relations to physically described inputs and outputs, and other mental states. For example, it is a platitude of folk psychology that bodily damage typically causes pain, and that pain typically causes aversion and characteristic pain-behaviour. The theory contained in idealized folk psychology can be expressed with the aid of a sentence (FP) consisting of a long conjunction of platitudes of this kind.

Lewis proposes to convert the implicit analyses of theoretical terms in FP to reductive analyses by constructing the *Carnap sentence* of FP. This is achieved by taking every theoretical term for a mental state, M_1, M_2, M_3 (etc.), and replacing it with a variable, *x*, *y*, *z* (etc.). The result is an open sentence that contains only logical vocabulary, variables and the physical terms used to describe inputs and outputs, which we can designate 'FP*'. The Carnap sentence of FP binds FP* under an existential quantifier ('there exist a unique *x*, *y*, *z*…') and links the variables in FP* with the theoretical terms they replaced. It states that *if* there exist a unique *x*, *y*, *z* (etc.) such that FP*, then *x* is M_1, *y* is M_2, *z* is M_3 (etc.). On this view, what it is to be in mental state M_i is to be in a state that occupies the M_i-role, as that role is defined by folk psychology.

Like Lewis, Davidson takes our ordinary practice of interpreting others as his starting point. He says, for instance, that 'talk apparently of thoughts and sayings belong[s] to a familiar mode of explanation of human behaviour and must be considered an organized department of common sense which may as well be called a theory,' and that one way to examine the nature of thought and talk 'is by inspecting the theory implicit in this sort of explanation' (Davidson 1975, p. 158). However, whereas Lewis holds that folk psychology gives rise to analyses of our intentional concepts, Davidson takes our ordinary practice to give rise to principles that are

[8] For overviews, see Schwarz (2015) and Weatherson (2016).

constitutive of our intentional mental states (cf. Davidson 1990, p. 317). This difference notwithstanding, both hold that our ordinary practice of interpretation gives rise to a set of a priori constraints on interpretation, by the application of which the radical interpreter can come to know *S* given knowledge of *PT*.

With this in place, it is easy to see the overall shape of the interpretationist's argument for the possibility of radical interpretation. If a radical interpreter is omniscient of the physical truths, he knows not only the microphysical truths, but also physical truths about the behaviour and dispositions of macrophysical objects, such as Karla. The interpreter knows, for instance, that certain physical states of the world typically cause Karla to be in state P_1, and that if she is in state P_1, she is disposed to be behave in way *B*, physically described. On the basis of knowledge of this kind, the interpreter can determine whether P_1 uniquely occupies any of the functional roles specified by folk psychology. If the interpreter *does* find that P_1 uniquely occupies the M_1-role, then together with the Carnap sentence for FP, this entails that P_1 is M_1. It follows that there are a priori entailments from the physical truths to the intentional truths, which are formulated as constraints on interpretation, such as a Principle of Charity.

Compelling though this proposal may be in its general outline, it is at best a promissory note. After all, there is no guarantee that the constraints on interpretation will be uniquely satisfied. For instance, on a common sense understanding of charity, the charity of an interpretation is in the eye of the beholder. If Karla is charitably interpreted by three different people, each with different intrinsic desires, beliefs, and evidence, she will undoubtedly be interpreted differently by each (cf. Eriksson and Hájek 2007, p. 199). Yet if the constraints on interpretation can be satisfied by multiple inconsistent interpretations, the radical interpreter is not in a position to deduce the semantic truths from his knowledge of the physical truths together with the constraints, and the possibility of deviant scenarios cannot be ruled out.

Decision theory holds out the promise of an inspired solution to this difficulty. The interpretationist's key claim is that decision theory articulates constraints on interpretation that are already implicit in folk psychology, and hence are constitutive of an important class of propositional attitudes.[9] As Lewis put it, decision theory 'is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference, and choice…the very core of our common-sense theory of persons, dissected out and elegantly systematized' (Lewis 1974, p. 337). Moreover, both Lewis (1974, p. 337) and Davidson (1985, 93ff; 1990, pp. 323–324) were

[9] Since he takes decision theory to *constitute* the attitudes, Lewis clearly subscribes to a mentalistic interpretation of decision theory, which treats credences and utilities as psychologically real, as opposed to a behaviourist interpretation, according to which decision theory only treats agents 'as if' they have credences and utilities, because these are merely representations of preferences or choices. More generally, if decision theory is to provide a foundation for radical interpretation, then any behaviourist, or 'as if' interpretation of decision theory is out of the picture. For unlike some economists, for instance, the interpretationist is not really interested in examining when we can represent someone *as if* she had certain credences and utilities; rather, the interpretationist wants to explain how we can discover what others *actually* desire and believe. (We thank a referee for pointing out the need to clarify this.) For a discussion of the contrast between these different interpretations of decision theory, see Okasha (2016). For an example of an economist who subscribes to the behaviourist ('as if') interpretation, see Gilboa (2009).

struck by the power of Bolker's (1966) representation theorem for Jeffrey's (1965) decision theory, which *proves* that if an agent's preferences satisfy certain minimal constraints—the Bolker–Jeffrey axioms—it is possible to deduce probability and utility functions that can be understood as representing the agent's degrees of belief and desire.[10] By appeal to this representation theorem, they motivate the claim that there is an a priori entailment from the truths about Karla's preferences to truths about her beliefs and desires, which in turn motivates the claim that there is an a priori entailment from *PT* to *S*—provided, of course, that an agent's preferences can be described in physical terms (a matter to which we shall presently return). When formulated as a constraint on interpretation, the *Rationalization Principle* (Lewis 1974, p. 337) tells the radical interpreter to assign preferences to Karla on the basis of information about her physically described choice behaviour in such a way as to satisfy the Bolker–Jeffrey axioms, and to assign a credence and a utility function to her that make her out to maximize expected utility.

Thus, interpretationists argue, a radical interpreter who is guided by decision theory is guaranteed to succeed at least in the first step of assigning beliefs and desires to agents: if the radical interpreter knows that Karla's preferences satisfy the Bolker–Jeffrey axioms, he can rely on the Bolker–Jeffrey representation theorem to deduce Karla's degrees of belief and desire; if *what it is* to have those beliefs and desires *just is* to be representable as such by the lights of decision theory, the radical interpreter simply cannot fail.[11] Since this serves to rule out the possibility of any deviant scenario in which Karla's preferences satisfy the constraints, but she lacks the beliefs and desires that decision theory assigns, the deviance condition has been met.

What of the circularity condition? Even if the radical interpreter may be able to deduce the truths about Karla's beliefs and desires if he knows her preferences, how can he come to know Karla's preferences on the basis of the physical information alone? After all, Karla's preference for tea over coffee is a contentful psychological state, which by hypothesis is initially unknown to the radical interpreter. In order to meet the circularity condition, it must be shown that the radical interpreter could deduce Karla's degrees of belief and desire from *PT* without recourse to any semantic or intentional information. Lewis has little to say about this aspect of the enterprise, suggesting merely that an agent's preferences might be knowable on the basis of 'raw behaviour' (Lewis 1974, p. 338).

Davidson is far more alive to this problem, proposing to solve it in two stages (Davidson 1990, p. 315). In the familiar second stage, the radical interpreter applies

[10] Davidson (1990, p. 323) discusses Frank Ramsey's decision theory and dismisses it as a non-starter for his purposes. The problem, as he sees it, is that Ramsey's approach involves presenting agents with gambles or wagers which are described in sentences in a language the agent understands. Thus, the radical interpreter cannot get started with Ramsey's method without first understanding the agent's language.

[11] Meacham and Weisberg (2011) argue that representation theorems only prove that if an agent satisfies the relevant constraints, she can be *represented* as having certain degrees of belief and desire, which does not on its own entail that the agent *does* have these (or any other) degrees of belief and desire. It only has the potential to do so against the background of the view that decision theory is constitutive of the attitudes.

the Principle of Charity to the interpretation of Karla's language, assigning meanings to the sentences of Karla's language in such a way as to maximize truth in the set of sentences she holds true. In the first stage, he appeals to Jeffrey's decision theory to explain how the radical interpreter can come to know which sentences Karla holds true, without yet knowing what those sentences mean (Davidson 1990, p. 317).

Davidson begins by replacing the propositions that serve as the objects of preference in Jeffrey's theory with *sentences* of the subject's language. He argues that the radical interpreter could come to know which sentences Karla prefers true by observing her pairwise choices between sentences—without first knowing what those sentences mean (Davidson 1985, p. 88; 1990, p. 317, p. 323).[12] However, in order to apply the Bolker–Jeffrey representation theorem to an interpretation of Karla, the radical interpreter must know the Boolean structure of the objects of Karla's preferences. So, he must first determine the meanings of the truth functional connectives in Karla's language. Here, too, Davidson appeals to the apparatus of decision theory to argue that it is possible for the radical interpreter to determine, on the basis of observations of Karla's choice behaviour, that some initially uninterpreted truth-functional connective in Karla's language is the Sheffer stroke, meaning 'not both X and Y', from which all other truth functional connectives can be determined (Davidson 1985, 1990). Given a set of sentences with a Boolean structure, together with information about Karla's preferences over sentences, the radical interpreter can then proceed to use the modified Bolker–Jeffrey representation theorem to determine which sentences Karla holds true. In this way, Davidson proposes to meet the circularity condition.

It is worth emphasizing that for both Lewis and Davidson, decision theory comes in at the ground level. In Davidson's case, the radical interpreter must first appeal to decision theory to determine which sentences Karla holds true, since this constitutes the data to which the Principle of Charity is applied. And although Lewis holds that the radical interpreter must begin by assigning beliefs and desires to Karla in accordance with the Rationalization Principle *and* the Principle of Charity, suggesting that the two principles are equally foundational, it is only with the Rationalization Principle that a plausible story can be told about how the radical interpreter might break into the intentional circle of belief, desire, and meaning, given only knowledge of her 'raw behaviour'. In contrast, to apply Lewis' Principle of Charity, the radical interpreter must already know a great deal about Karla's intentional states. This principle instructs the radical interpreter to assign beliefs that are rational in light of Karla's evidence according to some suitable inductive method (Lewis 1969, p. 534), such as Bayesian conditionalization (Lewis 1983, p. 374). So, if Karla acquires some evidence E, the radical interpreter should take her to assign a posterior probability to any hypothesis H that is equivalent to the prior conditional probability she assigned to H given E. But that means that to apply the Principle of Charity, the

[12] It is admittedly somewhat puzzling how this procedure is thought to be carried out. How can the radical interpreter tell whether Karla chooses one sentence over another because she believes it to be true, rather than because she wants it to be true? We set this difficulty aside here.

radical interpreter must first know both the content of E and the prior probability Karla assigns to H given E—neither of which he can be assumed to know at the outset.[13] Though Lewis (1974) cites further principles as constraints on interpretation, these must come in at a later stage, since they relate to the assignment of conventional meaning to the sentences of Karla's language, which given Lewis' account of convention in terms of belief (Lewis 1975), can only be applied once the radical interpreter has worked out what Karla believes. Thus, if the decision theoretic case for the possibility of radical interpretation breaks down, radical interpretation cannot get off the ground.

Before we turn to an exploration of the ways in which the decision-theoretic case for the possibility of radical interpretation ultimately breaks down, a brief note on indeterminacy is in order. Both Davidson and Lewis suggest that if it turns out that the physical truths together with the constraints on interpretation underdetermine which of several conflicting interpretations is correct, then Karla's intentional states are simply indeterminate in the respects in which those interpretations differ, which might appear to be inconsistent with the assumption of realism about beliefs, desires, and meanings. Lewis attempts to reconcile his remarks on indeterminacy with realism by suggesting that the only indeterminacy that will remain reflects genuine indeterminacy in Karla's intentional states, such as might be present in 'the confused desires of the compulsive thief' (Lewis 1974, p. 343). If interpretation is underdetermined to a more 'virulent' degree (Lewis 1974, p. 342), if it results in an indeterminacy that is *not* reflected in the intentional states Karla actually has, Lewis endorses the credo that we have just not found all of the constraints.

In contrast to Lewis, Davidson seems to be open to more extensive indeterminacy. For instance, he says: 'Because there are many different but equally acceptable ways of interpreting an agent, we may say, if we please, that interpretation or translation is indeterminate, or that there is no fact of the matter as to what someone means by his or her words' (Davidson 1991, p. 161). Yet, he suggests that this indeterminacy is benign, much like the difference between representing temperatures using Fahrenheit and Centigrade scales (Davidson 1991, p. 161). Critics have noticed that it is far from clear that the indeterminacy entailed by Davidson's theory is as superficial as he sometimes suggests (cf. Child 1994, p. 73; Hacking 1975, p. 155).

At any rate, since our primary concern here is not with exegesis, but with realist interpretationism, we impose strict limits on the extent of indeterminacy that is allowed—set by the semantic and intentional facts. If Karla has indeterminate beliefs about heaps and clouds, or indeterminate degrees of belief with respect to how likely it is that it will rain, this indeterminacy will be reflected in the true interpretation of her. But if she in fact *has* determinate attitudes, and the physical truths together with all of the constraints nevertheless underdetermine which interpretation of her attitudes is correct, then the possibility of deviant scenarios has been left open, and radical interpretation fails.

[13] Note that even if we allow the radical interpreter knowledge of the phenomenal truths, as Pautz (2013) proposes, she can only be represented as conditionalizing on the *content* of that evidence.

4 The trilemma

As we have seen, the appeal to decision theory plays a crucial role in what we take to be the strongest case for the possibility of radical interpretation. In this section, we look more closely at existing decision theories, and the representation theorems that have been proven for them, in in order to evaluate their suitability to play this role. Starting with Jeffrey's decision theory and representation theorem—to which both Davidson and Lewis appeal—and moving on to further existing theories and theorems, we show that none of them is well-suited to play its designated role.[14] The trouble is that a radical interpreter who relies on the existing representation theorems faces a *trilemma*, since for all these theorems, at least one of the following is true:[15]

1. *Underdetermination.* The theorems deliver a set of probability functions that disagree even in their *rankings* of the objects of belief, so a rational agent's preferences are compatible with several mutually incompatible interpretations of her beliefs.
2. *Inapplicability.* The theorems impose constraints on preference that are both too normatively and psychologically demanding, and hence do not come close to being satisfied by any actual agent nor even by ideally rational agents.

[14] Note that we focus only on *normative* decision theories here, setting aside *descriptive* decision theories, such as the well-known *prospect theory* (Kahneman and Tversky 1979). One reason is that these descriptive theories have been developed as generalizations of phenomena observed in empirical settings in which the investigator and the subjects of investigation share a common language. As a result, these theories clearly face both the underdetermination and the inaccessibility problems, and are ill-suited to figure in the interpretationist project. For instance, prospect theory does not directly infer subjective probabilities (i.e., degrees of belief) from an agent's preferences, but rather 'decision weights' that are *assumed* to correspond to *some* degrees of belief. Precisely *which* degrees of belief they correspond to is left underdetermined by the agent's preferences. Since prospect theory does not provide any alternative way to determine degrees of belief on the basis of physical information, knowledge of *PT* together with prospect theory underdetermine *S*. Moreover, prospect theory assumes that before a decision takes place, the agent engages in a process—the 'editing phase'—to simplify her decision-problem. Once again, the outcome of this process is not determined by the agent's preferences, and since prospect theory provides no alternative way to determine exactly how the agent has simplified her decision problem (without asking), it leaves this aspect of her state of mind inaccessible to a radical interpreter.

[15] Since Lewis endorsed *causal* decision theory in work where he was not concerned with radical interpretation (see, in particular, his 1981), it is worth noting that it too faces the trilemma. The reason, to put it somewhat crudely, is that the causal decision theories that Lewis and others have proposed all consist in embedding causal dependencies (or causal dependency hypotheses) within one of the traditional non-causal decision theories, most typically (and in Lewis' case) Jeffrey's decision theory. So, from the perspective of radical interpretation, these causal decision theories give rise to the underdetermination that Jeffrey's theory faces, and additionally make certain mental states inaccessible, given the special difficulties in inferring *causal* beliefs from behaviour. For instance, Joyce's method of arriving at a unique subjective probability function, which we discuss below, is meant to provide a foundation for his *causal* decision theory. In fact, his is arguably the most sophisticated causal decision theory that has been developed, and as we point out, it faces the inaccessibility problem. To avoid the inaccessibility problem, a causal decision theorist can either introduce additional rationality constraints that make her theory inapplicable to ordinary agents, or simply accept that belief is radically underdetermined by preference and choice.

3. *Inaccessibility.* The frameworks within which the theorems are proven contain objects the preference ranking of which the interpreter could not know on the basis of knowledge of the physical truths alone.

This trilemma arises as a direct result of the role decision theory must play for the interpretationist project to succeed. First, interpretationists take decision theory to provide a constitutive account of the attitudes, as specifying *what it is* for an agent to have beliefs, desires, or preferences *at all*. In order to fulfil this role, a decision theory must capture constraints on the attitudes that are minimal enough to be satisfied by all ordinary agents, with all of their foibles and imperfections. If a decision theory places overly demanding constraints on the attitudes, it fails to specify the necessary conditions for attitude possession, and gives rise to 'false negatives'—genuine beliefs, desires or preferences that are misclassified as non-attitudes. Second, interpretationists hope that decision theoretic representation theorems provide a bridge from knowledge of *PT* to knowledge of *S* that satisfies circularity and deviance conditions. For a representation theorem to satisfy the circularity condition, it must take as 'input' purely physical information about an agent, and to satisfy the deviance condition, it must yield as 'output' *only* those probability and utility functions that accurately represent the agent's beliefs and desires. On the input side, it must be possible for the radical interpreter to know whether an arbitrary agent satisfies the postulated constraints on the basis of knowledge of *PT* alone. On the output side, it must be possible for the radical interpreter to rule out all but the correct interpretation of the agent's beliefs and desires. As we shall see in more detail in what follows, those existing decision theories and theorems that specify constraints that are suitably minimal, and plausibly take as input purely physical information about an agent, severely underdetermine the radical interpreter's choice of interpretation, while those that limit underdetermination in their output do so by either postulating constraints that are too demanding to be applicable to ordinary agents, or whose inputs involve semantic or intentional information.

The representation theorem for Jeffrey's theory, which Lewis (1974, p. 337) and Davidson (1985, 93ff; 1990, pp. 323–324) both assumed in their defence of the possibility of radical interpretation,[16] suffers from the underdetermination problem (as Jeffrey himself was well aware). The reason is essentially an issue brought up by Joyce, namely, that since Jeffrey does not assume 'absurdly strong structural requirements on preference rankings' (Joyce 1999, p. 197), preferences in Jeffrey's theory only suffice to determine utility/probability *products*, but do not allow us to isolate the effect of the probability function from the effect of the utility function on the agent's preferences (and choice behaviour). Indeed, all versions of expected utility

[16] Jeffrey himself did not, however, claim that his theory could ensure the possibility of radical interpretation. (We thank a referee for encouraging us to clarify this.)

theory[17] assume that a rational person's preferences between alternatives correspond to a product of the utilities and probabilities of the alternative's outcomes.

More precisely, the problem is that even if an agent perfectly fits Jeffrey's theory—that is, even if her preferences perfectly satisfy the Bolker–Jeffrey axioms—there will still be a (non-singleton) set $\boldsymbol{P}$ of probability functions, such that for each function p in $\boldsymbol{P}$ it is true that the agent's preferences can be represented as maximizing expected utility (or 'desirability', to use Jeffrey's term) relative to p (+some utility function u); but for any p, q in $\boldsymbol{P}$ such that $p \neq q$, p and q will not agree on how to rank some propositions.[18] Hence, even if a radical interpreter knows all choices that Karla is disposed to make—assuming further that these dispositions match her preferences and that Karla's preferences satisfy the Bolker–Jeffrey axioms—such knowledge does not suffice to determine, for many propositions X and Y, which of these propositions Karla believes more strongly.[19] In other words, preferences in Jeffrey's theory do not entail a coherent comparative belief relation.

The result is an underdetermination of the interpretation of Karla that does not reflect any ordinary indeterminacy in her beliefs and desires. To see this, take any proposition X, that is neither a tautology nor a contradiction, and suppose that Karla does not have unbounded preferences (recall fn. 18). For any two probability functions, p and q, that represent Karla's preferences, the possible size of the difference between $p(\mathrm{X})$ and $q(\mathrm{X})$ depends on the product $p(\mathrm{X})u(\mathrm{X})$; the closer this is to 0, the less the difference will be. Now, there are, in Jeffrey's framework, two fixed points that are common to any probability/utility representation: A *tautology* (which is assumed to be neutral in value) is assigned a utility of 0 and a probability of 1 according to all functions, while a *contradiction* is assigned a probability of 0 according to all functions.[20] Therefore, if the utility of X is close to the utility of

[17] 'Expected utility theory' is interpreted here broadly to include any theory according to which a rational preference is representable as maximizing the expectation of some value function. Hence, any theory that has been proposed as a normative decision theory counts as a version of 'expected utility theory', given this terminology.

[18] There is a way of recovering, within Jeffrey's framework, a unique comparative belief ranking from an agent's preferences: by assuming that the utility function that represents the strength of the preferences is unbounded, in the sense that for any proposition X that the agent considers, there is a proposition Y that the agent prefers to X. The assumption that people have unbounded preferences seems psychologically questionable. At the very least, it is hard to see that unbounded preferences should be required for it to be possible to interpret an agent, and similarly required for an agent to have desires and beliefs at all, which an interpretationist view like the one we are considering would entail, if it were grounded in Jeffrey's decision theory. Moreover, unbounded utility functions lead to well-known problems in examples like the St. Petersburg Paradox; in fact, some (e.g. Joyce 1999) take the St. Petersburg Paradox to show that it would be irrational to have unbounded preferences.

[19] Another problem with relying on the Bolker–Jeffrey framework for radical interpretation is that the framework assumes a non-denumerable algebra of propositions. This is particularly a problem for Davidson's version of radical interpretation, according to which the Bolker–Jeffrey propositions are re-interpreted as sentences in the language of the agent to be interpreted. For as Rabinowicz (2002) points out, this means that the number of propositions needed for the Bolker–Jeffrey representation theorem exceeds the sentential resources of any language.

[20] The utility (or desirability) of the contradiction is not defined in Jeffrey's framework. In other words, one cannot have a defined conative attitude to an impossible proposition, on this view; one cannot want the impossible.

a tautology, or the probability of X is close that of the contradiction, then the difference between $p(X)$ and $q(X)$ will be small. But if X has, say, a middling probability and is more desirable than the tautology, then the difference between $p(X)$ and $q(X)$ can be considerable. Most importantly, for any two logically independent[21] propositions X and Y, Karla's attitudes to which satisfy the above constraints—i.e., the probability/utility product of each is not (close to) 0—it is possible to find two probability functions p and q that both represent her preferences, but for which, say, $p(X) < p(Y)$ but $q(Y) < q(X)$. In other words, if radical interpretation is based on a preference ordering that satisfies only the Bolker–Jeffrey axioms, then for any propositions X and Y, such that Karla is neither (almost) certain that both X and Y are false nor considers X and Y to be of (almost) neutral desirability, a radical interpreter will not be able to determine which of X and Y Karla believes to be more likely to be true.

To take an example, consider the following two propositions:

(X) It will snow in London in December 2020 and I will have a successful career.
(Y) It will snow in Copenhagen in December 2020 and I will have a successful career.

It is plausible that although Karla is much less convinced of the truth of both X and Y than the truth of a tautology, and while both X and Y are desirable to Karla, there is a fact of the matter as to whether Karla believes X or Y more strongly. But if there is such a fact, then Jeffrey's theory leaves it underdetermined. Though Karla may in fact believe Y more strongly than X, a radical interpreter guided by Jeffrey's theory cannot rule out the possibility that Karla believes X more strongly than Y. Thus, given the assumption of realism, and given that it is possible that Karla really does believe X more likely to be true than Y or vice-versa, there are semantic and intentional facts that the radical interpreter cannot come to know by appeal to Jeffrey's theory. For Lewis, this means that radical interpretation simply cannot get off the ground.

For Davidson, it means that the data on which the Principle of Charity operates—the set of sentences that Karla holds true—is radically underdetermined. This is because, on any plausible view of how holding true is connected to subjective probabilities, an interpretation of Karla based on Davidson's version of Jeffrey's theory will, for many sentences s, result in one probability function according to which Karla holds s true, and another probability function according to which Karla does not hold s true. Since Davidson's Principle of Charity tells the radical interpreter to assign meanings to the sentences of Karla's language that maximize truth in the set of sentences she holds true, if it is underdetermined which sentences Karla holds

[21] The reason we stipulate that the propositions are logically independent, is that if, say, Y = X-or-Z, then for any probability function p, $p(X) \leq p(Y)$; moreover, if Y = X-and-Z, then for any probability function p, $p(Y) \leq p(X)$.

true, it is underdetermined which interpretation is most charitable. It is unavoidably underdetermined which sentences Karla holds true because the set of propositions that get assigned probabilities and utilities in Jeffrey's theory form an *atomless Boolean algebra*,[22] so we can be sure that wherever we set the threshold, as long as it is below 1, there will be some proposition (which might be a complex disjunction or conjunction) that gets a probability just above the threshold according to one representing probability function and a probability just below the threshold according to another. So, *if* Davidson is right in thinking that we can translate the propositions in Jeffrey's framework into sentences, then Davidson's version of radical interpretation will fail to determine which of several probability functions represents an agent's beliefs.

It can be easily seen in a toy case how underdetermination in the representation of an agent's comparative beliefs induces underdetermination in the interpretation of the agent's language. Let p, q, be probability functions in $\boldsymbol{P}$ that represent Karla's preferences by the lights of Davidson's version of Jeffrey's theory, and let s_1, s_2, be sentences in $\boldsymbol{S}$: the set of sentences of Karla's language whose probabilities are not close to a that of a contradiction, and whose utilities are not close to that of the tautology. Suppose that $p(s_1)=0.6$ and $p(s_2)=0.4$, while $q(s_1)=0.4$ and $q(s_2)=0.6$, and that the threshold for holding a sentence true is around 0.5 (though note that our point holds for any threshold below 1). Then, if p represents Karla's degrees of belief, she holds s_1 true but not s_2, and if q represents her degrees of belief, she holds s_2 true, but not s_1. Now suppose that the radical interpreter sets out to maximize truth in the set of sentences Karla holds true. Relative to the stipulated choice of threshold, if he assumes p, he will arrive at an interpretation I_p that makes s_1 true but not s_2, whereas if he assumes q, he will arrive at an interpretation I_q that makes s_2 true but not s_1. Since I_p and I_q differ in their truth value assignments to s_1 and s_2, they are not equivalent. Thus, since it is underdetermined whether p or q represents Karla's degrees of belief, it is underdetermined whether I_p or I_q is most charitable.

Furthermore, we have assumed for simplicity a threshold that determines how degrees of belief map onto sentences held true. Yet, not only is the choice of threshold underdetermined by the physical truths, it is underdetermined whether *any* fixed threshold determines the function from degrees of belief to sentences held true at all. For instance, the function from degrees of belief to sentences held true could vary with context, the agent's desires, preferences or degrees of belief.[23] This only makes matters worse for the Davidsonian radical interpreter, since there are many sentences in $\boldsymbol{S}$, many probability functions in $\boldsymbol{P}$, and many functions from degrees

[22] Of course, as we noted in fn. 19, Rabinowicz (2002) points out that the set of sentences is at most countably infinite, and hence cannot form an atomless Boolean algebra. For this reason, this assumption of Jeffrey's theory does not straightforwardly carry over to Davidson's reformulation of it. However, as Rabinowicz also points out, the fact that sentences cannot form an atomless Boolean algebra undermines Davidson's proposal to replace propositions in Jeffrey's theory with sentences in the first place. We set this issue aside here.

[23] Giving up on a fixed threshold seems to be indicated by the need to deal with the lottery paradox, in which we do not think that a high degree of belief in the sentence 'my ticket will lose' corresponds to acceptance of the sentence as true. In contrast, in many ordinary cases, a high degree of belief in the sentence 'I won't be buying vacation property in the Bahamas next year' does correspond to acceptance of the sentence as true (cf. Hawthorne 2003).

of belief to sets of sentences the subject holds true (even allowing for vagueness), which combine to give a wide range of verdicts on which sentences Karla holds true. Such rampant underdetermination of which sentences Karla holds true in turn gives rise to rampant underdetermination of which interpretation of her is most charitable. Any way you slice it, radical interpretation based on Jeffrey's theory does not satisfy the deviance condition.

Decision theorists have adopted three main strategies to avoid this radical indeterminacy in Jeffrey's theory. Though these were not expressly formulated to address difficulties that arise for the interpretationist project—nor were they intended to apply to actual or ordinary agents—they are worth considering as potential solutions to those difficulties.[24] First, they have imposed further (both normative and structural) constraints on the agent's preferences (e.g. Savage 1954);[25] second, they have enriched the set of things between which the agent is assumed to have preferences (e.g. Bradley 1998); third, they have postulated primitive epistemic facts, in particular, a comparative belief relation that cannot be inferred from the preference relation (e.g. Joyce 1999). All three strategies cause trouble for radical interpretation.[26]

The problem with imposing further constraints on the agent's preferences is that the stronger the constraints we impose, the less likely it becomes that any actual agent comes close to satisfying them, and radical interpretation of ordinary agents becomes impossible. So, although in theory, imposing some such constraints means that we can infer, from the agent's preferences (or choice dispositions), a probability function and a utility function unique up to a choice of scale and starting point, we know that no actual agent will come sufficiently close to satisfying these constraints for radical interpretation to be possible.

Savage (1954) is the best-known advocate of the first strategy, that is, the strategy of deriving a unique probability function from a preference relation by imposing stronger constraints than Jeffrey does. A general problem with Savage's strategy, from the perspective of radical interpretation, is that he delivers a representation theorem by imposing such strong structural constraints on an agent's preferences that

[24] Note that none of our criticisms in this paper are directed towards decision theories per se, but their suitability to figure in the interpretationist's project. (We thank a referee for making us see the need to emphasise this.)

[25] Williams (2018, 2020) develops a strategy of this kind, and recommends that the radical interpreter assigns beliefs and desires in such a way as to maximize her substantive rationality, or 'reason-responsiveness' (Williams 2018, p. 48). However, as Hattiangadi (2020) argues, Williams' proposal leaves the interpretation of an agent's beliefs and desires underdetermined.

[26] In response to the permutation problem put forward by Putnam (1980), according to which the physical truths together with all the constraints underdetermine the assignment of properties to predicates in the interpretation of a language, Lewis (1984) appeals to an objective naturalness ordering of properties. Some properties, Lewis claims, are more natural than others, and the more natural a property, the more eligible it is to serve as the semantic value of a predicate. A similar response to the underdetermination problem discussed here is of no use, since the probability functions in the set that represents Karla's preferences need not differ with respect to naturalness, as is clear from the example given in the main text. Even Sider's (2011) more expansive notion of structure offers little hope of a solution, since there is no clear sense in which just one of the probability functions in the set carves nature at the joints, or comes closer to doing so than the others. (For a discussion of the role of naturalness in Lewis' later work, see Weatherson 2013.)

no actual (or even ideally rational) agent plausibly comes close to satisfying them. Moreover, even when the constraints are satisfied, they render radical interpretation impossible.

An important requirement of Savage's is that the set *A* of acts that an agent is assumed to have preferences between contains *all* possible functions from the set *S* of states of the world to the set *C* of possible consequences. A particularly worrying implication of this requirement is the so-called *constant act assumption*: for any consequence *c* in *C*, there is some act *a* in *A* that delivers *c* in any possible state in *S*, and which is equally desirable as *c*. So, for Savage's theorem to give hope for the possibility of radical interpretation, it must be possible to devise choice situations that reveal an agent's attitudes to such 'constant acts.' But the problem is that many such constant acts will be impossible; both physically impossible and epistemically impossible according to the agent in question.

In recent work, Gaifman and Liu (2018) show that one can relax Savage's constant action assumption by assuming that there are only *two* (non-equivalent) constant acts. This might suggest a way to render Savage's theory of some use to the radical interpreter. However, while the weaker assumption is in many ways a great improvement, it is still disastrous from the perspective of radical interpretation. Suppose, for instance, that Karla takes there to be a greater than zero chance of a meteorite strike that kills all of humanity sometime in the near future. Even on the weakened constant act assumption, there will have to be some action in her preference ordering that delivers the same consequence in the state where the meteorite strikes as in, say, any state where humanity's existence continues for another 200,000 years. But it is hard to see that such an action would be physically possible; let alone epistemically possible according to any sensible agent.

In the context of the interpretationist view that decision theory provides an account of the nature of belief and desire, the constant act assumption (even the weaker version) has the implausible implication that it is a necessary condition for an agent to have beliefs and desires at all that she has preferences with regard to acts that are physically impossible and/or epistemically impossible by her own lights. Moreover, even if this implausible hypothesis is true, the constant act assumption creates insuperable problems for a radical interpreter, who must devise choice situations that would reveal an agent's attitudes to impossible acts. Yet, if an act is physically impossible, then there is no physically possible circumstance in which an agent could perform the act. This difficulty may seem less acute for the Davidsonian radical interpreter, who initially establishes Karla's preferences over sentences, since it is physically possible to present Karla with choices between sentences that describe physically impossible circumstances. Nevertheless, if an act is epistemically impossible by Karla's lights, she may have no preference with regard to the truth of a sentence describing that act. Either way, the true interpretation of Karla remains severely underdetermined by the evidence available to the interpreter, and radical interpretation is thwarted at the outset.

Indeed, Savage's theory additionally faces the problem of *inaccessibility*. This is because *S*, *C* and *A* are sets that the *modeler* constructs in order to represent the agent, but do not necessarily correspond to how the agent herself conceptualizes states, consequences, and acts. This is clearly problematic from the perspective of

radical interpretation since, for instance, the functions in A may not correspond to the acts that the agent herself takes to be available to her, and it is impossible to know, on the basis of PT, whether they do. This aspect of the agent's mind is in effect inaccessible. All told, from the point of view of the radical interpreter, Savage's theory is a non-starter.[27]

Joyce (1999, pp. 138–145) suggests a very different way to recover a unique probability function from an extension of Jeffrey's theory. The suggestion, which Jeffrey himself had mentioned in passing, is to add to Jeffrey's framework a primitive comparative belief relation which, in addition to the preference relation, is taken to represent the attitudes of the agent. Crucially, the postulated comparative belief relation—which represents psychological states of being more confident in one proposition than another—cannot be derived from the preference relation (nor, in fact, from anything else).

It should be evident that, from the perspective of radical interpretation, this renders Joyce's solution to the underdetermination problem a nonstarter. In fact, Joyce explicitly distances himself from any attempts to reduce degrees of belief to preference (see e.g. Joyce 1999, pp. 89–90). And, indeed, the comparative belief relation in Joyce's framework cannot be known on the basis of 'raw behaviour.' Recall that the problem that Jeffrey's original theory poses for radical interpretation is that even if an agent's preferences satisfy the Bolker–Jeffrey axioms, her preferences at best entail a set $\boldsymbol{P}$ of pairwise inconsistent probability functions. Joyce's result simply shows that *if* in addition to having a preference relation that satisfies the Bolker–Jeffrey axioms, an agent has a comparative belief relation that satisfies the right constraints, *then* the two relations together entail a unique probability function in $\boldsymbol{P}$. But that means that all the radical interpreter can infer, from information about the agent's preferences, is that *if* the person in question has some comparative belief relation with the right structure, *then* that comparative belief function will be represented by exactly one probability function in $\boldsymbol{P}$. The problem is that the radical interpreter neither knows whether the agent in fact has such a relation, nor which probability function in $\boldsymbol{P}$ the relation—if there is one—corresponds to, on the basis

[27] A further problem with Savage's theory, from the perspective of radical interpretation, concerns his strict separation between objects of desire and objects of belief. We shall not discuss this issue in detail, but a problem that it raises for radical interpretation is that if Karla both believes that it will be sunny and finds it desirable that it will be sunny, say, then there is a fact about Karla's attitudes that a radical interpreter cannot come to know with the help of Savage's framework. For in that framework, we cannot *both* assign a utility value and a probability value to it being sunny.

A related problem with Savage's framework, is that to apply the framework the interpreter has to construct a partition of the possibility space for which it is true that Karla believes that the probability of each element in the partition is independent of which action she performs. But that is to assume from the start a part of what the radical interpreter is meant to discover.

of knowledge of the physical truths alone.[28] Once again, the underdetermination problem for the radical interpreter remains untouched.

Bradley (1998) proposes another type of solution to Jeffrey's underdetermination problem, which involves enriching the domain of the preference relation. In particular, he shows that adding certain types of (non-truth functional) indicative conditionals to Jeffrey's framework resolves the underdetermination.[29] As Bradley points out, the (quite standard) interpretation of the preference relation in Jeffrey's framework as an attitude to *news items* makes it very plausible that agents have preferences between indicative conditionals: 'One might prefer to learn, for instance, that if it's chicken for dinner then a white wine will be served than to learn that if it's beef, then beer will be served' (Bradley 1998, p. 188).

However, the inclusion of indicative conditionals in Jeffrey's decision theory undermines the use of the theory as a basis for radical interpretation.[30] For a radical interpreter to determine the agent's preference between conditionals, there would have to be possible choice situations that would reveal such preferences. But as Bradley's example illustrates, there are many (natural, common and simple) conditionals for which that would be impossible. For instance, although we can certainly construct choice situations that reveal an agent's preference between chicken and white wine on one hand, and beef and beer on the other, it would be impossible to devise a choice situation that reveals the agent's preference between 'if chicken, then white wine' and 'if beef, then beer'. Thus, even with the assistance of Bradley's extension of Jeffrey's theory, the information available to the radical interpreter remains insufficient to determine the true interpretation of the agent.

It might seem that Davidson's proposal to reformulate constraints on preference in terms of preferences over uninterpreted sentences fares better here. In this setting, perhaps all the radical interpreter would need to do to make use of Bradley's extension would be to present Karla with suitable sentences of her language, and observe her preferences between them. However, if the indicative conditional is not a truth functional connective, the radical interpreter cannot use Davidson's method to identify it, and thus cannot make use of Bradley's extension at the point at which it is most needed. Recall that Davidson's radical interpreter begins by identifying the Sheffer stroke in Karla's language, and uses that to derive the remaining truth functional connectives; if the indicative conditional is not truth functional, the radical interpreter simply cannot use the Sheffer stroke method to identify it. At best, the

[28] An interpretationist might respond to the above argument that the Principle of Charity, which Lewis formulates (Lewis 1974, pp. 336–337), would allow the radical interpreter to identify the comparative relation that correctly represents Karla's beliefs. The problem with this reply is that to identify Karla's actual comparative belief, one would initially need to know not just her life history of evidence and training, but also how she *interprets* the evidence that she has received, and what *prior* beliefs she had before receiving any evidence. But given what we have said above, it should be evident that neither Karla's interpretation nor her prior belief can be read off from her preferences (assuming Jeffrey's theory).

[29] What we say about Bradley's (1998) theorem also holds for Bradley and Stefánsson's (2017) theorem and the theory presented in Bradley (2017).

[30] In fairness to Bradley, it should be noted that his theorem was not intended as a basis for radical interpretation. (We thank a referee for making us see the need to emphasise this.)

Davidsonian radical interpreter might be able to interpret the indicative conditional in Karla's language at the second stage, when he applies the Principle of Charity to maximize truth in the sentences Karla holds true. But if the indicative conditional can only be identified at this later stage, the radical interpreter cannot make use of Bradley's extension to resolve the underdetermination inherent in Jeffrey's theory, which is needed to determine which sentences the subject holds true—the very *data* on which the Principle of Charity is to be applied.

Perhaps it will be suggested that the radical interpreter could simply *assume* that the indicative conditional is truth-functional. He could then use the Sheffer stroke method to identify it and present the agent with choices between the relevant sentences containing it. However, such a move would render Bradley's extension useless as far as deriving a unique probability function is concerned. If the interpreter wants to use Bradley's method to ensure such uniqueness, he must know Karla's preferences between sentences containing a non-truth functional indicative conditional. For, unless the interpreter assumes that the sentences in question have a non-truth functional structure—that is, without assuming that the agent's background algebra is a 'conditional algebra'—the interpreter is back in Jeffrey's indeterminacy.

5 Concluding remarks

In sum, given the information contained in *PT*, together with the existing decision theories and theorems, the radical interpreter's choice of interpretation is severely underdetermined, and any attempt to avoid this underdetermination either renders the interpretation based on decision theory inapplicable to ordinary agents altogether, or appeals to information that could not in principle be contained in *PT*. Though we do not take the foregoing considerations to show that no *possible* decision theory could bridge the gap between *PT* and *S*, it seems foolhardy to hope that the 'correct' decision theory is yet to be found; that we have not yet found all of the constraints. This is because the difficulties we have raised result from the structure of these theories and theorems and the role they must play in radical interpretation for the interpretationist project to succeed.

The radical interpreter needs a representation theorem that specifies the minimal constraints on preferences that an agent must satisfy in order to have attitudes at all. The constraints must be *minimal* because they must capture necessary conditions for the possession of the attitudes, and hence must be satisfied by all agents, not only those who are ideally rational. The constraints must be placed on *preferences* because in order to avoid circularity, the radical interpreter needs to be in a position to know whether Karla satisfies the rationality constraints on the basis of knowledge of *PT* alone. And while it is at least arguable that the radical interpreter could determine Karla's preferences from knowledge of her 'raw behaviour' as described in *PT*, it is not even remotely plausible that he could additionally determine her beliefs and desires directly from *PT,* without relying on further semantic or intentional information. Indeed, if the radical interpreter could directly determine Karla's beliefs and desires from *PT*, without going via knowledge of her preferences, the interpretationist would have no need to rely on representation theorems in the first place. So,

any representation theorem that serves the radical interpreter's needs must restrict itself to capturing minimal rationality constraints on preferences. However, if a theory places only minimal constraints on an agent's preferences, then the most that a radical interpreter can infer from knowledge of a rational person's preferences are utility/probability products, from which he cannot infer any particular probability and utility functions. In general, if all the radical interpreter has to work with are minimal rationality constraints on preference, the result is an underdetermination of interpretation far more radical than any realist interpretationist would be willing to countenance. And of course, if interpretation is underdetermined, the possibility of deviant scenarios is not ruled out.

On the other hand, if the radical interpreter is to avoid underdetermination, he must appeal to representation theorems that go beyond specifying minimal rationality constraints on preferences. Yet, as we have seen, hitherto proposed further constraints on the attitudes that are powerful enough to eliminate underdetermination are either too strong to be applicable to ordinary agents or can be applied to an agent only given knowledge of some semantic or intentional information not contained in *PT*. What the interpretationist needs is a representation theorem that specifies rationality constraints that strike exactly the right balance between being weak enough to be constitutive of the attitudes and being strong enough to determine them. Yet, the foregoing review of the existing decision theories and theorems leaves it far from obvious how this delicate balance in the specification of constraints is to be achieved, if it is achievable at all. Thus, the prospect of a decision theoretic argument for the possibility of radical interpretation is bleak.[31]

Acknowledgements Both Hattiangadi and Stefánsson gratefully acknowledge funding from Riksbankens Jubileumsfond (Hattiangadi through Pro Futura Scientia VI, Stefánsson through Pro Futura Scientia XIII).

Funding Open access funding provided by Stockholm University.

[31] This paper has been presented at the Stockholm-Boulder Workshop on Cognitive Value (Stockholm, 2019), at the Stockholm University CLLAM seminar (Stockholm, 2019), and at the Australian National University (Canberra, 2019). We are very grateful to the audience for their questions, comments, and suggestions. We have also benefitted from discussing parts of this paper with Lisa Bortolotti, Richard Bradley, Alan Hájek, Graham Oddie, Peter Pagin, David Papineau, Wlodek Rabinowicz, Daniel Stoljar, and J. Robert G. Williams. Finally, thanks to anonymous reviewers for comments and suggestions that helped us improve the paper.

References

Bolker, E. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society, 124*(2), 292–312.

Bradley, R. (1998). A representation theorem for a decision theory with conditionals. *Synthese, 116*(2), 187–229.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge: Cambridge University Press.

Bradley, R., & Stefánsson, H. O. (2017). Counterfactual desirability. *British Journal for the Philosophy of Science, 68*(2), 485–533.

Chalmers, D. J. (2006). Perception and the fall from Eden. In T. S. Gendler & J. Hawthorne (Eds.), *Perceptual experience* (pp. 49–125). Oxford: Oxford University Press.

Chalmers, D. J. (2009). The two-dimensional argument against materialism. In B. P. McLaughlin & S. Walter (Eds.), *Oxford Handbook to the philosophy of mind*. Oxford: Oxford University Press.

Chalmers, D. J. (2012). *Constructing the world*. Oxford: Oxford University Press.

Child, W. (1994). *Causality, interpretation and mind*. Oxford: Oxford University Press.

Davidson, D. (1973). Radical interpretation. *Dialectica, 27*(1), 314–328.

Davidson, D. (1975). Thought and talk. In S. Guttenplan (Ed.), *Mind and language* (pp. 7–23). Oxford: Oxford University Press.

Davidson, D. (1983). A coherence theory of truth and knowledge. In: D. Henrich (ed.), *Kant oder Hegel?* (Stuttgart: Klett-Cotta); Reprinted in Ernest LePore (ed.) *Truth and interpretation: perspectives on the philosophy of Donald Davidson* (pp. 307–319). Oxford: Basil Blackwell.

Davidson, D. (1985). A new basis for decision theory. *Theory and Decision, 18*(1), 87–98.

Davidson, D. (1986). A nice derangement of epitaphs. In E. LePore (Ed.), *Truth and interpretation: Perspectives on the philosophy of Davidson* (pp. 43–46). Basil: Blackwell.

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy, 87*(6), 238–279.

Davidson, D. (1991). Three varieties of knowledge. In A. P. Griffiths (Ed.), *A.J. Ayer Memorial Essays*, Royal Institute of Philosophy Supplement: 30 (pp. 153–166). Cambridge: Cambridge University Press.

Davidson, D. (2001). *Inquiries into truth and interpretation*. Oxford: Clarendon Press.

Dennett, D. (1987). *The intentional stance*. London: The MIT Press.

Eriksson, L., & Hájek, A. (2007). What are degrees of belief? *Studia Logica, 86*(2), 185–215.

Gaifman, H., & Liu, Y. (2018). A simpler and more realistic subjective decision theory. *Synthese, 195*(10), 4205–4241.

Gilboa, I. (2009). *Theory of decision under uncertainty*. Cambridge: Cambridge University Press.

Hacking, I. (1975). *Why does language matter to philosophy?* Cambridge: Cambridge University Press.

Hattiangadi, A. (2020). Radical interpretation and the aggregation problem. *Philosophy and Phenomenal Research, 101*(2), 283–303.

Hattiangadi, A. (2021). Substantive radical interpretation and the problem of underdetermination. *Analysis Reviews*. Early view.

Hawthorne, J. (2003). *Knowledge and lotteries*. Oxford: Oxford University Press.

Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 520–533). Oxford: Oxford University Press.

Jeffrey, R. (1965). *The logic of decision*. Chicago, IL: University of Chicago Press.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision-making under risk. *Econometrica, 47*(2), 263–291.

Kearns, S., & Magidor, O. (2012). Semantic sovereignty. *Philosophy and Phenomenological Research, 85*(2), 322–350.

Lewis, D. K. (1969). *Convention: A philosophical study*. London: Wiley-Blackwell.

Lewis, D. K. (1970). How to define theoretical terms. *Journal of Philosophy, 67*(13), 427–446.

Lewis, D. K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*(3), 249–258.

Lewis, D. K. (1974). Radical interpretation. *Synthese, 27*(3–4), 331–344.

Lewis, D. K. (1975). Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (pp. 3–35). Minneapolis, MN: University of Minnesota Press.
Lewis, D. K. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*(1), 5–30.
Lewis, D. K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*(4), 343–377.
Lewis, D. K. (1984). Putnam's paradox. *Australasian Journal of Philosophy, 62*(3), 221–236.
Lewis, D. K. (1994). Reduction of mind. In S. Guttenplan (Ed.), *Companion to the philosophy of mind* (pp. 412–431). London: Blackwell.
McCarthy, T. (2002). *Radical interpretation and indeterminacy*. Oxford: Oxford University Press.
McLaughlin, B., & Bennett, K. (2018). Supervenience. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/supervenience/.
Meacham, C. J. G., & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy, 89*(4), 641–663.
Mendelovici, A. (2018). *The phenomenal basis of intentionality*. Oxford: Oxford University Press.
Okasha, S. (2016). On the interpretation of decision theory. *Economics and Philosophy, 32*(3), 409–433.
Pautz, A. (2013). Does phenomenology ground mental content? In U. Kriegel (Ed.), *Phenomenal intentionality* (pp. 194–234). Oxford: Oxford University Press.
Putnam, H. (1980). Models and reality. *Journal of Symbolic Logic, 45*(3), 464–482.
Rabinowicz, W. (2002). Preference logic and radical interpretation: Kanger meets Davison. In P. Gärdenfors (Ed.), *11th international congress of logic, methodology and philosophy of science* (Vol. 2, pp. 213–233). Berlin: Springer.
Savage, L. (1954). *The foundations of statistics*. London: Wiley.
Schwarz, W. (2015). Analytic functionalism. In B. Loewer & J. Schaffer (Eds.), *The Blackwell companion to David Lewis* (pp. 504–518). London: Blackwell.
Shagrir, O. (2009). Strong global supervenience is valuable. *Erkenntnis, 71*(3), 417–423.
Sider, T. (2011). *Writing the book of the world*. Oxford: Oxford University Press.
Simchen, O. (2017). *Semantics, metasemantics, aboutness*. Oxford: Oxford University Press.
Weatherson, B. (2013). The role of naturalness in Lewis's theory of meaning. *Journal for the History of Analytical Philosophy, 10*(1), 1–19.
Weatherson, B. (2016). David Lewis. In E. Zalta (Ed.). *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/david-lewis/.
Williams, J. R. G. (2007). Eligibility and inscrutability. *Philosophical Review, 116*(3), 361–399.
Williams, J. R. G. (2016). Representational skepticism: The bubble puzzle. *Philosophical Perspectives, Special Issue: Metaphysics, 30*(1), 419–442.
Williams, J. R. G. (2018). Normative reference magnets. *Philosophical Review, 127*(1), 41–71.
Williams, J. R. G. (2020). *The metaphysics of representation*. Oxford: Oxford University Press.