Journal of Philosophical Logic (2024) 53:1321–1346
https://doi.org/10.1007/s10992-024-09767-2

# Carnap's Problem, Definability and Compositionality

**Pedro del Valle-Inclán**[1]

Received: 7 December 2023 / Accepted: 12 June 2024 / Published online: 26 June 2024


**Abstract**
In his *Formalization of Logic* (1943) Carnap pointed out that there are non-normal interpretations of classical logic: non-standard interpretations of the connectives and quantifiers that are consistent with the classical consequence relation of a language. Different ways around the problem have been proposed. In a recent paper, Bonnay and Westerståhl argue that the key to a solution is imposing restrictions on the type of interpretation we take into account. More precisely, they claim that if we restrict attention to interpretations that are (a) compositional, (b) non-trivial and (c) in the case of the quantifiers, invariant under permutations of the domain, Carnap's Problem is avoided. This paper has two goals. The first is to show that Bonnay and Westerståhl's solution to Carnap's Problem doesn't work. The second is to argue that something similar to their proposal seems to do the job. The problems with Bonnay and Westerståhl's approach trace back to issues concerning the (un)definability of subsets of the domain of first-order structures, as well as to the compositionality of first-order languages. After expanding on these problems, I'll propose a way to modify Bonnay and Westerståhl's account and solve Carnap's Problem.

**Keywords** Inferentialism · Categoricity · Carnap's Problem · Compositionality · Definability · Meta-Semantics

## 1 Introduction

It's a platitude to say that linguistic signs are conventional, that they mean what they do because speakers use them (or have used them) in a certain way. There is nothing wrong with the idea, as far as it goes, but it doesn't go far. 'Use determines meaning', understood this way, says too little to be useful.

Inferentialism is one way to flesh out the slogan. It claims that the meaning of words is determined by the way they are used *in inferences*. This is not a popular view, at least if applied to language at large, but it has found more acceptance when restricted to

✉ Pedro del Valle-Inclán
pedro.delvalleinclan@sns.it

[1] Department of Philosopy, Scuola Normale Superiore, Pisa, Italy

logical vocabulary. Words like 'or', 'some' and 'not' play a central role in arguments. And conversely, the way we use these words in arguments seems especially telling about what they mean. Inferentialists think this connection is tight enough for us to base a theory of meaning on it.[1]

There are different ways to be an inferentialist. Hard-line versions of the position –see [6, 20]– characterise meanings directly in terms of inference rules. To know the meaning of 'and', on these accounts, is simply to know that one can make certain inferences involving it (say, infer '$p$' and '$q$' from '$p$ and $q$' and vice versa). This type of inferentialism is fairly radical: it does without, or at least plays down, the usual notions of reference, truth-conditions and the like. A moderate form of inferentialism is favoured by Hacking [8], Garson [7] or Murzi and Topey [16]. Moderate inferentialists don't identify meanings with inference rules, but still claim that the meaning of logical vocabulary can be, in some sense, read off its role in inference.

What exactly is a 'role in inference', though? For instance, what is the role in inference of the classical connectives and quantifiers? Here's an idea: take some language $\mathcal{L}$ and consider the classical consequence relation $\vdash_{\mathcal{L}}$ on it.[2] $\vdash_{\mathcal{L}}$ gives full information about which sentences logically follow from which, alone or in combination with others, so we could take it to encode the inferential role of the logical vocabulary of $\mathcal{L}$. The idea is natural and straightforward, but a well-known result by Carnap weighs against it. In his [5] Carnap pointed out that there are 'non-normal' interpretations of classical logic: deviant interpretations of the connectives and quantifiers that are consistent with the classical consequence relation. In other words, if roles in inference are given by $\vdash_{\mathcal{L}}$, and meanings are what Carnap calls interpretations, then the meaning of classical logical vocabulary is *not* fixed by its inferential role. The situation is similar to the case of arithmetic: the Peano axioms have non-standard models, and so does, in a sense, classical logic itself. Let's call this Carnap's Problem.

Philosophers have come up with several workarounds to Carnap's Problem. Some propose to use multiple conclusion consequence relations [23], or bilateral ones [21, 24]. Others hold that inferential roles are given by the *rules* governing logical vocabulary, rather than by consequence relations [7]. Here I'll focus on a third kind of approach, recently put forward by Bonnay and Westerståhl [3]. According to Bonnay and Westerståhl some assignments of meanings to the expressions of a language can be ruled out from the get go. The 'space of possible interpretations' as they put it, is 'a priori restricted by universal semantic principles' [3, p. 721]. More precisely, they claim that if we restrict attention to interpretations that are (a) compositional, (b) non-trivial and (c) in the case of the quantifiers, invariant under permutations of the domain, Carnap's Problem is avoided.

This paper has two goals. The first is to show that Bonnay and Westerståhl's solution to Carnap's Problem doesn't work. The second is to argue that something similar to their proposal seems to do the job.

[1] From now on I'll use 'inferentialism' to mean inferentialism *as applied to logical vocabulary* (what's sometimes known as logical inferentialism). 'Logical vocabulary', here and in what follows, is short for 'connectives and quantifiers'.

[2] Where $\vdash_{\mathcal{L}}$ is understood 'syntactically', as a relation between sets of sentences and sentences.

There are two problems with Bonnay and Westerståhl's approach. The first concerns the main result of their paper, a characterisation of the interpretations of the universal quantifier ∀ that are consistent with the classical consequence relation of a language. I will show that their characterisation holds only for *second*-order languages, and not for first-order ones; for that reason their strategy doesn't fix the normal interpretation of first-order quantifiers. The underlying problem has to do with the (un)definability of subsets of the domain of a structure. It's well-known that given a first-order language and a structure for it there are, in general, subsets of the domain that can't be defined by any formula. This can be exploited to construct non-normal interpretations, and makes Carnap's Problem for first-order languages particularly challenging.

The second problem with Bonnay and Westerståhl's approach concerns the way they define interpretations. Although they hold that we should restrict attention to compositional interpretations, their normal interpretations of first-order languages aren't compositional after all. I will also show that if we redefine interpretations to avoid this problem, compositionality, non-triviality and invariance under permutations don't pin down the standard meaning of logical vocabulary. In this case the underlying problem is Bonnay and Westerståhl's demand for compositionality itself. The usual semantics for first-order logic is *not* compositional. While different compositional semantics for first-order languages are available, they all involve a large range of semantic values. This, in its turn, makes Carnap's Problem all the more difficult to solve: more semantic values means more possible interpretations, and therefore more non-normal ones that may be consistent with a given consequence relation. In the first-order case, then, demanding compositionality is counter-productive.

After expanding on these problems I'll propose a way to modify Bonnay and Westerståhl's solution that avoids them. Roughly put, I'll argue that there are plausible reasons to strengthen the notion of an interpretation being consistent with a consequence relation, and that this is enough to clinch the usual interpretation of classical logical vocabulary.

Here's the plan: Section 2 sets the stage by introducing Carnap's Problem for propositional languages and Bonnay and Westerståhl's solution to it. Section 3 looks at the interplay between Carnap's Problem and definability, and Section 4 at Carnap's Problem and compositionality. Section 5 then presents an alternative to Bonnay and Westerståhl's solution, Section 6 looks at different ways of defining interpretations, and Section 7 takes stock and concludes.

## 2 Carnap's Problem and Propositional Logic

Let's begin by sorting out the terminology. Take a propositional language $\mathcal{L}$ built up from atoms $p, q, r, \ldots$ and connectives in the usual way. A *valuation* for $\mathcal{L}$ is an assignment of truth-values 1 or 0 to all sentences, i.e. a function from $\mathcal{L}$-sentences to $\{1,0\}$. A *consequence relation* $\vdash$ for $\mathcal{L}$ is a binary relation between sets of $\mathcal{L}$-sentences and $\mathcal{L}$-sentences. A valuation $v$ is *consistent* with a consequence relation $\vdash$ if it makes all valid arguments truth-preserving: if $\Gamma \vdash \varphi$ and $v(\Gamma) = 1$, then $v(\varphi) = 1$.[3] A

[3] Here and in what follows '$v(\Gamma) = 1$' is short for $v(\gamma) = 1$ for all $\gamma \in \Gamma$.

valuation is $\diamond-$*normal* for a connective $\diamond$ if it interprets $\diamond$ in the intended way. $\wedge$-normal valuations, for instance, are such that $v(\varphi \wedge \psi) = 1$ iff $v(\varphi) = v(\psi) = 1$, and $\neg$-normal valuations are such that $v(\neg\varphi) = 1$ iff $v(\varphi) = 0$. Finally, a valuation is *normal* if it is $\diamond-$*normal* for all connectives $\diamond$ of $\mathcal{L}$.

What Carnap [5] pointed out is that some non-normal valuations are consistent with the classical consequence relation of any (truth-functionally complete) language. Say for concreteness that $\mathcal{L}$ has connectives $\wedge$, $\vee$, $\neg$ and $\rightarrow$, and consider the valuation $v_T$ such that $v_T(\varphi) = 1$ for all sentences $\varphi$. $v_T$ is consistent with the classical consequence relation $\vdash_{\mathcal{L}}$ of $\mathcal{L}$, but not $\neg$-normal. Or take $v_c$, the valuation such that $v_c(\varphi) = 1$ iff $\varphi$ is a classical tautology. It is easy to check that $v_c$ is normal only for $\wedge$ but, once again, consistent with $\vdash_{\mathcal{L}}$.

According to Bonnay and Westerståhl the key to a solution is imposing restrictions on the type of valuation we take into account. They begin by noting that 'as speakers, we know that our language [...] is going to have some true and some false sentences' [3, p. 733]. This means that the trivial valuation $v_T$ such that $v_T(\varphi) = 1$ for all $\varphi$ is ruled out from the start, by means of a 'semantic a priori'.

The second restriction they impose is compositionality. The usual, fast and loose formulation of the principle of compositionality says that the meaning of a complex expression depends on the meanings of its parts and the syntactic rules used to put them together. This formulation can be made more precise in a number of non-equivalent ways.[4] Bonnay and Westerståhl go for the following one: an assignment $\mu$ of semantic values to the expressions of some language is compositional if:

> **(PC):** For every n-ary syntactic rule $\mathcal{O}$ there is a semantic composition function $\mathrm{F}_{\mathcal{O}}$ such that for any well-formed expression $\mathcal{O}(e_1, ...e_n)$ we have $\mu(\mathcal{O}(e_1, ...e_n)) = F_{\mathcal{O}}(\mu(e_1), ...\mu(e_n))$.

In the case of propositional logic, where $\mu$ is a valuation $v$, **(PC)** just says that $v$ must treat each connective as a truth-function.

These two requirements, compositionality and non-triviality, are enough to rule out all non-normal interpretations of propositional languages. In fact, compositionality alone gets rid of all valuations that are non-normal for $\wedge$, $\vee$ or $\rightarrow$. The requirement of non-triviality is only needed to get rid of the valuation $v_T$, which is compositional and normal for all our connectives except $\neg$. Let's make this into a theorem, for future reference:

**Theorem 1** *Let $\mathcal{L}$ be a propositional language and $v$ a valuation for $\mathcal{L}$. If $v$ is compositional, non-trivial, and consistent with $\vdash_{\mathcal{L}}$, then $v$ is normal.*

***Proof*** See [3]. □

The situation can be summarised as follows: all non-normal valuations for propositional logic (except $v_T$) are non-compositional. Therefore, demanding compositionality and banning $v_T$ leaves us with all and only the normal valuations: Carnap's Problem is avoided. Now let's move on to the first-order case.

[4] See [17] for a detailed discussion and further references.

## 3 Carnap's Problem and Definability

Formulating Carnap's Problem for propositional languages is straightforward. Doing it for first-order languages is more delicate, and it will pay off to be cautious. Let's begin by being very clear about the languages at play. I'll assume that they have $\wedge$, $\neg$ and $\forall$ as logical vocabulary, but this is only for concreteness; any other signature would do. As for non-logical vocabulary, all languages will have infinitely many individual variables $x_1$, ..., $x_n$, ... and a stock of predicate symbols $P$, $R$, ... of finite arity. They may or may not contain, in addition, individual constants $c_1$, ...$c_n$, ... and function symbols $f_1$, ..., $f_n$, ..., also of finite arity.

Now, in order to state Carnap's Problem we have to do several things. First we have to define interpretations, and agree on what the normal and non-normal interpretations of first-order languages are. Then we have to explain what it is for an interpretation to be consistent with a consequence relation, and in particular with the *classical* consequence relation of a language. And, finally, since we're going to formulate Carnap's Problem the way Bonnay and Westerståhl do, we also have to consider what compositionality, non-triviality, and invariance under permutations amount to in this setting.

Let's begin with interpretations. When we talk about semantics for first-order logic, what comes to mind is the Tarskian definition of satisfaction of a formula $\varphi$ under a variable assignment $\sigma$ in a structure $\mathcal{M} = (D, I)$ –where $D$ is a domain and $I$ a function interpreting the non-logical vocabulary. Given such an $\mathcal{M}$, the usual clauses for satisfaction are:

(i) $\mathcal{M}, \sigma \models P(x_1, ...x_n)$ iff $(\sigma(x_1), ..., \sigma(x_n)) \in I(P)$.
(ii) $\mathcal{M}, \sigma \models \varphi \wedge \psi$ iff $\mathcal{M}, \sigma \models \varphi$ and $\mathcal{M}, \sigma \models \varphi$.
(iii) $\mathcal{M}, \sigma \models \neg\varphi$ iff $\mathcal{M}, \sigma \not\models \varphi$.
(iv) $\mathcal{M}, \sigma \models \forall x\varphi$ iff $\mathcal{M}, \sigma[a/x] \models \varphi$ for all $a \in D$.
(v) $\mathcal{M}, \sigma \models \exists x\varphi$ iff $\mathcal{M}, \sigma[a/x] \models \varphi$ for some $a \in D$.

where $\sigma[a/x]$ is the variable assignment that results from setting $\sigma(x) = a$ and leaving $\sigma$ otherwise unchanged. When $\forall$ and $\exists$ are treated as generalised quantifiers clauses (iv) and (v) are slightly modified. Quantifiers, in these contexts, denote sets of subsets of the domain: $\forall$ denotes the set $\{D\}$ containing only the domain $D$ of the structure, and $\exists$ denotes the set of all non-empty subsets of $D$ (let's call it $E_D$ for short). The corresponding clauses for satisfaction are:

(iv)$'$ $\mathcal{M}, \sigma \models \forall x\varphi$ iff $\{a \in D | \mathcal{M}, \sigma[a/x] \models \varphi\} \in \{D\}$.
(v)$'$ $\mathcal{M}, \sigma \models \exists x\varphi$ iff $\{a \in D | \mathcal{M}, \sigma[a/x] \models \varphi\} \in E_D$.

It should be clear that clauses (iv)-(v) are equivalent to (iv)$'$ -(v)$'$, in the sense that they give rise to the same satisfaction relation. Bonnay and Westerståhl prefer the latter formulation, and I will follow suit. They introduce the general notion of a first-order interpretation as follows:

> Since we assume compositionality, our interpretations amount to giving syntactically adequate semantic values to the logical and non-logical vocabulary. Since we furthermore assume non-triviality, we need not worry about the interpretation

> of connectives, which has to be standard by [Theorem 1]. Hence, our interpretations can be taken to be pairs of the form $(\mathcal{M}, Q)$ where $\mathcal{M}$ is a standard $\mathcal{L}$-structure [...] interpreting the non-logical vocabulary of $\mathcal{M}$, and $Q$ is a set of subsets of [the domain of the structure], interpreting $\forall$. [3, p. 729]

As I mentioned in the introduction, there's a problem with the appeal to compositionality in the paragraph above. I'll put it aside until the next section, so as to not juggle too many issues at once. For now let's follow Bonnay and Westerståhl to the letter, and assume that compositionality and non-triviality fix the standard meaning of connectives in the first-order case. In the rest of this section I'll show that, even if we take the standard meaning of connectives for granted, Bonnay and Westerståhl's restrictions are not enough to to pin down the normal interpretation of *quantifiers*. Let's see the details.

In Bonnay and Westerståhl's set-up an interpretation for a language $\mathcal{L}$ is a *weak model* $\mathcal{M}, Q$, where $\mathcal{M} = (D, I)$ is an $\mathcal{L}$-structure and $Q$ is a set of subsets of $D$. Given a weak model $\mathcal{M}, Q$, they define satisfaction as follows:[5]

(1) $\mathcal{M}, Q, \sigma \models P(x_1, ...x_n)$ iff $(\sigma(x_1), ..., \sigma(x_n)) \in I(P)$.
(2) $\mathcal{M}, Q, \sigma \models \varphi \wedge \psi$ iff $\mathcal{M}, Q, \sigma \models \varphi$ and $\mathcal{M}, Q, \sigma \models \varphi$.
(3) $\mathcal{M}, Q, \sigma \models \neg\varphi$ iff $\mathcal{M}, Q, \sigma \not\models \varphi$.
(4)$'$ $\mathcal{M}, Q, \sigma \models \forall x\varphi$ iff $\{a \in D | \mathcal{M}, Q, \sigma[a/x] \models \varphi\} \in Q$.

Clauses (2) and (3) are just the usual clauses for $\wedge$ and $\neg$. This is to be expected; as we've seen Bonnay and Westerståhl claim that compositionality and non-triviality fix the normal interpretation of connectives. Clause (4)$'$ is almost identical to (iv)$'$, the standard clause for $\forall$. The only difference is that in (4)$'$ the denotation of $\forall$ can be *any* set of subsets of $D$, and not just the usual one, $\{D\}$. In other words, the only possible source of non-normality in a weak model comes from deviant interpretations of $\forall$. A weak model $\mathcal{M}, Q$, then, is normal if $Q = \{D\}$, and non-normal otherwise.

Bonnay and Westerståhl also require that the interpretation of quantifiers be invariant under permutations. They define this notion in the usual way: a permutation $\pi$ of the domain $D$ of a structure is a bijection of $D$ onto itself. If $\pi$ is a permutation of $D$ and $X \subseteq D$, let $\pi(X) = \{\pi(a) \in D \,|\, a \in X\}$. Then $Q \subseteq \mathcal{P}(D)$ is invariant under permutations if $Q = \{\pi(X) \in \mathcal{P}(D) \,|\, X \in Q\}$ for all permutations $\pi$. It's easy to check that $\{D\}$, the normal interpretation of $\forall$, is indeed invariant under permutations. Lastly, Bonnay and Westerståhl take the classical consequence relation $\vdash_{\mathcal{L}}$ of a first-order language $\mathcal{L}$ to be a relation between sentences (closed formulas) only; this means that a weak model $\mathcal{M}, Q$ is consistent with the classical consequence relation $\vdash_{\mathcal{L}}$ for $\mathcal{L}$ if, whenever $\Gamma \vdash_{\mathcal{L}} \varphi$ and $\mathcal{M}, Q \models \Gamma$, we have that $\mathcal{M}, Q \models \varphi$ (no reference to variable assignments is needed, since we're dealing with sentences).

The main result of Bonnay and Westerståhl's paper is a characterisation of the interpretations of the universal quantifier $\forall$ that are consistent with the classical consequence relation of a language. This characterisation uses the notion of a principal filter closed under the interpretation of terms. Recall that a principal filter $F$ on a

[5] Of course, if $\mathcal{L}$ contains individual constants and function symbols, clause (i) is modified in the usual way to account for terms.

set $S$ is a non-empty set of subsets of $S$ such that for some $G \subseteq S$, we have that $F = \{X \subseteq S \mid G \subseteq X\}$. Given a weak model $\mathcal{M}, Q$ for a language $\mathcal{L}$, possibly containing individual constants and function symbols, we say that $Q$ is a principal filter closed under the interpretation of terms in $\mathcal{M}$ if:

(i) $Q$ is a principal filter on $D$ generated by some $G \subseteq D$.
(ii) For any n-ary function symbol $f$ of $\mathcal{L}$, if $a_1, ..., a_n \in G$, then $I(f)(a_1, ...a_n) \in G$.
(iii) For any individual constant $c$ of $\mathcal{L}$, $I(c) \in G$.

Bonnay and Westerståhl's characterisation is as follows:

(BW1) Let $\mathcal{L}$ be a language with $\forall$ primitive, let $\mathcal{M}, Q$ be a weak model for it. Then $\mathcal{M}, Q$ is consistent with $\vdash_{\mathcal{L}}$ iff $Q$ is a principal filter closed under the interpretation of terms in $\mathcal{M}$.

They combine this characterisation with an ancillary result:

(BW2) A principal filter $Q$ on $D$ is invariant under permutations iff $Q = \{D\}$.

In theory (BW1) and (BW2) should take care of Carnap's Problem. According to (BW1) consistency with the classical consequence relation forces us to interpret $\forall$ as a principal filter closed under the interpretation of terms. (BW2) then shows that the only such filter that's *also* invariant under permutations is $\{D\}$, the normal interpretation of $\forall$. Therefore (or so it seems) non-normal weak models are ruled out.

The reason Bonnay and Westerståhl's strategy fails is simple: (BW1) is false for first-order languages. They prove (BW1) only for languages with predicate *variables* (that is, for *second*-order languages without second-order quantifiers), and claim that the addition of predicate variables is a 'simplifying assumption' without which 'similar results would hold' [3, p. 729]. This, however, is a mistake. Some of the weak models that falsify (BW1) are both non-normal and invariant under permutations, and so the normal interpretation of quantifiers is not secured. To see why we need an extra bit of terminology: given a weak model $\mathcal{M}, Q$ for a language $\mathcal{L}$, the *extension* of a formula $\varphi$ relative to $\sigma$ and $x$ is $||\varphi||^{\mathcal{M},Q}_{\sigma,x} := \{a \in D \mid (\mathcal{M}, Q), \sigma[a/x] \models \varphi\}$, and a subset $X$ of $D$ is *definable* if $X = ||\varphi||^{\mathcal{M},Q}_{\sigma,x}$ for some $\varphi, \sigma$ and $x$. Now, the following lemma gives a counterexample to (BW1) that is simultaneously non-normal and invariant under permutations:

**Lemma 2** *There is a first-order language $\mathcal{L}_1$ with a non-normal weak model $\mathcal{N}, Q$ consistent with the classical consequence relation $\vdash_{\mathcal{L}_1}$.*

***Proof*** Take the language $\mathcal{L}_1$ with $\wedge$, $\neg$ and $\forall$ as logical constants and a unary predicate symbol $P$ as non-logical vocabulary. Consider the $\mathcal{L}_1$-structure $\mathcal{N} = (D, I)$ where $D$ is any set such that $|D| > 1$ and $I(P) = D$. Next we need a (non-normal) interpretation for $\forall$ to make $\mathcal{N}$ into a weak model. Say we interpret $\forall$ with $E_D$, the set of all non-empty subsets of $D$ –in other words, we'll (mis)interpret $\forall$ as $\exists$.

We now have a weak model $(\mathcal{N}, E_D)$. It's clear that $(\mathcal{N}, E_D)$ is non-normal, and it is easy to check that $E_D$ is invariant under permutations. The only thing left to show

is that $(\mathcal{N}, E_D)$ is consistent with the classical consequence relation $\vdash_{\mathcal{L}_1}$ of $\mathcal{L}_1$. A routine induction on formulas shows that for any $\varphi, \sigma$ and $x$:

($d_1$) $||\varphi||^{\mathcal{N},\{D\}}_{\sigma,x} = D$ or $||\varphi||^{\mathcal{N},\{D\}}_{\sigma,x} = \emptyset$.
($d_2$) $||\varphi||^{\mathcal{N},E_D}_{\sigma,x} = ||\varphi||^{\mathcal{N},\{D\}}_{\sigma,x}$.

A second induction on formulas then shows that for any $\varphi$ and $\sigma$, we have that $\mathcal{N}, \{D\}, \sigma \models \varphi$ iff $\mathcal{N}, E_D, \sigma \models \varphi$. All steps of this second induction are trivial except the quantifier one, which goes as follows:

$\mathcal{N}, \{D\}, \sigma \models \forall x\varphi$ iff by (4)′
$||\varphi||^{\mathcal{N},\{D\}}_{\sigma,x} = D$ iff by ($d_1$), ($d_2$)
$|\varphi||^{\mathcal{N},E_D}_{\sigma,x} = D$ iff by ($d_1$), ($d_2$)
$|\varphi||^{\mathcal{N},E_D}_{\sigma,x} \in E_D$ iff by (4)′
$\mathcal{N}, E_D, \sigma \models \forall x\varphi$.

Finally, $\mathcal{N}, \{D\}$ is a normal interpretation, and hence consistent with $\vdash_{\mathcal{L}_1}$. But by the above the normal weak model $\mathcal{N}, \{D\}$ is equivalent to the non-normal weak model $\mathcal{N}, E_D$, so $\mathcal{N}, E_D$ is also consistent with $\vdash_{\mathcal{L}_1}$. □

Lemma 2 is formulated for the very simple language $\mathcal{L}_1$, but it can be adapted to richer ones. If a language has function symbols and individual constants, the function $I$ can give them any value whatsoever, and if it has further predicate letters $R$ of arity $n$ we just need to set $I(R) = D^n$. It's also important to note that the fact that we (mis)interpreted $\forall$ as $\exists$ is largely irrelevant. For instance, let $E_n$ be the set of subsets of $D$ containing at least $n$ elements. Then Lemma 2 also applies if the denotation of $\forall$ is $E_n$ for any $n > 1$. And finally, Lemma 2 can be adapted to languages with $\exists$, rather than $\forall$, as a primitive symbol. Let a weak model $\mathcal{M}, Q$ for a language where $\exists$ is primitive be $\exists$-normal if $Q = E_D$. Then:

**Corollary 3** *There is a a language $\mathcal{L}_2$ with $\exists$ as a primitive symbol and a non $\exists$-normal weak model $\mathcal{N}, Q$ consistent with $\vdash_{\mathcal{L}_2}$.*

***Proof*** Let $\mathcal{L}_2$ be exactly as $\mathcal{L}_1$ but with $\exists$ primitive instead of $\forall$. It follows from Lemma 1 that $\mathcal{N}, \{D\}$ is a non $\exists$-normal weak model for $\mathcal{L}_2$ consistent with $\vdash_{\mathcal{L}_2}$. □

We can explain what's going on from a more general perspective by adapting a result from [1]. Consider once again the normal interpretation of $\exists$:

$$\mathcal{M}, E_D, \sigma \models \exists x\varphi \text{ iff } \{a \in D | \mathcal{M}, \sigma[a/x] \models \varphi\} \in E_D$$

There's a certain sense in which this interpretation overshoots the mark. We don't really need $\exists$ to denote *all* non-empty subsets of the domain in order to get the usual satisfaction relation. It's enough that it denotes all the *definable* subsets of the domain, those that are the extension of some formula. Let's make this observation more precise and more general. Given a weak model $\mathcal{M}, Q$, let $\mathrm{Def}(\mathcal{M}, Q)$ be the set of definable subsets of $D$. Then:[6]

[6] Lemma 4 is a modification of the Hull Theorem from [1]. The modification is needed to (a) adapt Antonelli's idea to the present setting and (b) avoid a problem with Antonelli's definition of the hull of a model, pointed out in [2].

**Lemma 4** *If $\mathcal{M}, Q$ and $\mathcal{M}, Q'$ are weak models such that $Q' = Q \cap Def(\mathcal{M}, Q)$, then $\mathcal{M}, Q, \sigma \models \varphi$ iff $\mathcal{M}, Q', \sigma \models \varphi$ for any $\varphi$ and $\sigma$.*

***Proof*** The proof is by induction on formulas. The weak models agree on the interpretation of atomic formulas and connectives, so all steps are trivial except the quantifier one. Let $\overline{Q}$ be whatever quantifier $Q$ and $Q'$ are meant to interpret. Then:

$\mathcal{M}, Q, \sigma \models \overline{Q}x\psi$ iff
$||\psi||^{\mathcal{M},Q}_{\sigma,x} \in Q \cap \text{Def}(\mathcal{M}, Q)$ iff (by induction hypothesis)
$|\psi||^{\mathcal{M},Q'}_{\sigma,x} \in Q \cap \text{Def}(\mathcal{M}, Q)$ iff by (by assumption)
$|\psi||^{\mathcal{M},Q'}_{\sigma,x} \in Q'$ iff
$\mathcal{M}, Q', \sigma \models \overline{Q}x\psi$.

□

Corollary 3 can now be seen as a special case of Lemma 4. The only definable sets of $\mathcal{N}, E_D$ are $D$ and $\emptyset$. Thus $\{D\} = E_D \cap \text{Def}(\mathcal{N}, E_D)$, and so the non $\exists$-normal model $\mathcal{N}, \{D\}$ is equivalent to the $\exists$-normal model $\mathcal{N}, E_D$.

Something similar happens with the normal interpretation of $\forall$:

$$\mathcal{M}, \{D\}, \sigma \models \forall x\varphi \text{ iff } \{a \in D | \mathcal{M}, \sigma[a/x] \models \varphi\} \in \{D\}$$

In this case we don't need $\forall$ to denote the set containing *only* $D$. We can, for instance, clutter its denotation with undefinable subsets and get an equivalent weak model:

**Lemma 5** *Let $\mathcal{M}, Q$ and $\mathcal{M}, Q'$ be weak models such that $Q' = Q \cup B$ for some $B \subseteq \mathcal{P}(D)$ such that $B \cap Def(\mathcal{M}, Q) = \emptyset$. Then $\mathcal{M}, Q, \sigma \models \varphi$ iff $\mathcal{M}, Q', \sigma \models \varphi$ for any $\varphi$ and $\sigma$.*

***Proof*** Again by induction on formulas. All steps are trivial except the quantifier one. For the left-to-right direction note that

$\mathcal{M}, Q, \sigma \models \overline{Q}x\psi \Rightarrow ||\psi||^{\mathcal{M},Q}_{\sigma,x} \in Q \Rightarrow$ (by induction hypothesis) $||\psi||^{\mathcal{M},Q'}_{\sigma,x} \in Q \Rightarrow ||\psi||^{\mathcal{M},Q}_{\sigma,x} \in Q \cup B \Rightarrow \mathcal{M}, Q', \sigma \models \overline{Q}x\psi$.

For the right-to left direction, note that

$\mathcal{M}, Q', \sigma \models \overline{Q}x\psi \Rightarrow ||\psi||^{\mathcal{M},Q'}_{\sigma,x} \in Q \cup B \Rightarrow$ (by induction hypothesis) $||\psi||^{\mathcal{M},Q}_{\sigma,x} \in Q \cup B \Rightarrow ||\psi||^{\mathcal{M},Q}_{\sigma,x} \in Q \Rightarrow \mathcal{M}, Q, \sigma \models \overline{Q}x\psi$.

□

Lemma 2 can then be seen as a special case of Lemma 5. The only definable sets of $\mathcal{N}, \{D\}$ are $D$ and $\emptyset$. Thus $E_D = \{D\} \cup B$ for a $B \subseteq \mathcal{P}(D)$ such that $B \cap \text{Def}(\mathcal{N}, \{D\}) = \emptyset$, and hence the non $\forall$-normal weak model $\mathcal{N}, E_D$ is equivalent to the $\forall$-normal weak model $\mathcal{N}, \{D\}$.

The results in this section also explain why Bonnay and Westerståhl's use of predicate variables is not a simplifying assumption, but a crucial part of their set-up. The non-normal weak models above exploit the fact that, given a first-order language and

a structure for it, there are often subsets of the domain that can't be defined by any formula.[7] Adding predicate variables to a first-order language boosts its expressive power, and therefore rules out a host of non-normal interpretations. Bonnay and Westerståhl's proof of (BW1) –for *second*-order languages– hinges on this, and cannot be adapted to first-order ones. Ultimately, this is why their strategy to solve Carnap's Problem doesn't scale up from the propositional to the first-order case. Even if we assume that the standard interpretation of connectives is fixed –and we shouldn't, as the next section will make clear– there are still non-normal interpretations of the quantifiers that fulfil all of Bonnay and Westerståhl's constraints.

## 4 Carnap's Problem and Compositionality

In the previous section I flagged a problem with Bonnay and Westerståhl's definition of an interpretation. In this section I'll explain the problem in detail, and explore some of its consequences.

The key fact to keep in mind is the following: the usual, two-valued semantics for first-order languages is *not* compositional. This calls for some explanation. When we talk about a semantics for first-order languages, as I said, we usually have in mind the Tarskian satisfaction relation defined by means of clauses (i)-(v) –or equivalently, (i)-(v)′– in the previous section. Strictly speaking, however, these clauses are are not an assignment of semantic values; they are the definition of a ternary relation $\models$ between structures, variable assignments, and formulas. The notion of compositionality only applies to assignments of semantic values, so it doesn't really make sense to say that a satisfaction relation is, or isn't, compositional. Still, there's a well-known way in which a satisfaction relation $\models$ determines a class of valuations: just set $v_\sigma^{\mathcal{M}}(\varphi) = 1$ if $\mathcal{M}, \sigma \models \varphi$, else $v_\sigma^{\mathcal{M}}(\varphi) = 0$, for any formula $\varphi$. As the notation suggests, this yields a valuation $v_\sigma^{\mathcal{M}}$ for each variable assignment $\sigma$ from $\mathcal{L}$ to $\mathcal{M}$.

Each satisfaction relation, then, defines a two valued semantics for first-order languages. But there's a catch: this semantics is not always compositional. Take a language $\mathcal{L}$, an $\mathcal{L}$-structure $\mathcal{M}$, and the standard satisfaction relation $\models$ between them. In any valuation determined by $\models$, the semantic value of a formula $\varphi \wedge \psi$ is a function of the semantic value of $\varphi$ and the semantic value of $\psi$: $v_\sigma^{\mathcal{M}}(\varphi \wedge \psi) = 1$ iff $v_\sigma^{\mathcal{M}}(\varphi) = v_\sigma^{\mathcal{M}}(\psi) = 1$, for any $\sigma$. Something similar happens with negation: $v_\sigma^{\mathcal{M}}(\neg\varphi) = 1$ iff $v_\sigma^{\mathcal{M}}(\varphi) = 0$ (again, for arbitrary $\sigma$). But the same does not hold for quantifiers: the truth value of $\forall x\varphi$ does not depend only on the truth value of $\varphi$. Put more formally, the value of $v_\sigma^{\mathcal{M}}(\forall x\varphi)$ is not a function of the value of $v_\sigma^{\mathcal{M}}(\varphi)$, but of the values of $v_{\sigma'}^{\mathcal{M}}(\varphi)$ for all appropriate $\sigma'$.[8] Indeed, it's easy to find structures $\mathcal{M}$ such that for some $\varphi$, $\psi$ and $\sigma$, we have that $v_\sigma^{\mathcal{M}}(\varphi) = v_\sigma^{\mathcal{M}}(\psi) = 1$ but $v_\sigma^{\mathcal{M}}(\forall x\varphi) = 0$ and

[7] Note, by the way, that undefinability is a pervasive phenomenon that applies to languages with more expressive power than I've considered above. For instance, a simple cardinality argument shows that given any countable $\mathcal{L}$ and any $\mathcal{L}$-structure $\mathcal{M} = (D, I)$ with a countably infinite $D$, some subsets of $D$ are not definable by any $\mathcal{L}$-formula.

[8] The appropriate $\sigma'$ are of course those of the form $\sigma[a/x]$ for some element $a$ of the domain.

$v_\sigma^{\mathcal{M}}(\forall x\psi) = 1$. In those cases no function on semantic values (i.e. no truth-function) corresponds to the syntactic operation generating $\forall x\varphi$ from $\varphi$ –and similarly for $\exists$.[9]

The fact that the valuations $v_\sigma^{\mathcal{M}}$ are not compositional is not a problem in and of itself. They assign truth-values to formulas in a natural and systematic way, following the recursive clauses for satisfaction. Nevertheless, this failure of compositionality *must* be a problem for Bonnay and Westerståhl. For them compositionality is a universal constraint that all interpretations, normal or not, must abide by. Unfortunately, they seem to have a two-valued semantics in mind when defining interpretations of first-order languages. Recall what they say:

> Since we assume compositionality, our interpretations amount to giving syntactically adequate semantic values to the logical and non-logical vocabulary. Since we furthermore assume non-triviality, we need not worry about the interpretation of connectives, which has to be standard, by [Theorem 1]. Hence, our interpretations can be taken to be pairs of the form $(\mathcal{M}, Q)$ where... [3, p. 729]

Theorem 1 says that *if our only semantic values are 0 and 1*, compositionality and non-triviality rule out all non-normal interpretations of the connectives. But then something must be wrong with the passage above. If the only semantic values are 0 and 1, then normal interpretations are not (always) compositional. If, on the other hand, there are semantic values besides 1 and 0, then the appeal to Theorem 1 is moot, and we can't assume that compositionality and non-triviality pin down the standard interpretation of connectives. This is the second problem with Bonnay and Westerståhl's approach: their definition of a first-order interpretation either violates compositionality or else begs the question against non-normal interpretations of the connectives.

It is possible to give compositional semantics for first-order languages, which leads to a natural question. Suppose we reformulate Bonnay and Westerståhl's notion of an interpretation so that (a) normal interpretations are always compositional and (b) we don't beg the question against non-normal interpretations of the connectives. Do compositionality, non-triviality and invariance under permutations pin down the normal interpretation of logical vocabulary?

The answer is 'no'. The reason, roughly put, is that compositional semantics for first-order languages require a large number of semantic values. This, in its turn, makes Carnap's Problem more difficult to solve, since more semantic values entail more possible interpretations, and therefore more unintended ones to rule out. As we'll see, this means that in the new setting Bonnay and Westerståhl's restrictions are not enough to fix the standard interpretation of *connectives*, let alone quantifiers.

The most common strategy to set up a compositional semantics for first-order languages is to take the semantic value $[\![\varphi]\!]^{\mathcal{M}}$ of a formula $\varphi$ in a structure $\mathcal{M}$ to be the

[9] Note that if we use *partial* valuations that assign truth-values to sentences (closed formulas) only, the resulting semantics is not compositional either. Set, for all sentences $\varphi$, $v^{\mathcal{M}}(\varphi) = 1$ if $\mathcal{M} \models \varphi$, else $v^{\mathcal{M}}(\varphi) = 0$ (no reference to assignments $\sigma$ is needed, since we are dealing with sentences). Compositionality, on Bonnay and Westerståhl's formulation, requires that for each syntactic operation $\mathcal{O}$ there should be a corresponding operation on semantic values. The syntactic components of sentences, however, are not in general sentences themselves. For instance, the syntactic components of $\forall x\, P(x)$ are all subsentential, so no operation on semantic values corresponds to the syntactic operation that generates $\forall x\, P(x)$.

set of variable assignments $\sigma$ such that $\mathcal{M}, \sigma \models \varphi$.[10] Here's some new notation: given a structure $\mathcal{M} = (D, I)$ for a language $\mathcal{L}$, $A^{\mathcal{M}}$ is the set of all variable assignments from $\mathcal{L}$ to $D$. Then first-order formulas can be interpreted as follows:[11]

(1) $[\![P(x_1, ...x_n)]\!]^{\mathcal{M}} = \{\sigma \in A^{\mathcal{M}} \mid (\sigma(x_1), ..., \sigma(x_n)) \in I(P)\}$
(2) $[\![\varphi \wedge \psi]\!]^{\mathcal{M}} = [\![\varphi]\!]^{\mathcal{M}} \cap [\![\psi]\!]^{\mathcal{M}}$
(3) $[\![\neg\varphi]\!]^{\mathcal{M}} = A^{\mathcal{M}} - [\![\varphi]\!]^{\mathcal{M}}$
(4) $[\![\exists x\varphi]\!]^{\mathcal{M}} = \{\sigma \in A^{\mathcal{M}} \mid \sigma \sim_x \sigma' \text{ for some } \sigma' \in [\![\varphi]\!]^{\mathcal{M}}\}$
(5) $[\![\forall x\varphi]\!]^{\mathcal{M}} = \{\sigma \in A^{\mathcal{M}} \mid \sigma' \in [\![\varphi]\!]^{\mathcal{M}} \text{ for all } \sigma' \text{ st. } \sigma \sim_x \sigma'\}$

The general idea is clear: the semantic values of formulas are sets of assignments, and conjunction and negation are interpreted as operations on semantic values (intersection and complementation, respectively). As pointed out in [26], however, the resulting semantics is not fully compositional either. The problem lies again with the interpretation of quantifiers –that is, with clauses (4) and (5). Compositionality, remember, demands that for each syntactic operation $\mathcal{O}$ there should be a corresponding operation on semantic values. The usual way of setting up the syntax of first-order languages has *one* syntactic operation that takes the quantifier $\forall$, any variable $x$, and any formula $\varphi$, and returns a formula $\forall x\varphi$ – and similarly for $\exists$. Clause (5), on the other hand, covertly defines one semantic operation *for each variable* $x$; the right-hand side of (5) refers to $x$ *qua* syntactic object –and so does (4).

Different (and somewhat strained) solutions to this problem have been put forward. One option is giving semantic values to variables. We could, for instance, take the value $[\![x]\!]^{\mathcal{M}}$ of $x$ in $\mathcal{M}$ to be the variable $x$ itself, and reformulate the interpretations of $\exists$ and $\forall$ accordingly, as functions that take a set of assignments and a variable as arguments, and map them to a set of assignments.[12] On this account the normal interpretations of $\wedge$ and $\neg$ would still be intersection and complementation, but the normal interpretations of quantifiers would be the following functions from $\text{Vars} \times \mathcal{P}(A^{\mathcal{M}})$ to $\mathcal{P}(A^{\mathcal{M}})$:

$f_{\exists}(x, Y) = \{\sigma \in A^{\mathcal{M}} \mid \sigma \sim_x \sigma' \text{ for some } \sigma' \in Y\}$
$f_{\forall}(x, Y) = \{\sigma \in A^{\mathcal{M}} \mid \sigma' \in Y \text{ for all } \sigma' \text{ st. } \sigma \sim_x \sigma'\}$

Another option is to slightly change the formulation of the syntax, so that for each variable $x$ we have a syntactic operation $\mathcal{O}_x$ that takes a string '$\forall x$' and a formula $\varphi$ and returns a formula $\forall x\varphi$ (similarly for $\exists$); on this account the string '$\forall x$' is then interpreted as a single unit. The normal interpretations of $\wedge$ and $\neg$ are intersection and complementation, as before, but the normal interpretations of quantifier-variable strings are functions from $\mathcal{P}(A^{\mathcal{M}})$ to $\mathcal{P}(A^{\mathcal{M}})$:

$f_{\exists x}(Y) = \{\sigma \in A^{\mathcal{M}} \mid \sigma \sim_x \sigma' \text{ for some } \sigma' \in Y\}$
$f_{\forall x}(Y) = \{\sigma \in A^{\mathcal{M}} \mid \sigma' \in Y \text{ for all } \sigma' \text{ st. } \sigma \sim_x \sigma'\}$

[10] See [11, 13, 15] or [10].

[11] Once again, if a language contains individual constants and function symbols clause (1) is modified in the obvious way to account for terms. The notation '$\sigma \sim_x \sigma'$' means that the assignment $\sigma$ differs from the assignment $\sigma'$ at most in the value it assigns to $x$.

[12] This is a simplification of [27, Ch. 10], which takes a similar but more elaborate route.

All of what I'll say applies, with obvious modifications, to either option. The second one saves some time and some brackets, so I'll take it as the official normal interpretation. Again, nothing hinges on this.

Now we have to adapt the rest of the notions involved in formulating Carnap's Problem. It will be useful to keep the simpler propositional case in mind while we do it, just to make sure we don't go astray. An interpretation for a propositional language is a valuation $v$ from formulas to $\{1, 0\}$. In the first-order case interpretations depend on structures $\mathcal{M} = (D, I)$. Given such an $\mathcal{M}$, an interpretation is a function from formulas to $\mathcal{P}(A^{\mathcal{M}})$.

A compositional valuation associates a truth-function to each connective. Thus, compositional valuations can be defined as tuples $v = (v_o, t_\wedge, t_\vee, t_\neg, t_\rightarrow)$, where $v_o$ is a valuation for atoms and the $t_\diamond$ are truth-functions. In the first-order case compositional interpretations can also be defined as tuples $\mathcal{T} = (f_o, F_\wedge, F_\neg, F_\forall, F_\exists)$. Here $f_o$ is a function from atoms to $\mathcal{P}(A^{\mathcal{M}})$, and the $F_\diamond$ are operations on $\mathcal{P}(A^{\mathcal{M}})$.[13]

In a normal, compositional valuation the $t_\diamond$ are the intended truth-functions. Similarly, in a normal, compositional interpretation $\mathcal{T} = (f_0, F_\wedge, F_\neg, F_\forall, F_\exists)$ the $F_\diamond$ are the intended operations. There's an additional wrinkle to take care of, though. Given the usual definition of satisfaction, the semantic value of an atomic formula $P(x_1, ...x_n)$ is closed under assignments that agree on the free variables $x_1, ...x_n$. Formally, this means that if $\sigma \in [\![P(x_1, ...x_n)]\!]^{\mathcal{M}}$ and $\sigma, \sigma'$ agree on the value of $x_1, ...x_n$, then $\sigma' \in [\![P(x_1, ...x_n)]\!]^{\mathcal{M}}$. Our interpretations are tuples $\mathcal{T} = (f_0, F_\wedge, F_\neg, F_\forall, F_\exists)$ where $f_o$ is an *arbitrary* function from atomic formulas to $\mathcal{P}(A^{\mathcal{M}})$, and arbitrary functions need not respect that closure condition. In order for an interpretation to be normal, then, we must also demand that $f_0$ obeys clause (1) above. And finally, a valuation is non-trivial if it does not assign the value 1 to all formulas. Similarly, $(f_o, F_\wedge, F_\neg, F_\forall, F_\exists)$ is non-trivial if it does not assign the value $A^{\mathcal{M}}$ to all formulas.

This takes care of the normal interpretations, as well as the non-normal, compositional and non-trivial ones. Now we need to address invariance under permutations. The notion of invariance in Section 3 works well if we take quantifiers to denote sets of subsets of $D$. In order to achieve compositionality, however, we've had to interpret quantifiers as operations on sets of assignments; the definition of invariance has to be adapted. Luckily there is a well-known, off-the-shelf way to do this due to [13]. If $\pi$ is a permutation of $D$ and $X$ is a set of variable assignments, let $\pi^*(X) = \{\pi \circ \sigma \in A^{\mathcal{M}} | \sigma \in X\}$. Then an n-ary operation $\mathcal{O}$ on $A^{\mathcal{M}}$ is invariant under permutations if $\pi^*(\mathcal{O}(X_1, ..., X_n)) = \mathcal{O}(\pi^*(X_1), ..., \pi^*(X_n))$ for all permutations $\pi$.

The last missing piece of the puzzle is consistency with a consequence relation $\vdash$. This is also easy to address. We'll say that $\mathcal{T}$ is consistent with $\vdash$ if, whenever $\Gamma \vdash \varphi$, we have that $\bigcap_{\gamma \in \Gamma} [\![\gamma]\!]^{\mathcal{T}} \subseteq [\![\varphi]\!]^{\mathcal{T}}$, where $[\![\psi]\!]^{\mathcal{T}}$ is the semantic value of $\psi$ on the interpretation $\mathcal{T}$.[14] Following Bonnay and Westerståhl $\vdash_{\mathcal{L}}$ is a relation between sentences only (although this doesn't really matter).

[13] Strictly speaking, given our official normal interpretation $F_\forall$ and $F_\exists$ are *sets* of operations (one for each variable).

[14] At the risk of stating the obvious: this definition of consistency just says that, given a valid argument, any assignment that makes the premises true must make the conclusion true too.

We have, at last, formulated all the notions related to Carnap's Problem in a way that's compatible with Bonnay and Westerståhl's assumptions. The following lemma shows that, in this setting, compositionality, non-triviality, and invariance under permutations don't fix the standard meaning of logical vocabulary.

**Lemma 6** *There is a first-order language $\mathcal{L}_3$ with a compositional, non-normal, non-trivial, invariant-under-permutations interpretation $\mathcal{T}$ that is consistent with $\vdash_{\mathcal{L}}$.*

***Proof*** Take the language $\mathcal{L}_3$ with $\wedge$, $\neg$ and $\forall$ as logical constants and a binary predicate symbol $R$ as non-logical vocabulary. Consider the $\mathcal{L}_3$-structure $\mathcal{M} = (D, I)$ where $D$ is any set such that $|D| > 1$ and $I(R) = D^2$. We will now construct an interpretation $\mathcal{T} = (f_o, F_\wedge, F_\neg, F_\forall, F_\exists)$ based on $\mathcal{M}$. Let $C \subset A^{\mathcal{M}}$ be the set of variable assignments such that $\sigma(x) = \sigma(y)$ for all variables $x, y$. Note that $C$ is indeed a proper subset of $A^{\mathcal{M}}$, since $|D| > 1$. Then $\mathcal{T}$ is defined as follows:

$f_0(R(x_1, x_2)) = \{\sigma \in A^{\mathcal{M}} \mid (\sigma(x_1), \sigma(x_2)) \in I(R)\}$
$F_\wedge(X, Y) = X \cap Y$.
$F_\neg(A^{\mathcal{M}}) = C$, and if $X \neq A^{\mathcal{M}}$ $F_\neg(X) = A^{\mathcal{M}}$.
$F_{\exists y}(X) = F_{\forall y}(X) = X$ for all variables $y$.

$\mathcal{T}$ is clearly compositional, so it remains to show that (a) $\mathcal{T}$ is non-normal, (b) $\mathcal{T}$ is non-trivial, (c) the $F_\diamond$ are invariant under permutations and (d) $\mathcal{T}$ is consistent with $\vdash_{\mathcal{L}_3}$.

The fact that (a) $\mathcal{T}$ is non-normal is obvious; $F_\neg$, $F_\forall$ and $F_\exists$ are not the intended operations. To see that that (b) $\mathcal{T}$ is non-trivial just note that $[\![\neg\exists x_1 \exists x_2 R(x_1, x_2)]\!]^{\mathcal{T}} = C \neq A^{\mathcal{M}}$. Next we need to prove that (c) the $F_\diamond$ are invariant under permutations. This is obvious for $F_\wedge$, which is just intersection, and also for the operations in $F_\exists$ and $F_\forall$, which are just the identity function. To show the same for $F_\neg$ first note that, for any permutation $\pi$ of the domain, $\pi^*(A^{\mathcal{M}}) = A^{\mathcal{M}}$ and $\pi^*(C) = C$. In other words, $A^{\mathcal{M}}$ and $C$, seen as 0-ary operations, are invariant under permutations. But then:

$F_\neg(\pi^*(A^{\mathcal{M}})) = F_\neg(A^{\mathcal{M}}) = C = \pi^*(C) = \pi^*(F_\neg(A^{\mathcal{M}}))$
$F_\neg(\pi^*(X)) = A^{\mathcal{M}} = \pi^*(A^{\mathcal{M}}) = \pi^*(F_\neg(X))$ for $X \neq A^{\mathcal{M}}$

where the second identity rests on the fact that if $X \neq A^{\mathcal{M}}$, then $\pi^*(X) \neq A^{\mathcal{M}}$ for any permutation $\pi$.

Finally, we need to prove that (d) $\mathcal{T}$ is consistent with $\vdash_{\mathcal{L}_3}$. A trivial induction on formulas shows that for any $\varphi$:

($d_1$) $[\![\varphi]\!]^{\mathcal{T}} = A^{\mathcal{M}}$ or $[\![\varphi]\!]^{\mathcal{T}} = C$.
($d_2$) $[\![\varphi]\!]^{\mathcal{T}} = A^{\mathcal{M}}$ iff $[\![\varphi]\!]^{\mathcal{M}} = A^{\mathcal{M}}$.
($d_3$) $[\![\varphi]\!]^{\mathcal{T}} = C$ iff $[\![\varphi]\!]^{\mathcal{M}} = \emptyset$.

Now we can prove the result by contraposition. Let $\Gamma \cup \{\varphi\}$ be a set of sentences, and suppose that $\bigcap_{\gamma \in \Gamma} [\![\gamma]\!]^{\mathcal{T}} \not\subseteq [\![\varphi]\!]^{\mathcal{T}}$. By ($d_1$) we must have that $[\![\gamma]\!]^{\mathcal{T}} = A^{\mathcal{M}}$ for all $\gamma \in \Gamma$ and $[\![\varphi]\!]^{\mathcal{T}} = C$. By ($d_2$) and ($d_3$) this means that $[\![\gamma]\!]^{\mathcal{M}} = A^{\mathcal{M}}$ for all $\gamma \in \Gamma$

and $[\![\varphi]\!]^{\mathcal{M}} = \emptyset$. Therefore, $\Gamma \nvdash_{\mathcal{L}_3} \varphi$. Note, incidentally, that if we had taken $\vdash_{\mathcal{L}_3}$ to be a relation between all formulas, open and closed, the proof would have still gone through. □

As before, Lemma 6 is stated for the very limited language $\mathcal{L}_3$, but can be adapted to richer ones: if a language has constants and function symbols, the function $I$ can give them any value whatsoever, and if it has further predicate letters $P$ of arity $n$, we just need to set $I(P) = D^n$.

There is a second, more important sense in which Lemma 6 can be generalised. It turns out that if we take the semantic values of formulas to be sets of assignments, there are non-normal interpretations *even if every subset of D (and every subset of $D^n$, for any n) is definable*.[15] The reason is simple: given a sufficiently large domain, certain sets of assignments can never be the semantic value of *any* formula on *any* normal interpretation. This allows us to construct non-normal interpretations that differ from normal ones only in the way they behave with respect to these 'extra' semantic values, and this is possible regardless of the interpretation of non-logical vocabulary. Let's spell out the details.

The semantic value $[\![\varphi]\!]^{\mathcal{M}}$ of an open formula $\varphi$ on the normal interpretation based on a structure $\mathcal{M}$ is the set of variable assignments $\sigma$ such that $\mathcal{M}, \sigma \models \varphi$, where $\models$ is the usual satisfaction relation. Clearly, whether $\mathcal{M}, \sigma \models \varphi$ or not hinges on the value $\sigma$ assigns to the (finitely many) free variables $\vec{x}$ of $\varphi$. We need some terminology to describe this. Given a structure $\mathcal{M}$ for $\mathcal{L}$ and a (non-empty) finite sequence of variables $x_1, ...x_n = \vec{x}$, we'll say that a set of assignments $Y \subseteq A^{\mathcal{M}}$ *depends on* $\vec{x}$ if there is a $\sigma$ and a $\sigma'$ that differ at most in the values they assign to some variables in $\vec{x}$, but such that $\sigma \in Y$ and $\sigma' \notin Y$. A set $Y \subseteq A^{\mathcal{M}}$ is *dependent* if it depends on some finite $\vec{x}$, and *independent* otherwise. Now, it is easy to check that:

**Observation 7** The semantic value $[\![\varphi]\!]^{\mathcal{M}}$ of an arbitrary formula $\varphi$ in a normal interpretation based on $\mathcal{M}$ is always $A^{\mathcal{M}}$, $\emptyset$, or a dependent subset of $A^{\mathcal{M}}$.

Crucially, Observation 7 holds regardless of whether every subset of the domain $D$ is definable. In addition:

**Observation 8** Given a structure $\mathcal{M}$ for $\mathcal{L}$, some independent subsets $Y$ of $A^{\mathcal{M}}$ are invariant under permutations.

For instance, given any structure $\mathcal{M}$, the set $C_\infty$ of assignments that give the same value to (at least) countably infinitely many variables is independent and invariant under permutations. By combining Observations 7 and 8 we get that:

**Lemma 9** *Let $\mathcal{L}$ be any first-order language. Then there is a compositional, non-normal, non-trivial, invariant-under-permutations interpretation $\mathcal{T}$ consistent with $\vdash_{\mathcal{L}}$.*

[15] Adapting the notion of definability to this setting is trivial: the extension of a formula $\varphi$ (relative to $\sigma$ and $x$, on an interpretation $\mathcal{I}$) is $||\varphi||^{\mathcal{I}}_{\sigma,x} := \{a \in D \mid \sigma[a/x] \in [\![\varphi]\!]^{\mathcal{I}}\}$, and a subset $X$ of $D$ is definable if $X = ||\varphi||^{\mathcal{I}}_{\sigma,x}$ for some $\varphi$, $\sigma$ and $x$. This is extended to cover subsets of $D^n$ in the obvious way.

***Proof*** Take any $\mathcal{L}$-structure $\mathcal{M} = (D, I)$ with $|D| \geq \omega$. By Observations 7 and 8 some independent subsets of $A^{\mathcal{M}}$ are invariant under permutations. Since $|D| \geq \omega$, we have that $C_\infty \neq A^{\mathcal{M}}$. Thus, some independent subsets of $A^{\mathcal{M}}$ are invariant under permutations *and* not the semantic value of any formula under the normal interpretation based on $\mathcal{M}$. Next we'll define a non-normal $\mathcal{T} = (f_0, F_\wedge, F_\neg, F_\forall, F_\exists)$ based on $\mathcal{M}$. Let $f_0$ interpret atoms normally, and let the $F_\diamond$ behave like the normal operations except when applied to subsets of $A^{\mathcal{M}}$ that are independent, invariant under permutations, and not the semantic value of a formula under the normal interpretation; for those sets (pairs of sets, in the case of $F_\wedge$), the $F_\diamond$ are the constant function to $C_\infty$. Then $\mathcal{T} = (f_0, F_\wedge, F_\neg, F_\forall, F_\exists)$ is compositional, non-normal and invariant under permutations. Moreover, since the $F_\diamond$ behave non-normally only for sets of assignments that are not the semantic value of a formula, and otherwise behave like the normal operations, $\mathcal{T}$ is non-trivial and consistent with $\vdash_{\mathcal{L}}$. Note that, as mentioned above, this construction works regardless of how the function $I$ interprets non-logical vocabulary □

It is time to take stock now. There are, I think, three morals to this section. The first is that there's a problem with Bonnay and Westerståhl's interpretations: they either violate compositionality or else beg the question against non-normal interpretations of the connectives. The second is that demanding compositionality does not, in general, help to avoid Carnap's Problem. Take the first-order case. If we want the usual two-valued semantics, then compositionality is too strong a requirement: it rules out normal valuations. If, on the other hand, we enrich our range of semantic values, then Carnap's Problem suddenly becomes much more difficult. And this takes us to the third, related moral. The more semantic values we throw into the mix, the harder Carnap's Problem (in general) becomes. Now let's see if we can find our way out of the woods.

## 5 Carnap's Problem and Logical Validity

So far I've focused on the shortcomings of Bonnay and Westerståhl's approach. In this section I'll change tack, and argue that a slight modification of their proposal gets around Carnap's Problem. The overall idea is simple: using second-order variables won't do, but if we look at our inferential practice, something similar to it to it can be justified.

The way forward is to focus on the notion of consistency with a consequence relation. Up until now we have identified consistency and truth-preservation: an interpretation was consistent with the classical consequence relation of a language if it made all classically valid arguments truth-preserving. A closer look at the way we draw inferences, though, suggests that this is too weak a requirement.

Take the inference from '7 is a counterexample to Goldbach's conjecture and 8 is a counterexample to Goldbach's conjecture' to '7 is a counterexample to Goldbach's conjecture'. We accept this inference despite not knowing the extension of the predicate 'is a counterexample to Goldbach's conjecture' (and regardless of what it happens to be). More generally, we accept all inferences of the form $P(c_1) \wedge P(c_2) \vdash P(c_1)$, despite not knowing the extension of every predicate $P$ and every individual constant

$c_1$ and $c_2$ (and regardless of what they happen to be). If we want to account for the inferential role of $\wedge$, then, it isn't enough to give it a semantic value that makes our inferences truth-preserving *given that P, $c_1$ and $c_2$ are interpreted a certain way*: we typically *don't* know the interpretation of all the non-logical vocabulary in an argument. To account for the inferential role of $\wedge$ we must give it a semantic value that makes valid arguments truth-preserving *regardless of how of non-logical vocabulary is interpreted*. This is not an isolated example, but the norm. Take any inference of the form $\forall x P(x) \vdash P(c)$. We also accept it without knowing the extension of $P$ or the reference of $c$, whatever they happen to be. To account for the role of $\forall$ in inferences, then, we must give it a semantic value that makes arguments of this form truth-preserving *no matter how P and c are interpreted*.[16]

Let me make the point more general: so far we've read semantics off consequence relations by looking at interpretations that make valid arguments truth-preserving. But to adequately represent the role of logical vocabulary, to do justice to the way we actually use it, is to give it semantic values that make valid arguments truth-preserving *regardless of the interpretation of non-logical vocabulary*. The notion of consistency with a consequence relation, then, should be strengthened accordingly.

The technical set-up this requires is straightforward. Taking sets of assignments as the semantic values of formulas, as we saw, makes Carnap's Problem worse. We'll go for a simpler, more natural option: the semantic values of formulas will be truth-values. The rest of modifications are obvious. In previous sections we built interpretations on top of structures $\mathcal{M} = (D, I)$. We're going to require truth-preservation across reinterpretations of non-logical vocabulary, so the function $I$ has to go. Everything else is largely as before: given a domain $D$, an interpretation is a tuple $\mathcal{T} = (F_\wedge, F_\neg, Q)$, where $F_\wedge$ and $F_\neg$ are truth-functions interpreting $\wedge$ and $\neg$, respectively, and $Q$ is a set of subsets of $D$, interpreting $\forall$. When we add a function $I$ for the non-logical vocabulary, formulas are evaluated in the obvious way:

$\mathcal{T}_\sigma(Px_1, ...x_n) = 1$ iff $(\sigma(x_1), ..., \sigma(x_n)) \in I(P)$.
$\mathcal{T}_\sigma(\varphi \wedge \psi) = F_\wedge[\mathcal{T}_\sigma(\varphi), \mathcal{T}_\sigma(\psi)]$.
$\mathcal{T}_\sigma(\neg\varphi) = F_\neg[\mathcal{T}_\sigma(\varphi)]$.
$\mathcal{T}_\sigma(\forall x\varphi) = 1$ iff $\{a \in D | \mathcal{T}_{\sigma[a/x]}(\varphi) = 1\} \in Q$.

Finally, an interpretation $\mathcal{T}$ is consistent$^+$ with $\vdash_\mathcal{L}$ if it is truth-preserving given *any* function $I$ for non-logical vocabulary. In other words, it is consistent$^+$ with $\vdash_\mathcal{L}$ if, given any $I$, we have that $\Gamma \vdash_\mathcal{L} \varphi$ and $\mathcal{T}_\sigma(\Gamma) = 1$ imply $\mathcal{T}_\sigma(\varphi) = 1$ (where $\Gamma$ and $\varphi$ can be formulas, not just sentences).[17]

[16] A different way to argue for the same point is to note that the extension of non-logical predicates changes over time, but we continue to endorse valid arguments regardless of how it changes. If I'm told that some red object has been destroyed, I don't pause to consider whether, given the new extension of 'red', the inference from 'Everything is red and sticky' to 'Everything is red' remains acceptable. To account for this fact, it seems, we must give $\wedge$ and $\forall$ semantic values that make the argument truth-preserving regardless of how non logical vocabulary is interpreted.

[17] Bonnay and Westerståhl use a similar approach, *mutatis mutandis*, when they consider Carnap's Problem in the context of possible worlds semantics for propositional logic. They say quantifying over interpretations in the definition of consistency is only done 'for simplicity' [3, p. 731], but it can be shown that it is a necessary condition for the strategy to work.

If we look at things this way, it's easy to show that the role of classical logical vocabulary in inferences does rule out non-normal interpretations:

**Lemma 10** *Let $\mathcal{T} = (F_\wedge, F_\neg, Q)$ be an interpretation (with underlying domain $D$) for a language $\mathcal{L}$. If $\mathcal{T}$ is consistent$^+$ with $\vdash_\mathcal{L}$, then $\mathcal{T}$ is normal.*

***Proof*** We'll work by contraposition:

Conjunction: Suppose e.g. that $F_\wedge(1, 1) = 0$. Take any $\varphi, \psi$ and $\sigma$ such that $\mathcal{T}_\sigma(\varphi) = \mathcal{T}_\sigma(\psi) = 1$ (there must be some such $\varphi$, $\psi$ and $\sigma$, because $\vdash_\mathcal{L}$ contains tautologies and $\mathcal{T}$ is consistent$^+$ with it, so tautologies must have value 1 under any $\sigma$). Since $F_\wedge(1, 1) = 0$, we have that $\mathcal{T}$ invalidates $\varphi, \psi \vdash_\mathcal{L} \varphi \wedge \psi$. Remaining cases are similar.

Negation: Suppose e.g. that $F_\neg(0) = 0$. Set $I(P) = \emptyset$ ($P$ is a predicate symbol of any arity. For simplicity we'll take to be unary) and let $\sigma$ be an arbitrary assignment. Since $I(P) = \emptyset$ we must have $\mathcal{T}_\sigma(Px) = 0$. Then $\mathcal{T}_\sigma(Px \wedge \neg Px) = 0$, since one conjunct is false. But $F_\neg(0) = 0$, so $\mathcal{T}_\sigma(\neg(Px \wedge \neg Px)) = 0$. Thus, $\mathcal{T}$ invalidates the tautology $\vdash_\mathcal{L} \neg(Px \wedge \neg Px)$. Remaining cases are similar.

Universal Quantifier: We have to show that $\{D\} = Q$. First, suppose for a contradiction that $D \notin Q$. Since $\vdash_\mathcal{L} \neg(Px \wedge \neg Px)$, we must have $\mathcal{T}_{\sigma'}(\neg(Px \wedge \neg Px)) = 1$ for *all* $\sigma'$. We'll use the abbreviation $\varphi := \neg(Px \wedge \neg Px)$ for readability. Now, let $\sigma$ be arbitrary. Then $\{a \in D | \mathcal{T}_{\sigma[a/x]}(\varphi) = 1\} = D \notin Q$, so $\mathcal{T}_\sigma(\forall x \neg(Px \wedge \neg Px)) = 0$. Therefore, $\mathcal{T}$ invalidates $\vdash_\mathcal{L} \forall x \neg(Px \wedge \neg Px)$, contradicting consistency$^+$. Next, suppose that there is some $C \subset D$ in $Q$. Let $I(P) = C$, and let $\sigma'$ be an assignment such that $\sigma'(x) \notin C$. Then $\{a \in D | \mathcal{T}_{\sigma[a/x]}(Px) = 1\} = C \in Q$, so $\mathcal{T}_\sigma(\forall x Px) = 1$. But $\mathcal{T}_\sigma(Px) = 0$, so $\mathcal{T}$ invalidates $\forall x Px \vdash_\mathcal{L} Px$, contradicting consistency$^+$.

□

A couple of comments are in order. First, it should be clear that quantifying over interpretations in the definition of consistency$^+$ has more or less the same effect as second-order variables in Bonnay and Westerståhl's account. The current option, however, has a clear motivation rooted in inferential practice. Moreover, it is a weaker assumption than the alternative. Second-order variables make every subset of $D$ and $D^n$ (for any $n$) definable, whereas it's not in general the case that, given a language $\mathcal{L}$ and a subset of $D$ or $D^n$, we can always find a function $I$ under which that subset is definable.

Secondly, the way I've defined interpretations presupposes what type of semantic value corresponds to each syntactic category: connectives are interpreted by truth-functions, and quantifiers by sets of subsets of the domain. This is something Bonnay and Westerståhl also take for granted:

> One must choose the semantic values of expressions belonging to a given syntactic category [...] Our hypothetical language learner already knows, or guesses, what kind of language is to be learnt: what the syntactic categories are, and what kinds of things expressions of these categories stand for. [3, p. 726]

I will say more about this assumption in the next section. For now, it's enough to note that it doesn't make interpretations compositional. This is a feature, not a bug: as we've seen, compositionality (in the strict sense preferred by Bonnay and Westerståhl) is too strong a requirement, and part of what gets them into trouble. It's also important to keep in mind that Bonnay and Westerståhl motivate the requirement of compositionality through the usual learnability argument: natural languages have infinitely many sentences, so speakers can't learn their meanings one by one. Compositionality, then, seems like 'our currently best explanation of learnability' [3, p. 725]. As it is often pointed out,[18] however, learnability arguments don't establish the need for compositionality in the strong sense demanded by Bonnay and Westerståhl, so nothing is lost if we abandon it.

With this in mind, it seems that the current proposal fares better than Bonnay and Westerståhl's. Their approach assumes compositionality, non-triviality, invariance under permutations, and doesn't work for first-order languages, since it is essentially tied to the use of second-order variables. The current approach starts from a similar notion of interpretation and only uses consistency$^+$, a substitute for predicate variables motivated by our inferential practice.

Finally, it may be useful to close the section by comparing the current proposal to another family of solutions to Carnap's Problem, those built around the notion of 'open-endedness'. In his [14] Vann McGee attempts to read classical semantics from certain *inference rules* for classical logic, rather than from the resulting consequence relation. According to McGee those rules are open-ended, meaning that 'they are valid [...] not only within the language $\mathcal{L}$, but will remain valid however the language may be enriched by the addition of new sentences' [14, p. 65]. Open-endedness is put to use as follows: McGee holds that, given any class of models, there is some 'mathematically possible language' which contains a sentence that is true in all and only those models. He then shows that any non-normal interpretation of a first-order language $\mathcal{L}$ invalidates a classically valid inference *when extended to a richer language* $\mathcal{L}'$, where $\mathcal{L}'$ contains new sentences that are true only in certain *ad hoc* classes of models; non-normal interpretations, then, are ruled out on the grounds that they violate open-endedness.

Clearly, open-endedness plays a similar role to quantification over interpretations of non-logical vocabulary: it bypasses the expressive limitations of first-order languages, and ensures that each non-normal interpretation invalidates a classically valid inference. At the same time, open-endedness is stronger and more problematic than consistency$^+$. McGee assumes that interpretations which correctly account for the role of logical vocabulary make valid inferences truth-preserving *across all mathematically possible extensions of a language*. But firstly, it's not clear what the range of mathematically possible extensions of a language consists in, and the extent to which we need to modify the notion of a model as we enrich first-order languages with arbitrary sentences. And secondly, it is dubious that acceptance of the rules of classical logic is (in general) open-ended. For instance, it's fairly common to hold that

[18] See e.g. [18, 25].

classical logic has to be abandoned when first-order languages are extended in certain ways (say, with vague predicates, or with truth-predicates).[19] In contrast, here we only assume that, given our inferential practice, an adequate interpretation must make valid arguments truth-preserving under *well-behaved* reinterpretations of the non-logical vocabulary *that we already have.*[20]

Murzi and Topey [16] have recently put forward a solution similar to McGee's. Like McGee, they attempt to read classical semantics from rules of inference, and take those rules to be open-ended. They cash out open-endedness more modestly than McGee, though: for Murzi and Topey rules must remain valid across all extensions of a language obtained by adding new predicate letters and individual constants. This is, again, stronger than consistency$^{+}$, the assumption that valid inferences must remain truth-preserving across reinterpretations of the *current* non-logical vocabulary. Murzi and Topey also assume which type of semantic value can be given to quantifiers.[21] And finally, the inference rules from which they read the semantics are quite unusual: they use a higher-order natural deduction system in the style of [22]. In standard natural deduction systems the rules conclude and discharge formulas. In higher-order systems, on the other hand, rules can conclude and discharge *other rules*. Say that a rule that discharges a formula is of level 1, and a rule that discharges a rule of level $n$ is itself of level $n + 1$. Then Murzi and Topey's system allows one to conclude and discharge rules of arbitrary finite level.

It's unclear that a system of this type is a faithful representation of our use of logical vocabulary. This, I think, makes it unsuitable to show that the meaning of logical constants is determined by the way we (actually) use them. Murzi and Topey arguably only show that classical semantics can be read off non-standard rules that deviate from inferential practice. The proposal in this section, on the other hand, reads the semantics of logical vocabulary from the classical consequence relation itself. It is blind to the format and specific formulation of whatever rules we take to govern logical vocabulary, and for that reason is immune to this sort of problem.[22]

[19] There are other problems with McGee's approach. The proof-system –due to Mates [12]– from which he reads the semantics of logical vocabulary is only complete for languages containing infinitely many individual constants, and he imposes some extra constraints on interpretations in order to secure the normal interpretation of quantifiers. I will say a smidgen more about this in the next section.

[20] Where 'well-behaved' just means that the re-interpretations give standard, unproblematic semantic values to non-logical vocabulary, and don't smuggle in vagueness, self-reference and the like.

[21] Their language sometimes suggests they take the denotation of $\forall$ to be a *subset* of the domain of quantification, rather than a set of subsets [see16, p.3407]. If this is so, it's unclear how the approach could be extended to other quantifiers like $\exists$.

[22] Julien Murzi has kindly pointed out to me that although [16] uses higher-order rules, their approach can avoid Carnap's Problem with more modest assumptions. It turns out that they can read the usual semantics from any natural deduction system *that allows for empty succedents* (i.e. empty conclusions). Note, however, that this is still a restriction to a specific, and fairly non-standard, type of proof-system. In contrast (see below) the current approach reads the usual semantics from *any* standard calculus for classical logic.

In fact, the present approach also allows us to read classical semantics off any standard calculus for classical logic. I'll take $\forall$ as an example, but the cases for $\wedge$ and $\neg$ are similar. Consider these typical natural deduction rules (in sequent notation) for $\forall$:

$$\frac{\Gamma \Rightarrow \varphi(y/x)}{\Gamma \Rightarrow \forall x\varphi}(\forall \text{ I}) \qquad \frac{\Gamma \Rightarrow \forall x\varphi}{\Gamma \Rightarrow \varphi(y/x)}(\forall \text{ E})$$

Rule ($\forall$ I) has the usual restrictions: $y$ must be free for $x$ in $\varphi$, $\varphi(y/x)$ is the result of replacing all free occurrences of $x$ in $\varphi$ for $y$, and $y$ must not occur free in $\Gamma$ or $\forall x\varphi$. Now, given a model $\mathcal{M} = (D, I)$, a sequent $\Delta \Rightarrow \psi$ is *valid* in $\mathcal{M}$ if, for any assignment $\sigma$, we have that $\mathcal{M}, \sigma \models \Delta$ implies $\mathcal{M}, \sigma \models \psi$. Similarly, a rule **R** *preserves validity* in $\mathcal{M}$ if whenever the premiss-sequents of an application of **R** are valid in $\mathcal{M}$, so is the conclusion-sequent. Murzi, Topey, and McGee read semantics off a rule **R** by looking at which interpretations are such that **R** preserves validity in them (the 'rule analogue' of consistency with a consequence relation). We can adapt the current approach to this setting. It's clear that we often draw inferences according to ($\forall$ I) and ($\forall$ E) –and take them to be valid– without knowing the interpretation of the non-logical vocabulary involved. Therefore, to adequately represent the inferential role of $\forall$, we must give it a semantic value that makes inferences sanctioned by ($\forall$ I) and ($\forall$ E) valid regardless of the interpretation of non-logical vocabulary. Let's make this more formal. Given an interpretation $\mathcal{T}$ and some $I$ for the non-logical vocabulary, we'll say that $\Delta \Rightarrow \psi$ is valid in $\mathcal{T}$ if $\mathcal{T}_\sigma(\Delta) = 1$ implies $\mathcal{T}_\sigma(\psi) = 1$ (for any $\sigma$). Moreover, an interpretation $\mathcal{T} = (F_\wedge, F_\neg, Q)$ is consistent$^+$ with a *rule* **R** if, given *any* $I$, whenever the premiss-sequents of **R** are valid in $\mathcal{T}$, so is its conclusion-sequent. Then it's easy to show that:

**Lemma 11** *Let $\mathcal{T} = (F_\wedge, F_\neg, Q)$ an interpretation (with underlying domain $D$) for a language $\mathcal{L}$. If $\mathcal{T}$ is consistent$^+$ with ($\forall$ I) and ($\forall$ E) then $Q = \{D\}$.*

***Proof*** We'll assume for simplicity that $\mathcal{L}$ has a unary predicate $P$, and work by contraposition. Suppose $D \notin Q$. Let $I(P) = D$. Then we have that $\mathcal{T}_\sigma(Px) = 1$ and $\mathcal{T}_\sigma(\forall x Px) = 0$ for any assignment $\sigma$. In other words, $\Rightarrow Px$ is valid, and $\Rightarrow \forall x Px$ is not. But we can infer $\Rightarrow \forall x Px$ from $\Rightarrow Px$ by ($\forall$ I), so $\mathcal{T}$ is not consistent$^+$ with ($\forall$ I). Similarly, suppose there is some $C \subset D$ in $Q$. Let $I(P) = C$, and let $\sigma'$ be an assignment such that $\sigma'(x) \notin C$. Then we have that $\mathcal{T}_\sigma(\forall x Px) = 1$ for any $\sigma$, but $\mathcal{T}_{\sigma'}(Px) = 0$. Therefore $\Rightarrow \forall x Px$ is valid but $\Rightarrow Px$ is not. Since we can infer the latter from the former using ($\forall$ E), $\mathcal{T}$ is not consistent$^+$ with ($\forall$ E). □

A similar argument works for any other standard set of natural deduction (or sequent) rules for $\forall$, and the corresponding result for the connectives is easy to prove. To summarise, then: the approach in this section also seems to improve on solutions to Carnap's Problem built around open-endedness. It avoids the problematic aspects of the claim that our acceptance of classical rules of inference is open-ended, and it allows one to read classical semantics from standard calculi for classical logic.

## 6 Carnap's Problem and Interpretations

Let's take a step back and put what we have done so far in perspective. We began with a relatively simple question: whether the inferential role of (classical) logical vocabulary rules out its non-normal interpretations. This question is sensible only insofar as we can explain three things:

(i) What we mean by 'interpretations'.
(ii) What we mean by 'inferential roles'.
(iii) What we mean inferential roles 'ruling out' interpretations.

The literature on Carnap's Problem contains very different answers to (i)-(iii). We have already seen the two standard ways of spelling out what inferential roles amount to: in terms of consequence relations and in terms of inference rules. The approach I've put forward works with both. We have also seen that 'ruling out' is usually cashed out in terms of truth-preservation. I've argued that it is better explained in terms of truth preservation across all reinterpretations of non-logical vocabulary, or consistency$^{+}$. This takes care of (ii) and (iii); now it's time to say more about (i). In this section I'll go through some ways interpretations have been defined, and motivate my own approach.

The literature seems split into two camps. According to the first, interpretations are assignments of truth-conditions to formulas. According to the second, interpretations are assignments of semantic values to subsentential expressions (and derivatively, to formulas). I followed Bonnay and Westerståhl down the second route, so let's see some examples of the first type for contrast.

Garson [7] approaches truth-conditions in terms of valuations. He distinguishes a 'substitutional' and an 'objectual' reading of quantifiers. On the substitutional reading a valuation $v$ is $\forall$-normal when:[23]

(s$\forall$) $v(\forall x\varphi) = 1$ iff $v(\varphi[y/x]) = 1$ for all variables $y$ of the language.

The question that immediately arises is whether (s$\forall$) is a plausible notion of *normal* interpretation. And arguably, it isn't. Note that a first-order semantics in terms of (s$\forall$) doesn't so much as mention a domain of quantification, which puts it quite far from the usual way(s) of interpreting quantifiers. But the worry goes deeper. First, (s$\forall$) doesn't really match the standard truth-conditions of quantified sentences: in a (normal) classical model, we may have that $\mathcal{M}, \sigma \models Py$ for all variables $y$ but $\mathcal{M}, \sigma \not\models \forall x\, Px$. And secondly, as Garson himself points out [7, p. 215], (s$\forall$) induces a non-compact consequence relation. To see this, fix a language with a unary predicate $P$ and no individual constants or functions symbols, and consider the argument from all sentences of the form $Py$, for $y$ a variable of the language, to $\forall x\, Px$:

$$\{P[y/x] \mid y \in \text{Vars}\} \vdash \forall x\, Px$$

This argument is truth-preserving on all valuations that satisfy (s$\forall$), and therefore valid on Garson's 'normal' substitutional semantics. The result of taking any finite

[23] As usual, if a language contains individual constants and function symbols, (s$\forall$) can be modified in the obvious way to account for terms.

subset of $\{P[y/x] \mid y \in \text{Vars}\}$ as premises, on the other hand, is not. Needless to say, no standard proof-system for classical logic is strongly complete for this substitutional semantics.

The objectual reading of quantifiers is more involved. Garson begins by introducing the notion of a 'hybrid formula'. Given a domain $D$, hybrid formulas are recursively defined so that any well-formed-formula is a hybrid formula, and the result $\varphi[d/x]$ of substituting an object $d \in D$ for all free occurrences of $x$ in a hybrid formula $\varphi$ is a hybrid formula itself. Garson then considers valuations that range *over hybrid formulas*, and takes $v$ to be $\forall$-normal, given a domain $D$, when:

(o$\forall$) $v(\forall x\varphi) = 1$ iff $v(\varphi[d/x]) = 1$ for all $d \in D$.

Although (o$\forall$) is, perhaps, more natural than (s$\forall$), there's still reason to doubt it's adequacy. We were after a definition of normal valuations over a language $\mathcal{L}$. What we've been given, however, is a definition of normal valuations over a 'hybrid' (i.e. extended) language $\mathcal{L}'$. To avoid this sort of problem Brîncuş [4] restricts attention to countable domains all of whose of elements are named by a constant, which effectively collapses Garson's objectual and substitutional readings. But this doesn't really solve the issue: he is just left without a notion of normal interpretation applicable to all languages.[24]

McGee [14] takes interpretations to be Tarski-style clauses relating sentences (closed formulas) and models. A first reason to worry is that it's not clear what notion of model he has in mind. As we've seen, McGee wants to consider 'all mathematically possible extensions' of a given language, which presumably stretches the notion of a model accordingly, but we are not told how. There seem to be other problems as well. The usual clause stating the truth-conditions of $\forall$ is the well-known:

(iv) $\mathcal{M}, \sigma \models \forall x\varphi$ iff $\mathcal{M}, \sigma[a/x] \models \varphi$ for all $a \in D$.

Or equivalently, when $\forall$ is seen as a generalised quantifier:

(iv)$'$ $\mathcal{M}, \sigma \models \forall x\varphi$ iff $\{a \in D \mid \mathcal{M}, \sigma[a/x] \models \varphi\} \in \{D\}$.

In order to avoid the complications that arise from considering variable assignments and satisfaction relations for open formulas, McGee states his normal interpretation of universally quantified sentences in terms of other sentences only. If $c$ is an individual constant, say that two models are $c$-variants when they differ only in the value they assign to $c$. McGee's normal interpretation of the universal quantifier is:

(m$\forall$) If $c$ does not appear in $\varphi$, then $\forall x\varphi(x)$ is true in $\mathcal{M}$ iff $\varphi[c/x]$ is true in every $c$-variant of $\mathcal{M}$. [14, p. 71]

Note, however, that this clause is *not* equivalent to the usual (iv) and (iv)$'$: it yields the classical truth-conditions for quantified sentences only in languages with an infinite

[24] A reviewer asks how Lemmas 10 and 11 relate non-normal valuations. The approaches are difficult to compare and, for the reasons I've just explained, I doubt Garson's (s$\forall$) and (o$\forall$) are adequate definitions of normality. But in any case, given a normal interpretation (in my sense) $\mathcal{T}$, we have that for any $I$, $\mathcal{T}_\sigma(\forall x\varphi) = 1$ iff $\mathcal{T}_{\sigma[a/x]}(\varphi) = 1$ for all elements $a$ of the domain. In the special case where all elements are named by a constant, this is equivalent to $\mathcal{T}_\sigma(\forall x\varphi) = 1$ iff $\mathcal{T}_\sigma(\varphi[t/x]) = 1$ for all terms of the language.

number of individual constants.[25] In other words, we are saddled again with a notion of normal interpretation that is neither genuinely normal nor applicable to all languages.

Things are different on the second approach, where interpretations are assignments of semantic values to subsentential expressions. The usual set-up (at the propositional level) has interpretations assign truth-functions to connectives (see Peacocke [19], Hodes [9] and Hacking [8]). Bonnay and Westerståhl extend this to the first-order case by taking quantifiers to be interpreted by sets of subsets of the domain, and I have followed suit. The resulting notion of normal interpretation is both recognisably normal and applicable to all first-order languages, unlike Garson's and McGee's. In that sense, I take it, it is a better way of clarifying (i), a better notion of interpretation.

Admittedly, this second approach takes for granted that each syntactic category is interpreted by means of the usual semantic type. Now, some degree of stipulation is just inevitable: if we take interpretations to be assignments of semantic values, we have to specify what sorts of things semantic values are. And more generally, *any* way of defining interpretations rules out some options from the get-go.[26] The notion of interpretation I've used is a natural, straightforward generalisation of standard first-order semantics. But still, some could complain that it is not 'general enough'. I'm not sure of how much a definition of interpretations can take for granted before it is 'too much'. And while I *do* think that a reasonable definition of interpretations should be applicable to languages of arbitrary signature, I don't expect the approach I've settled on to be the only way to achieve this. But in any case, readers who think my definition of interpretations is too narrow can relativise the results of Section 5. What Section 5 shows is that *if* connectives and quantifiers are interpreted with the usual semantic types, then they must be given their normal semantic values. Or put more suggestively: that if connectives and quantifiers are interpreted *as* connectives and quantifiers, they must be interpreted normally. This is not the last word on Carnap's Problem. It couldn't be: there are as many 'Carnap's Problems' as there are ways to define interpretations. But it does, I hope, help to clarify the relation between what logical constants mean and how they are used in arguments.

## 7 Concluding remarks

Moderate inferentialists hold that the meaning of logical vocabulary is determined by the way it's used in inferences. There are multiple ways of making this claim more precise, but they all have to grapple with the same problem: ruling out non-normal interpretations. Bonnay and Westerståhl's way around the problem doesn't work, for two reasons. The first is that we can exploit the undefinability of subsets of the domain of a structure to create non-normal interpretations that satisfy all of Bonnay

[25] If a language has finitely many individual constants, there will be quantified sentences which contain all of them, rendering (m∀) inapplicable. Incidentally, in the course of his argument McGee smuggles in two further assumptions. The first [14, p. 70] is that there is no model in which all sentences are true, the analogue of Bonnay and Westerståhl's non-triviality. The second [14, p. 71] is that if a constant $c$ does not occur in $\psi$, the class of models in which $\psi$ is true must be closed under $c$-variants.

[26] For instance, when defining interpretations Garson and McGee assume that there are two truth-values, no gaps and no gluts; by doing so they rule out supervaluational and many-valued semantics from the outset.

and Westerståhl's constraints. The second is that the normal, two-valued semantics for first-order logic doesn't satisfy their requirement of compositionality, and more complex semantics make Carnap's Problem intractable.

Something close to Bonnay and Westerståhl's solution, however, works better. The key is to keep in mind that interpretations that do justice to the way we use logical constants must assign them semantic values that make valid arguments truth-preserving regardless of how non-logical vocabulary is interpreted. And once we do this, it's easy to show that the classical consequence relation (or any standard proof-system for classical logic) rules out non-normal interpretations.

**Acknowledgements** I would like to thank two anonymous reviewers from this journal, as well as Bogdan Dicher, Bruno Jacinto, Julian J. Schlöder, Julien Murzi and Bas Kortenbach for their very helpful comments on this material. Earlier versions of this paper were presented at the *LanCog Seminar* in the University of Lisbon, the *Logic Seminar* of the Scuola Normale Superiore, and at *Logica 2023* and *PhD's in Logic 2023*. I'm also grateful to the audiences of these events for their valuable feedback.

**Author Contributions** Not applicable

**Funding** Open access funding provided by Scuola Normale Superiore within the CRUI-CARE Agreement.

**Availability of data and materials** Not applicable

## Declarations

**Ethical Approval** Not applicable

**Competing interests** No competing interests



## References

1. Antonelli, A. (2013). On the general interpretation of first-order quantifiers. *Review of Symbolic Logic, 6*(4), 637–658.
2. Antonelli, A. (2017). Completeness and decidability of general first-order logic. *Journal of Philosophical Logic, 46*(3), 233–257.
3. Bonnay, D., & Westerstahl, D. (2016). Compositionality solves carnap's problem. *Erkenntnis, 81*(4), 721–739.
4. Brîncuş, C. C. (2024). Categorical quantification. *The Bulletin of Symbolic Logic*, (pp. 1–27).
5. Carnap, R. (1943). *Formalization of Logic*. Cambridge, Mass.,: Harvard university press.
6. Dosen, K. (1989). Logical constants as punctuation marks. *Notre Dame Journal of Formal Logic, 30*(3), 362–381.
7. Garson, J. W. (2013). *What Logics Mean: From Proof Theory to Model-Theoretic Semantics*. New York: Cambridge University Press.
8. Hacking, I. (1979). What is logic? *Journal of Philosophy, 76*(6), 285–319.

9. Hodes, H. (2004). On the sense and reference of a logical constant. *The Philosophical Quarterly, 54*(214), 134–165.
10. Janssen, T. M., & Partee, B. H. (2011). Compositionality. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of Logic and Language* (pp. 417–473). Amsterdam: North-Holland.
11. Kreisel, G., & Krivine, J. L. (1967). *Éléments de Logique Mathématique Théorie des Modèles*. Dunod.
12. Mates, B. (1972). *Elementary Logic*. Oxford University Press.
13. McGee, V. (1996). Logical operations. *Journal of Philosophical Logic, 25*(6), 567–580.
14. McGee, V. (2000). Everything. In G. Sher & R. Tieszen (Eds.), *Between Logic and Intuition: Essays in Honor of Charles Parsons* (pp. 54–78). Cambridge University Press.
15. Monk, J. D. (1976). *Mathematical Logic*. New York: Springer Verlag.
16. Murzi, J., & Topey, B. (2021). Categoricity by convention. *Philosophical Studies, 178*(10), 3391–3420.
17. Pagin, P., & Westerstahl, D. (2010). Compositionality I: Definitions and variants. *Philosophy Compass, 5*(3), 250–264.
18. Pagin, P., & Westerstahl, D. (2010). Compositionality II: Arguments and problems. *Philosophy Compass, 5*(3), 265–282.
19. Peacocke, C. (2004). Understanding logical constants: A realist's account. In T. J. Smiley, & T. Baldwin (Eds.), *Studies in the Philosophy of Logic and Knowledge*, (p. 163). Published for the British Academy by Oxford University Press.
20. Peregrin, J. (2014). *Inferentialism: Why Rules Matter*. London and New York: Palgrave-Macmillan.
21. Rumfitt, I. (2000). Yes and no. *Mind, 109*(436), 781–823.
22. Schroeder-Heister, P. (1984). A natural extension of natural deduction. *The Journal of Symbolic Logic, 49*(4), 1284–1300.
23. Shoesmith, D. J., & Smiley, T. (1978). *Multiple Conclusion Logic*. Cambridge, England / New York London Melbourne: Cambridge University Press.
24. Smiley, T. (1996). *Rejection. Analysis, 56*(1), 1–9.
25. Szabó, Z. G. (2012). The case for compositionality. In *The Oxford Handbook of Compositionality*. Oxford University Press.
26. Wehmeier, K. F. (2018). The proper treatment of variables in predicate logic. *Linguistics and Philosophy, 41*(2), 209–249.
27. Zimmermann, T. E., & Sternefeld, W. (2013). *Introduction to Semantics*. Berlin, Boston: De Gruyter Mouton.