Philosophy and Phenomenological Research
Vol. CI No. 2, September 2020
doi: 10.1111/phpr.12578

Radical Interpretation and The Aggregation Problem

ANANDI HATTIANGADI
Stockholm University

This paper takes issue with Lewis' influential argument for the supervenience of the semantic on the non-semantic based on the possibility of radical interpretation. Radical interpretation is possible only if an ideal being, who is omniscient about the non-semantic truths, can deduce the semantic truths a priori. The radical interpreter appeals to a set of criteria of interpretation choice, such as most notably some kind of Principle of Charity.

It is argued in this paper that the radical interpreter faces an insoluble aggregation problem: the radical interpreter must jointly apply several criteria for evaluating interpretations in order to determine which interpretation is best overall. First, the situation of the radical interpreter is formally modeled using the machinery of social choice theory. Second, it is argued that either Arrow's impossibility theorem or a variant of it applies to the situation of the radical interpreter. The upshot is that radical interpretation is impossible, and Lewis' argument for semantic supervenience fails.

1. Introduction

This paper takes issue with Lewis' influential argument for Semantic Supervenience, the popular view that the semantic truths supervene on the non-semantic truths.[1] The semantic truths are those concerning the meaning or reference of sentences, utterances and sub-sentential expressions, as well as those concerning the contents of thoughts and concepts. For example, it is a semantic truth that the name 'Stockholm' denotes Stockholm, and it is a semantic truth that the belief that snow is cold has the content *that snow is cold*. To say that the semantic truths supervene on the non-semantic truths is roughly to say that any metaphysically possible world that is just like our world in all non-semantic respects is just like our world in semantic respects. Lewis'

[1] There are many ways to characterise supervenience more precisely. For an overview, see McLaughlin and Bennett (2014). A more precise characterisation is given in §2.

argument for Semantic Supervenience goes as follows (Cf. Lewis, 1974; Lewis, 1983; Lewis, 1984; Lewis, 1994):

A Priori Argument for Semantic Supervenience

(1) The semantic truths supervene on the non-semantic truths if and only if there is an a priori entailment from the non-semantic truths to the semantic truths.

(2) There is an a priori entailment from the non-semantic truths to the semantic truths.

———————

(3) Therefore, the semantic truths supervene on the non-semantic truths.

The first premise of the argument reflects Lewis' overall analytic functionalist approach to the reduction of mind (Lewis, 1970; Lewis, 1994). In 'Radical Interpretation' (1974), he defends the second premise: that there is an a priori entailment from the non-semantic truths to the semantic truths. Here, a radical interpreter is an ideally rational being who is omniscient about the non-semantic truths, yet ignorant of the semantic truths, and who sets out to discover the semantic truths about an arbitrary subject—call her 'Karla'—on the basis of knowledge of non-semantic truths and a priori resources alone. These a priori resources consist in a set of criteria for the evaluation of interpretations, such as most famously some kind of Principle of Charity (Cf. Davidson, 1973; Lewis, 1974). Lewis argues that by means of applying these criteria, the radical interpreter is able to select the correct interpretation of Karla, and thereby deduce the semantic truths from the non-semantic truths without appeal to any further empirical information.[2]

As we shall see, Lewis' optimism about the possibility of radical interpretation is misguided. There is a strong case to be made that radical interpretation is impossible, and hence that the a priori argument for Semantic Supervenience fails. Though the possibility of radical interpretation has been questioned previously (Hale & Wright, 1997; Field, 1978; Putnam, 1980; Quine, 1960; Williams, 2007), there is an obstacle to radical interpretation that has hitherto gone unnoticed: the aggregation problem.

The aggregation problem arises because the radical interpreter must apply several criteria for interpretation choice at once. As is well known, the Principle of Charity is insufficient on its own to enable the radical interpreter to deduce the correct interpretation (Davidson, 1973; Field, 1978; Lewis, 1983; Lewis, 1984; Putnam, 1980; Quine, 1950; Sider, 2011; Williams, 2007). The obvious solution is to add criteria. But this in turn gives rise to the aggregation problem.

In order to select an interpretation, the radical interpreter needs some sort of rule by which to aggregate the rankings of interpretations given by several criteria in order to determine which interpretation is best overall. The trouble is that there are at least as

[2] Though radical interpretation is as closely associated with Davidson (1973) as with Lewis, there are subtle differences between their approaches to radical interpretation and their overall accounts of the reduction of mind. None of these differences is germane to the present discussion—the aggregation problem arises just as much in a Davidsonian setting. However, to simplify the exposition, I focus on Lewis here.

many logically possible aggregation rules as there are logically possible rankings of logically possible interpretations.[3] As we shall see in §3, the situation of the radical interpreter can be formally modeled using the machinery of social choice theory. This raises the question whether some of the results of social choice theory carry over to the context of radical interpretation. Of particular interest is Arrow's impossibility theorem, which states that there is no aggregation rule that satisfies a small number of plausible constraints (Arrow, 1951;1963).[4] In §4-5, we shall see how both Arrow's theorem, and a variant of it applied by Morreau to theory choice in the sciences, can be applied to the situation of the radical interpreter. But before we consider the impossibility of radical interpretation, it will be instructive to take a closer look at the argument for its possibility.

2. The A Priori Argument for Semantic Supervenience

Let's say that an interpretation assigns meanings or contents to representational vehicles described in non-intentional, non-semantic terms. These representational vehicles might be patterns of neural activation, inscriptions, or vocalizations. What are meanings and contents more exactly? Lewis held initially that these are propositions, understood as sets of possible worlds, and that the extensions of predicates and concepts are sets of possible objects. On this view, interpretations map representations to *intensions*, which are functions from possible worlds to truth-values or extensions.[5] Lewis later came to view this picture of meaning and content as inadequate, particularly for its inability to provide a semantics for indexicals, such as 'I' and 'now', and self-locating thought (Lewis, 1979; see also Kaplan, 1989a; Kaplan, 1989b). To accommodate Lewis' later view, representations can be associated with two-dimensional intensions: functions from scenario-world pairs $\langle v, w \rangle$ to truth values or extensions. A scenario can be modelled as a centered world—a metaphysically possible world marked with an agent, time, location, and so forth, at its 'center' (Cf. Chalmers, 2006a).[6] We can leave it open for the time being whether meanings and contents are one-dimensional or two-dimensional intensions.[7]

It is worth noting that Lewis was a semantic realist, who held that the semantic truths are discovered rather than 'constructed' by the radical interpreter; that there is an

---

[3] For every logically possible interpretation, there is a logically possible aggregation rule that ranks it as strictly more choice-worthy overall than every other interpretation. Thanks to H. Orri Stefánsson for this point.

[4] For applications of Arrow's Theorem and variants of it to new domains (outside of social choice theory), see Hurley, 1985; Morreau, 2010; Morreau, 2014; Morreau, 2015; Okasha, 2009; Okasha 2011; and Stegenga, 2013.

[5] Note that this picture may need to be complicated in order to accommodate vagueness. For instance, in some cases, an interpretation may not assign a precise extension, but a vague one.

[6] There are of course further options. One might, for instance, take interpretations to map representations to structured propositions or, with Davidson (1973), hold that interpretations take the form of a T-theory, stating truth conditions for every sentence of the language.

[7] The distinction turns out to be relevant to the question whether the ranking of interpretations by the naturalness criterion is necessary or contingent, which in turn has a bearing on which variant of Arrow's theorem is applicable in the context of radical interpretation. See §4.

antecedent fact of the matter what Karla believes, desires, and intends, and what the expressions of her language mean. In Lewis' (1984) terms, the 'intended' interpretation is the one that gets the semantic facts right.[8] Though the issue here is ultimately metaphysical—'do the non-semantic facts determine the semantic facts?'—it is viewed through an epistemological lens—'can knowledge of the non-semantic facts give rise to knowledge of the semantic facts?' The radical interpreter merely serves as a device of dramatization. Thus, the situation of the radical interpreter can be likened to that of a scientist who is omniscient of all possible observations of the universe, and must select the true theory of the unobservable universe from the set of empirically adequate theories without knowing in advance which theory is true. In both cases, to avoid underdetermination, there must be principles that bridge the known and the unknown. And in both cases, the initially unknown truths serve as an external condition of adequacy: just as the scientist can go wrong by selecting a false theory of the universe, so too can the interpreter go wrong by selecting a false interpretation of Karla.

What does it mean to say that the semantic truths supervene on the non-semantic truths? Lewis (1994) endorsed a minimal supervenience thesis, according to which, necessarily, any minimal duplicate of our world in relevant non-semantic respects is a semantic duplicate of it.[9] Lewis (1974) takes the class of non-semantic truths on which the semantic truths supervene to be the fundamental microphysical truths. Others hold that the semantic truths do not supervene on the physical truths alone, but on some expanded class of non-semantic truths, including for instance the phenomenal truths (Chalmers, 2006b; Horgan & Tienson, 2002; Pautz, 2013; Pitt, 2004). For present

---

[8] Strictly speaking, Lewis (1974) claims that these criteria, together with the non-semantic facts, entail a small class of interpretations. He suggests that this small class of interpretations models indeterminacy, for instance in vague expressions, such as 'thin', and in the 'confused desires of the compulsive thief' (Lewis 1974:343). However, there is a difficulty with modelling indeterminacy in this way, having to do with higher-order vagueness. If the vagueness of the expression 'thin' is modelled by a class of interpretations, each of which assigns a precise extension to 'thin', then if it is indeterminate whether it is indeterminate whether Tim is thin, this will need to be modelled by yet another class of classes of interpretations. And so on, for each order of vagueness. (I am grateful to an anonymous referee for this insight.) However, we need not assume that vagueness is best modelled in the way that Lewis proposes. For instance, instead of thinking that each interpretation assigns precise extensions to every predicate, and then allow for classes of interpretations to be true, one might suppose instead that the (uniquely true) intended interpretation assigns vague meanings to vague expressions, where vagueness is modelled in whatever turns out to be the best way to do so. Higher order vagueness can then be modelled within the intended interpretation. The central thought is that if the intended interpretation gets the semantic facts right, and some expressions do in fact have vague meanings, then the intended interpretation will assign vague meanings to those expressions. For these reasons, I will assume that there is a unique intended interpretation which accurately represents all of the semantic truths, including the indeterminacy of expressions and attitudes that are, in fact, indeterminate.

[9] This supervenience thesis can be articulated more precisely as follows: Let N be a sentence in a canonical language stating of all the non-semantic truths, let T be a 'that's all' statement to the effect that nothing more exists than is needed to satisfy N, and let S be a statement of the semantic truths. The minimal supervenience thesis is that it is metaphysically necessary that $NT \rightarrow S$. (This formulation follows Chalmers 1999; Chalmers 2012. For a variant, see Jackson 1998; Chalmers, 1999; Chalmers, 2012.)

By endorsing a minimal supervenience thesis, the physicalist is able to respond to objections premised on the metaphysical possibility of non-physical 'extras'. For instance, suppose that semantic physicalism is characterized as the view that any physical duplicate of our world is a semantic duplicate of it. Now consider a world, w, that is a physical duplicate of our world but contains non-physical extras, such as ghosts who are capable of thought. If the ghosts' thoughts have contents, then w is a physical duplicate of our world where the semantic truths differ. See Magidor and Kearns (2012) for several further cases along these lines.

purposes, we can simply allow the base to include all non-semantic truths. The crucial proviso is that any sentences that are included in the base be free of semantic terms or any terms that are equivalent in meaning to semantic terms, so as to avoid trivializing the supervenience thesis.

The central focus of this paper is the second premise of the a priori argument for Semantic Supervenience, which states that there *is* an a priori entailment from the non-semantic truths to the semantic truths. For Lewis (1974), to say that there is an a priori entailment here is just to say that an ideally rational being who is omniscient about the non-semantic truths and fully competent with the intentional concepts defined by folk psychology is in a position to deduce the semantic truths without recourse to any further empirical information (Cf. Chalmers and Jackson, 2002; Chalmers, 2012). Such a being is a radical interpreter. So, there is an a priori entailment from the non-semantic truths to the semantic truths if and only if radical interpretation is possible.

Lewis (1970, 1994) defends the a priori entailment thesis in general terms by appeal to analytic functionalism, according to which the theoretical postulates of folk psychology give rise to implicit analyses of mental concepts, such as the concepts of belief, desire, and meaning.[10] More specifically, in defense of the possibility of radical interpretation, Lewis argues that the implicit folk psychological definitions of belief, desire, and meaning, generate a priori criteria of interpretation choice, which in conjunction with the non-semantic truths, jointly entail the intended interpretation.[11] To begin with, Lewis puts forward six criteria, including, most notably, the *Principle of Charity* and the *Rationalization Principle* (Lewis, 1974). The Principle of Charity concerns the assignment of contents to Karla's beliefs: they ought to be assigned in such a way as to make Karla's beliefs rational in the light of her evidence. The Rationalization Principle says to assign credence and value functions in such a way as to satisfy the axioms of decision theory and assign the highest expected value to Karla's behavior.[12]

In his later work, Lewis realized that his original six criteria were insufficient to entail the intended interpretation. In 'Putnam's Paradox' (1984), Lewis considers the permutation problem raised by Putnam and others, that from any consistent interpretation x, it is possible to construct a competing interpretation y, that is equivalent to x in the assignment of truth values to sentences of the subject's 'global theory', but which assigns deviant extensions to lexical items. For instance, such a wayward interpretation might assign

---

[10] This kind of analytic functionalism about phenomenally conscious mental states has come under attack (Cf. Block, 1978; Chalmers, 1996; Jackson, 1982; Loar, 1990; Nagel, 1974). However, analytic functionalism about belief, desire and meaning has escaped relatively unscathed.

[11] Lewis (1974) calls these 'constraints' on interpretation, and suggests that the intended interpretation is the one that perfectly satisfies all of the constraints, i.e., that is perfectly charitable, perfectly rationalizing, and so forth. However, because ordinary humans are fallible and imperfectly rational, the intended interpretation, which describes Karla's state of mind with all its defects, is not perfectly charitable, perfectly rationalizing, and so forth. Thus, the intended interpretation does not perfectly match all of the constraints. It is more charitable therefore to treat them as criteria for the selection of interpretation, than as constraints.

[12] In addition, Lewis puts forward the *Truthfulness Principle*, which concerns the assignment of contents to Karla's utterances: it says that they should be assigned in such a way that Karla typically utters a sentence *s* of her language *L* only if she believes *s* true in *L*. The *Manifestation Principle* says that Karla's beliefs should be manifest in her dispositions to speech behaviour. The *Generativity Principle* requires that meaning assignments to sentences of Karla's language be finitely specifiable, reasonably uniform and simple, and the *Triangle Principle* says that assignments of contents to Karla's beliefs and desires remain fixed irrespective of the language in which they are assigned.

the property of being grue to 'green', where something is grue iff it is observed before some distant future time *t* and found to be green, or is blue otherwise (Goodman, 1990). As many philosophers have noticed, a deviant assignment in one place can be 'cancelled out' by deviant assignments elsewhere (Davidson, 1973; Field, 1978; Kripke 1982; Putnam, 1980; Quine, 1960; Williams 2007).

Lewis' solution to the permutation problem invokes *naturalness* (Lewis, 1983; 1984). Not all properties are created equal, he says; the property of being green is more natural than the property of being grue. Lewis therefore suggests a further Naturalness criterion for interpretation choice: all else being equal, if the properties interpretation *x* assigns to predicates are more natural than the properties *y* assigns to them, then *x* is to be preferred over *y*.

Others have suggested alternative criteria. For instance, Davidson (1974) is well-known for endorsing a Principle of (Alethic) Charity that involves maximizing truth in the subject's system of beliefs; Williamson (2004) has endorsed a Principle of Epistemic Charity, in place of both Lewisian and Davidsonian versions, which involves maximizing an agent's knowledge; and several others have suggested some kind of Causal Criterion for assignments of reference (Field, 1978; Hawthorne, 2007; McCarthy, 2002; Schiffer, 1978; Williams, 2007). Given the possible enlargement of the base to include phenomenal truths (Chalmers 2006b; Mendelovici 2018), one might wish to include criteria that pertain to the relations between conscious experiences and intentional mental states. For now, we can leave it open exactly what the criteria are.

3. The Framework

The radical interpreter needs a method for applying several criteria at once. Unfortunately, Lewis is not as explicit as one might like about this stage of the enterprise. He suggests three methods and indicates his preference among them (Lewis, 1974). But this is not good enough. There are many more logically possible methods than those Lewis considers, all of which are fully compatible with the non-semantic truths. Moreover, if the radical interpreter is to deduce the semantic truths from the non-semantic truths a priori, his choice of aggregation rule must be based on a priori considerations of some kind, since the non-semantic truths have all been accommodated by the individual criteria. Without an account of how to select an aggregation method, the choice of interpretation remains underdetermined by the non-semantic information and a priori resources alone.

As I have suggested, the situation of the radical interpreter can be fruitfully compared to that of a scientist selecting between rival, empirically adequate theories. Just as the scientist might appeal to criteria of theory choice, such as simplicity, scope, and fit, the radical interpreter appeals to criteria for interpretation choice, such as Charity, Rationalization, and Naturalness. And just as theory choice in the sciences can be fruitfully compared to social choice (Morreau, 2014; Morreau, 2015; Okasha, 2011), so too can radical interpretation. To be picturesque, we can imagine that the radical interpreter is aided by a committee, where each criterion is represented by an individual member of the committee with a one-track mind: Ms Charity cares only about charity, Ms Rationalization cares only about rationalizing an agent's behavior, Ms Naturalness cares only about naturalness, and so forth. Each of these individuals can be assumed to be omniscient about the non-semantic truths, and ranks the interpretations according to their

charity, rationality, naturalness and so forth. The radical interpreter's aim is to aggregate these individual rankings of interpretations into an overall ranking.

With this analogy in mind, we can use the machinery of social choice theory to model the situation of the radical interpreter. Let $X = \{x,y,z,\ldots\}$ be the set of all logically possible interpretations, and let $N = \{1,2,3,\ldots, n\}$ be the set of individual criteria—Charity, Rationalization, Naturalness, and so forth. Assume that each criterion determines a binary relation on X that is transitive and complete (a weak order), designated xR_iy, (e.g. 'interpretation x is at least as charitable as interpretation y'). From xR_iy we can define a strict relation,[13] designated xP_iy, (e.g. 'interpretation x is strictly more charitable than y'), and we can define a relation of indifference, designated xI_iy (e.g. 'x is equally as charitable as y').[14] Let's say that a profile is an n-tuple of criterion-relative individual orderings $<R_1, R_2,\ldots, R_n>$. The radical interpreter is concerned with the choice of a function F, from profiles of rankings by individual criteria $<R_1, R_2,\ldots, R_n>$ over interpretations in X, to an overall ranking of interpretations, xRy.

How are these rankings determined, more exactly? Consider, to begin with, Lewis' version of the Principle of Charity, which says to assign beliefs to Karla that are rational in light of her evidence according to some suitable inductive method (Lewis, 1979:534), such as Bayesian conditionalization (Lewis, 1983:374). We can start by characterizing what it is for an assignment to be perfectly charitable in a context[15] and then use that as a benchmark against which less charitable interpretations can be measured. Suppose that interpretation x assigns to Karla at time t_1 a prior probability distribution $\boldsymbol{P}_p$ over propositions that represents Karla's beliefs at t_1. If Karla acquires some evidence E between t_1 and t_2, x assigns a final probability distribution $\boldsymbol{P}_f$ at t_2 that represents Karla's beliefs at t_2, updated by the new information, E. We can say that x's assignment of $\boldsymbol{P}_f(B)$ to Karla at t_2 is perfectly charitable if and only if the final probability of B is equal to the prior probability of B given E, that is, if and only if $\boldsymbol{P}_f(B) = \boldsymbol{P}_p(B/E)$. With this in place, we can order interpretations of Karla at a context with respect to how far they deviate from one that is perfectly charitable, that is, by the difference between their assignments of $\boldsymbol{P}_f(B)$ and $\boldsymbol{P}_p(B/E)$ at that context: the greater the difference between $\boldsymbol{P}_f(B)$ and $\boldsymbol{P}_p(B/E)$ assigned by an interpretation at a context, the less charitable it is in its assignment at that context.[16] The overall charitability of an interpretation x is some suitable aggregate of the charitability of assignments that x makes at individual contexts. Of course, this means that Ms Charity faces an aggregation problem of her own. We can assume, for the sake of argument, that this aggregation problem can be resolved.[17,18]

---

13 xP_iy iff $xR_iy \wedge \neg yR_ix$.

14 xI_iy iff $xR_iy \wedge yR_ix$.

15 I set aside here any issues that may arise as a consequence of the contextual variability of the contents of certain expressions or sentences (Cf. Hawthorne 2007). The reference to a context is needed merely to isolate a particular set of token beliefs, desires and utterances with regard to which fragments of interpretations can be compared.

16 The difference can be measured either in absolute terms, or as a ratio.

17 Indeed, it is plausible to assume that when aggregating the overall charity of interpretations, based on measures of their charity at particular times, the individual rankings that serve as input to the aggregation function can be meaningfully compared, since they are all rankings by the same covering value—charity. If this is so, then Ms Charity may avoid Arrovian impossibility. See Sen (1970) on comparisons of well-being as a way to avoid Arrow's impossibility theorem.

18 Perhaps there are some interpretations that assign no probabilities at all. In this case, we can simply stipulate that they are maximally uncharitable.

Next, consider the Rationalization Principle, which says to assign to Karla credence and value functions, $\boldsymbol{C}$ and $\boldsymbol{V}$, that satisfy the axioms of decision theory and make her behavior out to maximize expected value. Once again, we can characterize more precisely what it is for an interpretation to rationalize behavior perfectly in its assignment at a given context, and then measure how far an interpretation deviates from the ideal. Suppose that Karla performs some act A at time t_1. An interpretation x perfectly rationalizes Karla's behavior at this context just in case it assigns credence and value functions such that the expected value of performing act A is at least as great as the expected value of performing any other act available to the agent at t_1, where an act's expected value is the probability-weighted average of the values of its possible outcomes. Once again, it is possible to measure the extent to which an interpretation rationalizes Karla's behavior at a context by the extent to which it deviates from an interpretation that perfectly rationalizes behavior at that context: the greater the difference between the expected value of A and the act that has the highest expected value given the assignments of $\boldsymbol{C}$ and $\boldsymbol{V}$, the less it rationalizes Karla's behavior in the context. Ms Rationalization then aggregates the rationality of $\boldsymbol{C}$ and $\boldsymbol{V}$ at each context into an overall measure of the rationality of x.

Finally, the Naturalness Principle pertains to the assignment of properties to predicates. According to Lewis, there is an elite class of perfectly natural properties, including, for instance, the fundamental physical properties. Properties which are not perfectly natural on this view can nevertheless be ordered by their degree of naturalness. For instance, neither the property of being green nor the property of being grue is perfectly natural, but the property of being green is more natural than the property of being grue. Lewis suggests that naturalness can be measured by 'length of definition' in a canonical language which contains exactly one predicate for each perfectly natural property. The thought is that the definition of 'grue' in such a canonical language is longer than the definition of 'green' because it contains more disjuncts, which reflects the fact that the members of the extension of 'grue' are more disparate than the members of the extension of 'green' in perfectly natural respects. Ms Naturalness then determines the overall naturalness of an interpretation as some kind of aggregate of the naturalness of its assignment of properties to predicates and concepts at each context.[19]

4. The Aggregation Problem

The radical interpreter and the social choice theorist are in a similar predicament—faced with a large number of possible choice functions, they must both find a way to narrow these down until only a few suitable functions remain. This, as Amartya Sen puts it, is 'something of an exercise in brinkmanship. . .In order to choose between the different possibilities through the use of discriminating axioms, we have to introduce *further* axioms, until only one possible procedure remains. . .We have to go on cutting down alternative possibilities, moving—implicitly—*towards* an impossibility, but then stop just before all possibilities are eliminated, to wit, when one and only one option remains.'

[19] One might worry that the ranking will be incomplete, since it is unclear how length of definition is to be measured in the case of definitions of infinite length (Sider 2011). I will set this worry aside here, since it is more charitable to assume that the rankings by individual interpretations are transitive and complete. If individual rankings are incomplete, the overall ranking of those interpretations that are not ranked by some criterion is indeterminate, and the radical interpreter will not know how to rank them.

(Sen, 1999:353-354) The trouble is that one constantly risks toppling into the abyss of impossibility.

The best-known impossibility theorem in social choice theory is Arrow's (1951;1963), which states that if $|X|>2$, there exists no social welfare function F satisfying the following constraints:

> **Ordering.** For any profile $<R_1,R_2,\ldots, R_n>$ in the domain of F, R is transitive and complete.
>
> **Weak Pareto.** For any profile $<R_1, R_2,\ldots, R_n>$ in the domain of F, if for all $i \in N$, xP_iy, then xPy.
>
> **Independence of Irrelevant Alternatives.** For any two profiles $<R_1, R_2, \ldots, R_n>$ and $<R^*_1,R^*_2,\ldots, R^*_n>$ in the domain of F and any $x,y \in X$, if for all $i \in N$, R_i's ranking of x and y coincides with R_i*'s ranking of x and y, then xRy iff xR^*y.
>
> **Non-Dictatorship.** There is no individual $i \in N$ such that, for all profiles $<R_1, R_2, \ldots, R_n>$ in the domain of F and all $x, y \in X$, if xP_iy then xPy.
>
> **Unrestricted Domain.** The domain of F is the set of all logically possible profiles.

Are these Arrovian constraints defensible in the context of radical interpretation? First, consider Ordering, which in the context of radical interpretation says that the radical interpreter's overall ordering R is transitive and complete. This seems plausible. First, if R is incomplete, then it could turn out to be the case that there is some pair of interpretations x and y, such that it is neither the case that x is at least as choice-worthy as y overall nor the case that y is at least as choice-worthy as x overall. Were such a situation to arise, the radical interpreter's choice between x and y would be under-determined by the non-semantic truths together with the a priori constraints on interpretation. If some interpretations are not ranked, the radical interpreter has no a priori basis for selecting an interpretation from the set of possible interpretations, and radical interpretation is impossible. Second, if R were intransitive, the selection of an interpretation would be unacceptably *ad hoc*. For instance, if xPy, yPz and zPx, then there is no interpretation that can be selected as best overall, since for each interpretation there is another that is strictly more choice-worthy than it.[20]

Second, consider Weak Pareto. When applied to radical interpretation, Weak Pareto says that if x is strictly more charitable, rationalizing, natural (etc.) than y, then x is strictly more choice-worthy than y overall. This constraint ensures that at least when the strict rankings of criteria are unanimous, the overall ranking cannot go against the consensus of individual criteria. Indeed, Weak Pareto can be viewed as a completeness constraint on the set of criteria. And this constraint is unassailable in the context of radical

[20] One might question whether it is really necessary for the radical interpreter to determine a binary relation R on interpretations that is transitive and complete, since all he needs to do is select a winner, and might do well enough with a rule that assigns a choice function to each profile. Such a rule would tell him which interpretation is best overall without saying anything about how suboptimal interpretations are ranked. Relaxing Ordering in this way makes no difference at all, however, since an impossibility theorem can be proven in this setting as well (See Kelly 1988). The remaining constraints are close cousins of the constraints given above.

interpretation. Since it is assumed that all of the non-semantic information has already been accounted for by the individual criteria, any aggregation rule that went against a consensus among the criteria in *N* would have to be justified by appeal to some further criterion, not included in *N*. But then, this further criterion could simply be added to *N*.[21]

Third, consider Independence of Irrelevant Alternatives. In the context of radical interpretation, this constraint says that whether *x* is at least as choice-worthy as *y* overall depends *exclusively* on how *x* and *y* are ranked with respect to Charity, Rationalization, Naturalness, and so forth. This seems prima facie plausible because any pairwise comparison should depend on the relevant qualities of the interpretations being compared—their charity, naturalness, rationality, and so forth—not on further considerations, such as how other interpretations are ranked.

It might be objected, however, that Independence of Irrelevant Alternatives also rules out taking account of information that might plausibly be deemed relevant, such as information about *how much* more charitable *x* is than *y*, or *how much* less natural *y* is than *x*. Indeed, Independence of Irrelevant Alternatives has been questioned in the context of social choice on the grounds that it unduly restricts the information available to social choice by disallowing inter-personal comparisons. Moreover, as Sen (1970) famously demonstrated, if we enrich the information available to social choice in such a way as to make interpersonal comparisons of well-being of certain kinds possible, then aggregation rules can be found which do satisfy suitably reformulated versions of Arrow's conditions. The imposition of this constraint on aggregation is further called into question by the fact that we often *do* make interpersonal comparisons of the relevant kinds, suggesting that such comparisons are possible.[22]

In contrast, in the context of radical interpretation, Independence of Irrelevant Alternatives seems to be on safer ground. For one thing, it is far from obvious how to make sense of inter-criterial comparisons of the choice-worthiness of interpretations. Suppose the radical interpreter knows that interpretation *x* is twice as charitable as *y*, but one third as natural. How should this information influence his overall ranking of *x* and *y*? It is difficult to say. The trouble is that we have no general principles concerning how charity and naturalness should be traded off against one another. Indeed, we don't even have a rough idea how these trade-offs are to be made. Moreover, our ordinary practice of interpretation offers little guidance here, since we do not in general make inter-criterial comparisons of the choice-worthiness of interpretations of this kind; we do not interpret one another on the basis of rich information about ratio or interval differences between the charity, naturalness, rationality, (etc.) of interpretations. Instead, we typically interpret one another on the basis of extensive, background *semantic* information—which of course is off limits to the radical interpreter. For these reasons, even if Independence of Irrelevant Alternatives may be questionable in the context of certain kinds of social choice, the grounds for its questionability appear to be absent in the context of radical interpretation.

Fourth, consider Non-Dictatorship. This constraint says that there should not be an individual constraint that unilaterally dictates which interpretation is best overall, irrespective of how interpretations are ranked by the other constraints. In the context of social choice, this constraint is defensible on the grounds of fairness. In the context of radical

---

[21] I am grateful to an anonymous referee for this line of defense of Weak Pareto in the context of radical interpretation.

[22] This point was raised by John Broome (personal communication).

interpretation, this constraint too is defensible, though on different grounds. In this context, if any criterion is a dictator, the interpretation choice function will deliver the wrong result.

For instance, it has been suggested that the naturalness criterion 'trumps' the rest (Sider, 2011). If this amounts to the claim that naturalness is a dictator, then if interpretation x is strictly more natural than y, it follows that x is strictly more choice-worthy than y overall, irrespective of how x and y are ranked by any other criterion. However, we can easily see that any F that coincides with the naturalness criterion will deliver the wrong result—it will not rank the intended interpretation as best overall. To illustrate, suppose that interpretation x is the intended interpretation of Karla, assigning the property of being green to 'green', the set of emeralds to 'emerald' and so forth. Now, suppose that y is a 'micro-world' interpretation (Hawthorne, 2007), assigning microscopic denizens of some distant part of the universe as referents of all names in Karla's language, and perfectly natural properties of those microscopic particles as extensions for all of the predicates, such as 'green', 'emerald', and so forth. The naturalness criterion ranks y above x. If naturalness is a dictator of F, F too ranks y above x overall. Since x is by hypothesis the intended interpretation, a radical interpreter who treats naturalness as a dictator is destined to select a false interpretation of Karla. This point generalizes to all other constraints. Since Karla's beliefs are not perfectly rational, the intended interpretation is not perfectly charitable, and Charity will rank an interpretation that is perfectly charitable as strictly more choice-worthy than the intended interpretation. If Charity is a dictator of F, then a perfectly charitable but false interpretation of Karla ranks above the intended interpretation. In general, because Karla is imperfect, for each criterion, there is some interpretation that is ranked above the intended interpretation by the lights of that criterion. For this reason, no criterion can be a dictator.

Finally, consider Unrestricted Domain. In the context of social choice, this constraint is motivated on the grounds that there are no prior constraints on voter's preferences—they can choose to rank candidates any way that they like. Voters' rankings are in this sense contingent. In the context of radical interpretation, it is somewhat difficult to determine whether all of the criteria are similarly contingent, not only because it has not been established exactly what the criteria are, but also because certain detailed questions concerning how the criteria are understood have implications for whether or not they are contingent. We can work through a few criteria, to see that some are straightforwardly contingent, while others are only contingent given certain assumptions.

First, consider the Alethic Charity principle put forward by Davidson. This criterion is plausibly contingent, since its rankings of interpretations depend on contingent features of the world: whether an interpretation of Karla maximizes truth in her system of beliefs depends to a large extent on what the contingent facts of her world are. For instance, consider a world w that differs from our world in non-semantic truths concerning the use of 'cat' and 'dog', where 'there is a cat' is typically used when there is a dog present and 'there is a dog' is typically used when there is a cat present. If we let x be an interpretation that assigns the set of all cats as the extension of 'dog', and y be an interpretation that assigns the set of all dogs as the extension of 'dog', we should expect that if Karla's world is like our world, then y is more charitable than x, but if Karla's world is like w, then x is more charitable than y. Since there is no restriction on what the

contingent non-semantic truths might be, there is no restriction on how Alethic Charity ranks interpretations across different possible worlds.

Is Lewisian Charity similarly contingent? In this case, the answer is not entirely straightforward. As suggested above, the Charity of an interpretation can be measured by the relation between its assignment of prior and final probabilities, given the evidence. Whether this measure delivers a ranking of interpretations that is necessary or contingent depends on how the evidence is characterized. If Karla's evidence is characterized in non-semantic terms—for instance, as a state with a certain phenomenology, or as a physical state—then Lewisian Charity is contingent, since the charity of an interpretation varies depending on contingent, non-semantic truths about Karla's evidence. In contrast, if the evidence is characterized in semantic terms—for instance, as an experience with a certain semantic content—then Lewisian Charity is necessary, since the charity of an interpretation depends on the values of three variables assigned by interpretations themselves: prior probability, final probability, and the semantic content of the evidence. Since these relations do not vary with contingent, non-semantic truths, rankings by Lewisian Charity are in this case necessary.[23]

Finally, consider the naturalness criterion.[24] Whether or not this criterion is contingent turns on whether interpretations assign one-dimensional or two-dimensional intensions to predicates. If interpretations assign one-dimensional intensions to predicates, then the naturalness criterion is necessary, because a one-dimensional intension assigns a set of actual and possible objects to a predicate, which does not vary with contingent features of the world. Furthermore, according to Lewis, the naturalness ordering of properties is necessary rather than contingent: it is necessary that the property of being green is more natural than the property of being grue. Thus, if interpretation *x* assigns the property of being green to a predicate, while interpretation *y* assigns the property of being grue, then all else being equal, it is necessary that *x* is a more natural interpretation than *y*.

However, if interpretations assign two-dimensional intensions, the naturalness criterion is plausibly contingent. Recall that a two-dimensional intension is a function from scenario-world pairs to truth values or extensions. If an interpretation assigns a two-dimensional intension to a predicate, the extension it assigns to that predicate can vary with contingent, non-semantic features of the world in which the agent finds herself. For instance, consider an interpretation *x* that assigns to 'water' a two-dimensional intension that delivers H_2O as the extension of 'water' if Karla's world is like our world, but XYZ as the extension of 'water' if Karla's world is like Putnam's (1975) Twin Earth, where the clear, colorless, odorless liquid that fills the lakes and streams is XYZ, rather than H_2O. If we suppose that H_2O is more natural than XYZ, the naturalness of *x* varies depending on contingent, non-semantic truths about the world: if Karla's world is like our world, her term 'water' picks out H_2O, whereas if her world is like Twin Earth, her term 'water' picks out XYZ (Cf. Chalmers, 2006b). Now, consider a comparison between interpretation *x* and interpretation *y* that assigns H_2O as the extension of 'water' whether or not Karla's world is like ours or like Twin Earth. In that case, arguably, if Karla's world is

[23] It is worth noting that Lewis does not provide any account of how to deduce the content of a subject's evidence from the non-semantic truths, which suggests that he takes the truths about the subject's evidence to be stated in non-semantic terms.

[24] I am grateful to an anonymous referee for pointing out that on a one-dimensional semantics, the naturalness criterion is necessary.

like our world, x is equally as natural as y, and if Karla's world is like Twin Earth, then y is more natural than x. If interpretations assign two-dimensional intensions, the naturalness principle is contingent.

Thus, it seems that given certain assumptions—that evidence is characterized in nonsemantic terms, and that interpretations assign two-dimensional intensions—it is plausible that Unrestricted Domain is met in the context of radical interpretation, at least on the assumption that the three criteria in play are Alethic Charity, Lewisian Charity, and Naturalness. Given these assumptions, Arrow's theorem applies to the context of radical interpretation.

However, it is controversial whether interpretations assign one-dimensional or two-dimensional intensions, and there is no consensus with regard to which criteria are to be relied upon. Neither of these matters can be settled here. Yet, since Unrestricted Domain is not satisfied if a single criterion is necessary, this leaves it unresolved whether Arrow's theorem is applicable in the present context.

5. Strong Neutrality and Richness

Fortunately, we need not leave the possibility of radical interpretation unresolved, since there are variants of Arrow's theorem which assume a weaker domain constraint. For instance, both Parks (1976) and Kemp and Ng (1976) independently put forward impossibility theorems where the domain consists of a single profile—the preferences that voters actually have. Morreau (2014, 2015) applied an impossibility theorem to the case of theory choice in the sciences which falls somewhere in between, allowing for some variation in the profiles in the domain of F, but not requiring that the domain is unrestricted.[25] And there is a strong case to be made that Morreau's impossibility theorem applies to the situation of the radical interpreter.

Like Arrow, Morreau assumes that individual orderings are transitive and complete. In addition, he assumes two Arrovian constraints: Weak Pareto and Non-Dictatorship.[26,27] However, in order to balance the weaker domain constraint, in place of Independence of Irrelevant Alternatives, he assumes Strong Neutrality:

> **Strong Neutrality**. For all profiles $\langle R_i \rangle_{i \in N}$ and $\langle R'_i \rangle_{i \in N}$ in the domain of F, and for all interpretations, w, x, y, and z in X: if for all i, xR_iy iff zR'_iw, and yR_ix iff wR'_iz then: xRy iff $zR'w$ and yRx iff $wR'z$.

Strong Neutrality requires that F treats no differently any two pairs of interpretations that are ranked in the same way by all criteria; it requires that F be insensitive to all information besides the individual rankings, including the nature or identities of the items that are ranked. In social choice, Strong Neutrality may seem implausibly strong, since it

---

25 For details, see Morreau, 2014; Morreau 2015. See also Hurley, 1985; Roberts, 1980.

26 Ordering is left in the background in Morreau, 2014.

27 It is worth noting that the motivation for Non-Dictatorship has been questioned in the single profile and restricted domain cases (Hylland, 1986; Morreau, 2016). The thought is that in a single profile case, the fact that one individual's strict preferences dictate the social ordering does not preclude the social ordering being democratic or fair. However, this objection does not carry over to the context of radical interpretation, where the motivation for Non-Dictatorship has to do with the fact that if the interpretation choice function were to preserve the strict rankings of any individual criterion, it would deliver the wrong result. This motivation for Non-Dictatorship holds equally in single profile and restricted profile cases.

rules out choice procedures that differ across contexts, such as for instance requiring a majority in one context and a super-majority in another. Yet it may well be appropriate to have different choice procedures in different contexts, since one context may be a vote on which color to paint the parlor, while the other may be a vote on which candidate should become president (Morreau, 2014:1264-1265). In social choice, the stakes can vary radically, making it sensible to choose aggregation rules that are more or less risk averse depending on what is at stake.[28]

In contrast, Strong Neutrality is more easily defended in the context of radical interpretation. First, there is no analogous variability in what is ranked—interpretations do not differ from one another in the way that paint colors differ from presidential candidates. Second, when it comes to radical interpretation, what is at stake is always the same: knowing what an agent believes, desires and means.[29] Third, if radical interpretation is possible, the radical interpreter's choice of interpretation *must* be based exclusively on the orderings by individual criteria, since these orderings accommodate all the relevant non-semantic information. So, Strong Neutrality seems distinctly more plausible in the context of radical interpretation than in the context of social choice.

In place of Unrestricted Domain, Morreau's theorem assumes that the domain of F is rich. Roughly put, a rich domain is one where for every suitable 'pattern' of ways in which three arbitrary interpretations may be ranked, there is a profile in the domain in which three interpretations are ranked in that way. More precisely, let α, β, and γ be three variables that can stand for interpretations in X. There are four ways in which these three variables may be weakly ordered, up to alphabetic variation:

(1) $\alpha = \beta = \gamma$,

(2) $\alpha = \beta < \gamma$,

(3) $\alpha < \beta = \gamma$,

(4) $\alpha < \beta < \gamma$.

A pattern is an n-tuple $\langle \mathcal{P}_i \rangle_{i \in N}$, where each $\mathcal{P}_i$ is one of the foregoing orderings of the variables. A pattern is suitable for a domain if there are as many orderings of interpretation variables as there are choice criteria in the profiles in the domain. For example, if there are two criteria, then each profile in the domain consists of a pair of orderings of interpretations, and suitable patterns are pairs of orderings of interpretation variables. A profile $\langle R_i \rangle_{i \in N}$, realizes a pattern $\langle \mathcal{P}_i \rangle_{i \in N}$ if there is a structure-preserving mapping from the variables in $\langle \mathcal{P}_i \rangle_{i \in N}$ to interpretations in X. We can now characterize richness as follows (Morreau 2014):

> **Richness**. For every suitable pattern $\mathcal{P}$ of three variables, there is some profile in the domain of F that realizes $\mathcal{P}$. (Morreau 2014:1264)

---

[28] Thanks to H. Orri Stefánsson for this point.

[29] Thanks to H. Orri Stefánsson for this point.

The impossibility result is this: suppose the set of criteria for the choice of interpretations is finite. No interpretation-choice rule with a rich domain satisfies Strong Neutrality, Weak Pareto and Non-Dictatorship.[30]

As we have already seen, there is a strong case to be made that in the context of radical interpretation, Weak Pareto, Strong Neutrality, and Non-Dictatorship are plausible constraints on aggregation. Moreover, it is possible to show that the domain of F is rich, at least in a toy case involving two criteria. Intuitively, this can be seen by noting that the two rankings are independent of one another. Indeed, richness can be shown even in cases of criteria that are logically related, and hence seem at first blush to be dependent. For instance, since knowledge entails truth, it may seem as though Epistemic Charity and Alethic Charity cannot rank interpretations in any which way, and thus fail to realize some patterns.[31] However, closer inspection suggests that this is not the case. Intuitively, the reason is that though knowledge entails truth, ignorance can involve either false beliefs or true beliefs that are unconfirmed by the agent's evidence. This means that the number of true beliefs an interpretation assigns to an agent can vary independently of the number of known beliefs it assigns.

To illustrate, suppose that for $1 \leq i \leq 4$, interpretations a_i, b_i and c_i assign no knowledge to Karla whatsoever. Every belief that each of these interpretations assigns to Karla is either false or unconfirmed by the evidence, so Epistemic Charity ranks them $a_i = b_i = c_i$. However, it is not necessarily the case that every belief assigned to Karla is *both* false and unconfirmed by the evidence. Indeed, suppose that some of these interpretations assign true beliefs to Karla that are not supported by Karla's evidence—let's call these unknown true beliefs. We can then see that though Epistemic Charity may rank these interpretations equally, Alethic Charity can rank them in any which way: If a_1, b_1, and c_1 assign no true beliefs to Karla, Alethic Charity ranks them $a_1 = b_1 = c_1$. If c_2 assigns n unknown true beliefs, while a_2 and b_2 assign no true beliefs, Alethic Charity ranks them $a_2 = b_2 < c_2$. If b_3 and c_3 assign n unknown true beliefs, while a_3 assigns no true beliefs, Alethic Charity ranks them $a_3 < b_3 = c_3$. If a_4 assigns no true beliefs, b_4 assigns n unknown true beliefs, and c_4 assigns $n+1$ unknown true beliefs, Alethic Charity ranks them $a_4 < b_4 < c_4$. That is, for $1 \leq i \leq 4$, Epistemic Charity ranks interpretations $a_i = b_i = c_i$, while Alethic Charity ranks them as follows:

(AC1) $a_1 = b_1 = c_1$,

(AC2) $a_2 = b_2 < c_2$,

(AC3) $a_3 < b_3 = c_3$,

(AC4) $a_4 < b_4 < c_4$.

Thus, it is at least plausible that the domain **D** of F is rich. (For a proof that **D** is rich in this toy case, see 'Appendix'.) Though it is beyond the scope of this paper to show that **D** is rich for any set of plausible criteria, the foregoing suggests that it is, since the rankings induced even by criteria that are logically related can be shown to be independent

[30] For a proof, see Morreau, 2014, 'Appendix'.

[31] I am grateful to Gustaf Arrhenius and an anonymous referee for pointing this out.

from one another, and since the set of interpretations to be ranked includes all logically possible interpretations. Yet, if **D** is rich, and if the foregoing constraints on interpretation choice hold, then radical interpretation is impossible.[32]

6. Conclusion

As we have seen, the radical interpreter faces an intractable problem when attempting to aggregate the rankings supplied by individual criteria of interpretation choice: there is no function F that meets a small set of plausible constraints on aggregation, whether these are Arrow's original constraints, or those applicable in a setting in which the domain of F is restricted. This means that even if the radical interpreter is omniscient about the semantic truths, and fully cognizant of the rankings of interpretations determined by each criterion of interpretation choice, she cannot deduce the intended interpretation without recourse to further semantic information. In other words, radical interpretation is impossible. And if radical interpretation is impossible, then the a priori argument for Semantic Supervenience fails.[33]

Appendix

In a two-criterion case involving only Epistemic Charity and Alethic Charity, the domain **D** is rich just in case sixteen suitable patterns are realized:

I. $(\alpha = \beta = \gamma,\ \alpha = \beta = \gamma)$,

II. $(\alpha = \beta = \gamma,\ \alpha = \beta < \gamma)$,

III. $(\alpha = \beta = \gamma,\ \alpha < \beta = \gamma)$,

IV. $(\alpha = \beta = \gamma,\ \alpha < \beta < \gamma)$,

V. $(\alpha = \beta < \gamma,\ \alpha = \beta = \gamma)$,

---

[32] Of course, though it is plausible that the domain will remain rich even with the addition of further criteria, its richness is not guaranteed. This suggests a route to a viable Lewisian meta-semantics: it would have to use criteria for which there is no rich domain of interpretations. However, such a route seems prima facie unattractive. As the toy case involving Epistemic and Alethic Charity has shown, **D** is rich even when two criteria are logically related. For a Lewisian metasemantics to be viable, then, there must be additional criteria that are so closely related that they rule out certain patterns of rankings of interpretations. One worry that might arise on such an approach is that the inclusion of such closely related criteria would amount to an unacceptable 'double counting', as there would be if the profile included two Alethic Charity criteria. Thus, even if this route to a Lewisian metasemantics may remain open, it is not obvious that it is truly viable.

[33] This paper builds on a suggestion that was sketched in Hattiangadi 2015. Earlier versions of this paper have been presented at Stockholm University, Oslo University, Umeå University, the Swedish Collegium of Advanced Studies, Uppsala University, and at a workshop on radical interpretation at Leeds University organized by Robbie Williams. I have benefitted tremendously from conversations with the audiences at all of these sessions. I am especially grateful to Gustaf Arrhenius, John Broome, Krister Bykvist, David Chalmers, Christian List, Wlodek Rabinowicz, H. Orri Stefánsson, and Robbie Williams, all of whom read and commented on earlier drafts. Special thanks go to H. Orri Stefánsson for assistance with the proof that **D** is rich in the two-criterion case. The paper was improved immeasurably by the thorough, helpful, and detailed comments provided by an anonymous referee for this journal.

VI. ($\alpha = \beta < \gamma$, $\alpha = \beta < \gamma$),

VII. ($\alpha = \beta < \gamma$, $\alpha < \beta = \gamma$),

VIII. ($\alpha = \beta < \gamma$, $\alpha < \beta < \gamma$),

⋮

We can see that these patterns are realized as follows.

First, suppose that for $1 \leq i \leq 4$, a_i, b_i and c_i assign no knowledge to Karla whatsoever, so Epistemic Charity ranks them:

($EC_{1\text{-}4}$) $a_i = b_i = c_i$.

However Alethic Charity ranks them as follows:

(AC_1) a_1, b_1, and c_1 assign no true beliefs to Karla, so $a_1 = b_1 = c_1$.

(AC_2) c_2 assigns some unknown true beliefs, while a_2 and b_2 assign no true beliefs, so $a_2 = b_2 < c_2$.

(AC_3) b_3 and c_3 assign n unknown true beliefs, while a_3 assigns no true beliefs, so $a_3 < b_3 = c_3$.

(AC_4) a_4 assigns no true beliefs, b_4 assigns n unknown true beliefs, and c_4 assigns n +1 unknown true beliefs, so $a_4 < b_4 < c_4$.

Second, suppose that for $5 \leq i \leq 8$, a_i and b_i assign no knowledge to Karla, but c_i assigns n known beliefs to her, so Epistemic Charity ranks them:

($EC_{5\text{-}8}$) $a_i = b_i < c_i$.

Since knowledge entails truth, it follows that for $5 \leq i \leq 8$, c_i assigns at least n true beliefs to Karla. With this in place, suppose that Alethic Charity ranks them as follows:

(AC_5) a_5 and b_5 assign n unknown true beliefs, so $a_5 = b_5 = c_5$.

(AC_6) a_6 and b_6 assign no true beliefs, so $a_6 = b_6 < c_6$.

(AC_7) a_7 assigns no true beliefs, and b_7 assigns n unknown true beliefs, so $a_7 < b_7 = c_7$.

(AC_8) a_8 assigns $n - 2$ unknown true beliefs, and b_8 assigns $n - 1$ unknown true beliefs, so $a_8 < b_8 < c_8$.

Third, suppose that for $9 \leq i \leq 12$, a_i assigns no known beliefs to Karla, while b_i and c_i assign her n known beliefs, so Epistemic Charity ranks them:

($EC_{9\text{-}12}$) $a_i < b_i = c_i$.

Since knowledge entails truth, it follows that for $5 \leq i \leq 8$, b_i and c_i assign at least n truths, and suppose Alethic Charity ranks these interpretations as follows:

(AC_9) a_9 assigns n unknown true beliefs, so $a_9 = b_9 = c_9$.

(AC_{10}) a_{10} assigns n unknown true beliefs, and b_{10} assigns 1 unknown true belief in addition to n known true beliefs, so $a_{10} = b_{10} < c_1$.

(AC_{11}) a_{11} assigns no unknown true beliefs, so $a_{11} < b_{11} = c_{11}$.

(AC_{12}) a_{12} assigns no unknown true beliefs, and c_{12} assigns 1 unknown true belief, so $a_{12} < b_{12} < c_{12}$.

Fourth, suppose that for $13 \leq i \leq 16$, a_i assigns no knowledge to Karla, b_i assigns her n known beliefs, and c_i assigns her $n + 1$ known beliefs, so Epistemic Charity ranks them:

($EC_{13\text{-}16}$) $a_i < b_i < c_i$

Once again, since knowledge entails truth, it follows that for $13 \leq i \leq 16$, b_i assign n true beliefs, and c_i assign n +1 true beliefs. However suppose that Alethic Charity ranks them as follows:

(AC_{13}) a_{13} assign $n + 1$ unknown true beliefs, b_{13} assigns 1 unknown true belief, so $a_{13} = b_{13} = c_{13}$.

(AC_{14}) a_{14} assigns n unknown true beliefs, so $a_{14} = b_{14} < c_{14}$.

(AC_{15}) a_{15} assigns no true beliefs to Karla, and b_{15} assigns 1 unknown true belief to her, so $a_{15} < b_{15} = c_{15}$.

(AC_{16}) a_{16}, b_{16}, and c_{16} assign no unknown true beliefs to Karla, so $a_{16} < b_{16} < c_{16}$.

Now, call the set containing all of the rankings by Epistemic Charity **EC**, and the set containing all of the rankings by Alethic Charity **AC**. The relevant domain **D** is the set of all ordered pairs, or profiles, $\langle a, b \rangle$ such that $a \in \mathbf{EC}$, $b \in \mathbf{AC}$. We can see that all 16 patterns given above are realized, and **D** is rich:

($a_1 = b_1 = c_1$, $a_1 = b_1 = c_1$) realizes I,

($a_2 = b_2 = c_2$, $a_2 = b_2 < c_2$) realizes II,

($a_3 = b_3 = c_3$, $a_3 < b_3 = c_3$) realizes III,

($a_4 = b_4 = c_4$, $a_4 < b_4 < c_4$) realizes IV,

($a_5 = b_5 < c_5$, $a_5 = b_5 = c_5$), realizes V,

⋮

Works Cited

Arrow, K. 1951/1963. *Social Choice and Individual Values*. New York: Wiley.

Block, N. 1978. 'Troubles with Functionalism.' *Minnesota Studies in the Philosophy of Science* 9:261–325.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

—— 1999. 'Materialism and the Metaphysics of Modality.' *Philosophy and Phenomenological Research*, 59(2):473–496.

—— 2006a. 'The Foundations of Two Dimensional Semantics.' M. Garcia-Carpintero & J. Macia, eds. Oxford: Oxford University Press: 55–140.

—— 2006b. 'Perception and the Fall from Eden.' In T. Gendler & J. Hawthorne, eds. *Perceptual Experience*. Oxford: Oxford University Press, pp. 49–125.

—— 2012. *Constructing the World*. Oxford: Oxford University Press.

—— & Jackson, F. 2002. 'Conceptual Analysis and Reductive Explanation.' *The Philosophical Review*, 110(3):315–360.

Davidson, D. 1973. 'Radical Interpretation.' *Dialectica* 27(3/4): 313–328.

Field, H. H. 1978. 'Mental Representation.' *Erkenntnis* 13:9–61. Reprinted in H.H. Field, *Truth and the Absence of Fact*. Oxford: Oxford University Press, 2001, pp. 30–67.

Goodman, N. 1990. *Fact, Fiction and Forecast*. Cambridge, Mass.: Harvard University Press.

Hale, B., and Wright, C. 1997. 'Putnam's model-Theoretic Argument against Metaphysical Realism.' In C. Wright and B. Hale, eds., *A Companion to the Philosophy of Language*. Oxford: Blackwell, pp. 427–57.

Hattiangadi, A. 2015. 'Metasemantics out of Economics?' In A. Reisner and I. Hirose eds. *Weighing and Reasoning: A Festschrift for John Broome*. Oxford: Oxford University Press, pp. 52–60.

Hawthorne, J. 2007. 'Craziness and Metasemantics.' *The Philosophical Review*, 116(3): 427–440.

Horgan, T. and Tienson J. 2002. 'The Intentionality of Phenomenology and the Phenomenology of Intentionality.' In *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers. Oxford: Oxford University Press.

Hurley, S. 1985. 'Supervenience and the Possibility of Coherence.' *Mind* 94: 501–525.

—— 1998. *From Metaphysics to Ethics: A Defence of Coneptual Analysis*. Oxford: Oxford University Press.

Hylland, A. 1986. 'The Purpose and Significance of Social Choice Theory: some general remarks and an application to the "Lady Chatterley problem".' *Elster and Hylland* 1986: 45–73.

Jackson, F. 1982. 'Epiphenomenal Qualia.' *Philosophical Quarterly* 32:127–136.

—— 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Kaplan, D. 1989a. 'Demonstratives.' In*Themes from Kaplan*. Oxford: Oxford University Press, pp. 481–563.

—— 1989b. 'Afterthoughts.' In*Themes from Kaplan*. Oxford: Oxford University Press, pp. 565–612.

Kelly, J. 1988. *Social Choice Theory: An Introduction*. Berlin: Springer-Verlag.

Kemp, M. C. & Ng, Y.-K. 1976. 'On the Existence of Social Welfare Functions, Social Orderings and Social Decision Functions.' *Economica New Series*, 43(169): 59–66.

Lewis, D. 1970. 'How to Define Theoretical Terms.' *Journal of Philosophy* 67:427–46.

—— 1974. 'Radical Interpretation.' *Synthese* 23:331–344.
—— 1979. 'Attitudes *de dicto* and *de se*.' *The Philosophical Review*, 88(4):513–543.
—— 1983. 'New Work for a Theory of Universals.' *Australasian Journal of Philosophy* 61: 343–77.
—— 1984. 'Putnam's Paradox.' *Australasian Journal of Philosophy*. 62(3): 221–236.
—— 1994. 'Reduction of Mind.' In *A Companion to the Philosophy of Mind*, ed. S. Guttenplan. Oxford: Blackwell, pp. 412–31.
List, C. 2001. 'A Note on Introducing a 'Zero-Line' of Welfare as an Escape-Route from Arrow's Theorem.' Pacific Economic Review, 6, special section in honour of Amartya Sen, pp. 223-232.
Loar, B. 1990. 'Phenomenal States.' In N. Block, O. Flanagan and G. Guzeldere, eds. *The Nature of Concsiousness*. Cambridge, MA: MIT Press.
Magidor, O. And Kearns S. 'Semantic Sovereignty.' *Philosophy and Phenomenal Research*, 85(2):322–350.
McCarthy, T. 2002. *Radical Interpretation and Indeterminacy*. Oxford: Oxford University Press.
Mendelovici, A. 2018. *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.
Morreau, M. 2010. 'It Simply Does Not Add Up: Trouble with Overall Similarity.' *Journal of Philosophy*, 107:469–490.
—— 2014. 'Mr. Fit, Mr. Simplicity, Mr. Scope: from Social Choice Theory to Theory Choice.' *Erkenntnis* 79:1253–1268.
—— 2015. 'Theory Choice and Social Choice: Kuhn Vindicated.' *Mind*, 124(493):239–262.
—— 2016. 'Arrow's Theorem*', The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/arrows-theorem/>.
Nagel, T. 1974. 'What is it Like to be a Bat?' *Philosophical Review* 83:435–50.
Okasha, S. 2009. 'Individuals, Groups, Fitness and Utility: Multi-Level Selection Meets Social Choice Theory.' *Biology and Philosophy*, 24:561–584.
—— 2011. 'Theory Choice and Social Choice: Kuhn versus Arrow.' *Mind*, 120:83–115.
Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.
Parks, R. P. 1976. 'An Impossibility Theorem for Fixed Preferences: A Dictatorial Bergson-Samuelson Welfare Function.' *The Review of Economic Studies* 43(3):447–450.
Pautz, A. 2013. 'Does Phenomenology Ground Mental Content?' In Kriegel, ed. *Phenomenal Intentionality*. Oxford: Oxford University Press, pp. 194–234.
Pitt, D. 2004. 'The Phenomenology of Cognition, Or, What is it Like to Think that P?' *Philosophy and Phenomenological Research* 69: 1–36.
Putnam, H. 1975. 'The Meaning of Meaning.' *Philosophical Papers, Vol. II: Mind, Language, and Reality*. Cambridge: Cambridge Univeresity Press.
—— 1980. 'Models and Reality.' *Journal of Symbolic Logic* 45: 421–444.
Quine, W. V. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Roberts, Kevin 1980: 'Social Choice Theory: The Single-profile and Multi-profile Approaches.' *Review of Economic Studies*, 47: 441–50.
Schiffer, S. 1978. 'The Basis of Reference.' *Erkenntnis* 13: 171–206.
Sen, A. 1970. *Collective Choice and Social Welfare*. San Fransisco: Holden-Day.

—— 1999. 'The Possibility of Social Choice.' *The American Economic Review*, 89 (3):349–378.

Sider, T. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Stegenga, J. 2013. 'An Impossibility Theorem for Amalgamating Evidence.' *Synthese* 190: 2391–2411.

Williams, J. R. G. 2007. 'Eligibility and Inscrutability.' *The Philosophical Review*, 116 (3): 361–399.

Williamson, T. 2004. 'Philosophical 'Intuitions' and Scepticism about Judgment.' *Dialectica* 58(1): 109–153.