What Is Carnap's Problem and How Can We Solve It?

Julien Murzi
University of Salzburg
Salzburg, Austria
julien.murzi@plus.ac.at

Brett Topey
University of Salzburg
Salzburg, Austria
brett.topey@plus.ac.at

Abstract

On a plausible approach to logical metasemantics, our dispositions to treat a logical expression's I- and E-rules as valid determine its contribution to the truth conditions of sentences in which it appears. Carnap's so-called Categoricity Problem is that the rules in question don't seem to fix a unique interpretation of our logical vocabulary: there appear to be deviant interpretations of both the connectives and the quantifiers that are compatible with the validity of their rules. And although standard responses are available to Carnap's problem as it applies to propositional logic (by appeal, e.g., to bilateral rules or to a local notion of validity), the case of the quantifiers is more difficult. Here we develop a more precise account of how Carnap-style arguments work than has ever before been given, one that makes clear why certain such arguments succeed while others fail. In so doing, we demonstrate that despite recent criticisms, the account of the categoricity of the quantifiers we offered in previous work isn't threatened by any of the alleged deviant interpretations that have been discussed in the literature: each of these either is incompatible with the validity of the quantifier rules or else results in an illegitimate Carnap-style argument.

1 Introduction

Carnap's so-called Categoricity Problem is a metasemantic problem for moderate inferentialism: the plausible (by our lights) view that use determines meaning. More specifically, in the case of a logical operator \$, moderate inferentialists hold that our dispositions to treat \$'s I- and E-rules as valid categorically determine \$'s contribution to the truth conditions of the sentences in which it appears. Moderate inferentialism is inferentialist, in that it's a view on which the interpretation of a given expression is a product of the inference rules that govern its use, but it's also moderate, in that it's based on an entirely orthodox reference-based view of what interpretations are. It can also, unlike many philosophical views, be *tested*: since both the standard reference-based interpretation of our logical expressions and their I- and E-rules are relatively well understood, one can check, for any logical operator \$, whether the rules governing its use really do determine its interpretation. Carnap's problem, then, is that there seem to be deviant interpretations of both the connectives and the quantifiers that are nevertheless compatible with the validity of their I- and E-rules – i.e., that the validity those rules seemingly fails to categorically determine the interpretation of our logical expressions, in which case moderate inferentialism can't be correct. Consider, for instance, the trivial valuation: a valuation v^* that makes every sentence true, including 'A' and '$\neg A$'. This valuation is incompatible with the standard truth conditions for negation, and yet it appears to be compatible with the validity of all natural deduction rules, including the I- and E- rules for the negation sign.

Moderate inferentialism, though, is a highly unspecific doctrine: it entails that the validity of I- and E-rules fixes the interpretation of logical expressions, but it doesn't tell us how to formalise logic, nor does it tell us how to understand the notion of validity. And as we now know, Carnap's problem is highly sensitive to both the way logic is formalised and the way validity is understood. For instance, bilateral and multiple-conclusion formalisations of classical propositional logic (CPL) are categorical (see, e.g., Smiley 1996; Rumfitt 2000; Raatikainen 2008;

Murzi and Hjortland 2009), and so indeed can be standard single-conclusion formalisations, provided that validity is understood as *local* (see, e.g., Garson 2013; Murzi and Topey 2021). So there are plausible responses available to Carnap's problem for CPL.

Of course, natural language contains quantifiers in addition to propositional connectives, and so a full solution to Carnap's problem must generalize to first-order logic (FOL), and arguably also to second-order logic (SOL) and beyond. But here things are more complex. Bonnay and Westerståhl (2016) have argued that on certain semantic assumptions, the rules of FOL categorically determine the interpretation of the first-order quantifiers. Their claim, though, is highly problematic on at least two counts. First, the semantic assumptions in question are arguably illicit: the moderate inferentialist's claim was that our use of logical expressions determines their interpretation *on its own*, not that it does so when certain semantic facts are already in place. Second, as we show, Bonnay and Westerståhl's work depends on a mistaken understanding of what it is for a model to be deviant – a mistake that's inherited by del Valle-Inclán's (2024) recent critique of their proposal. Further complicating matters is the fact that we have argued in previous work – see our "Categoricity by Convention" (2021) – that the validity of the rules of FOL and SOL determines their standard, unrestricted interpretations, provided those rules are understood as *open-ended* – a conclusion that, according to Brîncuş (2024a) and Brîncuş (2024b), is disproved by deviant interpretations of the quantifiers given by Carnap (1943) and Garson (2013).

Here we offer a more precise understanding of Carnap-style objections to moderate inferentialism than has appeared in the literature before, an understanding that makes it clear why certain Carnap-style arguments succeed – in the sense that they genuinely show that certain ways of formalising logic, or certain notions of validity, are unavailable to moderate inferentialists – while others fail. In so doing, we demonstrate that despite recent criticisms, our own account of the categoricity of the quantifiers isn't threatened by the alleged deviant interpretations of the quantifiers offered by Carnap, Garson, Briîncuş, or del Valle-Inclán: each of these either is incompatible with the validity of the quantifier rules (as we formalise them) or else results in an illegitimate Carnap-style argument. The result is that moderate inferentialism, as we develop it, turns out to be tenable after all.

2 Background

We said that, on our view, the interpretation of a logical expression e is determined by our dispositions to treat the basic rules for e as valid. But what is a rule and what is our conception of validity? If e is a logical expression, we take e's basic rules to be its I- and E-rules in a *sequent-style single-conclusion natural deduction system*. For instance, we formalise the basic rules for $\rightarrow$ as follows:

$$\frac{\Gamma, \varphi \vdash \psi}{\Gamma \vdash \varphi \rightarrow \psi} \rightarrow\text{-I} \qquad \frac{\Gamma \vdash \varphi \rightarrow \psi \qquad \Delta \vdash \varphi}{\Gamma, \Delta \vdash \psi} \rightarrow\text{-E}$$

As for validity, there are three possible accounts:

(i) validity as preservation of truth;

(ii) validity as preservation of sequent satisfaction (local validity); and

(iii) validity as preservation of sequent validity (global validity).

Simple truth preservation doesn't apply to metainferences such as $\rightarrow$-I, $\neg$-I, $\vee$-E etc. So our choice is restricted to (ii) and (iii), which apply to (meta)rules whose inputs and outputs are sequents.

We begin by defining the notion of sequent satisfaction:

Definition 2.1 (Sequent satisfaction)**.** A valuation v satisfies a sequent $\Gamma \vdash \varphi$ iff either v makes some $\psi \in \Gamma$ false or v makes φ true.

We can then define global and local validity as follows:

Definition 2.2 (Global and local validity)**.** A metarule

$$\frac{\Gamma_1 \vdash \varphi_1 \quad \dots \quad \Gamma_n \vdash \varphi_n}{\Delta \vdash \psi}$$

- is locally valid with respect to a class of valuations V iff it preserves satisfaction – i.e. iff for all valuations $v \in V$, if v satisfies every $\Gamma_i \vdash \varphi_i$, it also satisfies $\Delta \vdash \psi$; and
- is globally valid with respect to V iff it preserves validity – i.e. iff, if all $v \in V$ satisfy every $\Gamma_i \vdash \varphi_i$, then all $v \in V$ satisfy $\Delta \vdash \psi$.

We take validity to be local validity, essentially because, in our setting, it is a generalisation of truth preservation. After all, if we let $\Gamma_1, \dots, \Gamma_n$ and Δ be empty, on the local account, a meta-rule

$$\frac{\Gamma_1 \vdash \varphi_1 \quad \dots \quad \Gamma_n \vdash \varphi_n}{\Delta \vdash \psi}$$

is valid if and only if the inference from $\varphi_1, \dots, \varphi_n$ to ψ preserves truth in v for all v's, as, we think, it should be.

Following Tennant (1999), we interpret '$\bot$' as a logical punctuation sign (formally, $\bot = \varnothing$). On this assumption, the local conception of validity gets us the determinacy of CPL, via the following theorem proved by Garson (2013):

Theorem 2.3 (Garson's Local Validity Theorem)**.** *The rules of* CPL *are locally valid with respect to a class of vlauations V only if all members of V obey the classical truth tables.*

Let us now move on to FOL. In general, establishing that our treating certain rules of inference as valid categorically determines particular truth-conditional meanings for our expressions involves showing that, among all the possible ways of assigning truth values to formulae – i.e., all *valuations* or *general models* – the *only* ones that are compatible with the validity of the rules are those that respect the truth-conditional meanings in question (where respecting those truth-conditional meanings amounts to satisfying the associated semantic clauses). But since we are interested here in the question of the categoricity of our *logical* expressions in particular, we can take for granted that nonlogical expressions have meanings of the usual sort, with each constant designating some member of a domain of objects, each predicate having as its extension some set of tuples of objects from the domain, and so on. So we can let the set of general models be the set of triples $\langle D, I, [\![]\!] \rangle$, where D and I are a domain and interpretation function of the usual sort and where $[\![]\!]$ is a function from pairs of formula and variable assignment to truth values that assigns truth values to atomic formulae in the usual way but may assign truth values to logically complex formulae in any way whatsoever. In this setting, solving Carnap's problem amounts to proving, just from the assumption that our logical operators' I- and E-rules are valid across some set of general models, that all of those models satisfy the usual truth-conditional semantic clauses for the operators in question.

We interpret natural deduction rules as being rules for formulae rather than for sentences. Thus, we adopt the following version of $\forall$-I:

$$\frac{\Gamma \vdash \varphi}{\Gamma \vdash \forall x \varphi} \forall\text{-I}$$

where x doesn't appear free in Γ. And since we now allow arbitrary formulae to occur in sequents, we must generalise our definitions of (sequent) satisfaction and validity.

Definition 2.4 (Sequent satisfaction relative to a variable assignment)**.** Let α be a variable assignment over v's domain. Then, α satisfies$_\alpha$ $\Gamma \vdash \varphi$ iff either for some $\psi \in \Gamma$, $[\![\psi]\!]_{v,\alpha} \neq 1$ or $[\![\varphi]\!]_{v,\alpha} = 1$.

Definition 2.5 (Sequent satisfaction, generalised)**.** A valuation v satisfies $\Gamma \vdash \varphi$ iff, for every variable assignment α over v's domain, α satisfies$_\alpha$ $\Gamma \vdash \varphi$.

(Note that this generalised definition gives the previous one as a special case when all the formulae in the sequent are sentences.)

Definition 2.6 (Local validity, generalised)**.** A meta-inference is locally valid with respect to a class of valuations V iff it preserves sequent satisfaction for all $v \in V$ – i.e. iff for all $v \in V$ either the premiss sequents are not all satisfied by v or the conclusion sequent is satisfied by v.

We accept, with McGee, that the first-order rules are '*open-ended*, meaning that the rules are valid... not only within the language... but they will remain valid however the language may enriched by the addition of new sentences' (2000, p. 66). On this assumption, it is possible to show that the local validity of ∀-I determines the standard, objectual interpretation of ∀:

(∀) $[\![\forall v\varphi]\!]_{\mathcal{M}} = 1$ iff $[\![\varphi[o/v]]\!]_{\mathcal{M}} = 1$ for all $o \in D$.

3 Three Types of Carnap-Style Argument

The various broadly Carnapian arguments for the non-categoricity of logic that have appeared in the literature can be understood as falling under one of three general types. Arguments of two of these types can, depending on their details, be successful, but arguments of the third type turn out to be entirely illegitimate. We proceed by distinguishing among these three types of non-categoricity argument and discussing them in turn.

3.1 Type I Arguments

In Type I arguments, a general model is presented that's compatible with the validity of the rules but incompatible with the standard semantic clauses for our logical operators. Insofar as the general model that's presented genuinely has these features, a Type I argument will certainly be successful: any such general model is straightforwardly a counterexample to the claim that the validity of the rules guarantees that the operators have their standard meanings.

The trivial valuation v^*, for example, is usually deployed in service of a Type I argument: insofar as that valuation (understood as a general model) is compatible with the validity of the rules – which it is, if the rules are the usual single-conclusion rules and their validity is understood by appeal to the usual consequence relation – the rules fail to pin down the standard semantic clauses for the negation and disjunction signs, since v^* is compatible *only* with deviant interpretations of those operators.

Of course, as suggested above, there are various ways to avoid this implication by building additional structure into either our formalisation of logic or our notion of validity. This is why our local-validity-based account of the categoricity of CPL, for instance, isn't threatened by this sort of Type I argument.

3.2 Type II Arguments

Type II arguments begin not from a general model but directly from a deviant set of semantic clauses for the logical operators; here what's shown is simply that the validity of the rules fails to rule out the deviant interpretation in question. To show this requires examining *all* general models compatible with that interpretation in order to determine whether the rules are valid across those general models. It also requires ensuring that the interpretation is genuinely deviant – i.e., that at least one general model compatible with that interpretation is *incompatible* with the standard interpretation. (An alternative semantic clause for the negation sign according to which '$\neg A$' is true in a model just in case 'A' is false *and* '$A \to A$' is true, for instance, is not deviant in any genuine sense.) Insofar as a Type II argument meets both constraints, it will be successful.

The argument from Carnap's deviant semantic clause for the universal quantifier, for example, is a Type II argument: all (propositionally standard) general models compatible with that deviant interpretation – an interpretation on which '$\forall x Fx$' is true just in case every object in the domain is in the extension of 'F' *and* the object designated by 'b' is in the extension of 'G' – turn out also to be compatible with the validity of the rules (if those rules are given the sort of axiomatic formalisation that was common in Carnap's time).

Again, though, it's possible to avoid this result simply by building more structure into our formalisation of logic. On our own formalisation of the I- and E-rules for the universal quantifier, for instance, the categoricity of FOL isn't threatened by this sort of Type II argument.

3.3 Type III Arguments

Type I arguments work by holding fixed a general model and considering all possible interpretations of the logical operators that are compatible with it, and Type II arguments work by holding fixed an interpretation of the operators and considering all possible general models that are compatible with it. Both strategies are legitimate. In Type III arguments, though, elements of those strategies are illegitimately combined: restrictions are placed both on general models and on interpretations of the operators in such a way that no demonstration of non-categoricity is genuinely made available.

This is perhaps best clarified by example. So consider Briîncuş's claim that the universal introduction rule remains valid given Carnap's deviant interpretation of the universal quantifier if, in addition, "we take 'Gb' to be true" (2024, p. 349). This claim is intended to serve as an objection to various attempts to secure the categoricity of the quantifier, including our own. This, though, is a Type III argument: Briîncuş presents a deviant semantic clause for the quantifier while ruling out by fiat certain general models compatible with that interpretation. When *all* general models compatible with that interpretation are taken into account, as in a Type II argument, it's evident that the universal introduction rule, as we formalise it, isn't valid across those general models: some of the general models satisfy 'Fx' but fail to make true 'Gb', in which case '$\forall x Fx$' turns out to be false on Carnap's deviant interpretation despite the fact that the validity of the rule would guarantee its truth.

4 More on Type III Arguments: A Case Study

Del Valle-Inclán's critique of Bonnay and Westerståhl's proposal goes via a Type III argument as well. In this case, though, the mistake isn't del Valle-Inclán's; it originates with Bonnay and Westerståhl themselves, who build too much structure into their notion of a general model and so are committed to describing as deviant certain valuations that are in fact compatible with the standard interpretation of the logical operators. In particular, they take a general model not

merely to assign truth values to logically complex formulae but to come with an interpretation of the universal quantifier built in. So del Valle-Inclán proceeds by presenting a general model that otherwise is compatible with the standard interpretation of the universal quantifier and then stipulating that the quantifier has a deviant interpretation. More precisely, he constructs a general model on which, for every predicate, either everything in the domain is in the extension of that predicate or nothing is, so that a universally quantified sentence (standardly interpreted) will be true just in case its existentially quantified counterpart (standardly interpreted) is, and then he stipulates that we "(mis)interpret $\forall$ as $\exists$" (2024, p. 7), in which case the general model is of course deviant.

Notice, though, that if we take a general model (as we should) merely to assign truth values to logically complex formulae, it's clear that del Valle-Inclán is here holding fixed *both* a general model *and* an interpretation of the universal quantifier. And if we hold fixed only the general model, as in a Type I argument, it's evident that the general model is in fact *not* deviant: since '$\forall$' satisfies the standard semantic clause for the existential quantifier, it will *also* satisfy the standard clause for the universal quantifier (since, again, the general model is designed in such a way that those two interpretations will correspond to the same assignment of truth values to quantified sentences).

5 Concluding Remarks

In short, it turns out that, by attending carefully to the differences between the three types of Carnap-style argument, we can see that the strategy for solving Carnap's problem we offered in "Categoricity by Convention" (2021) remains promising despite some recent suggestions to the contrary.

Acknowledgements. Funding for this research was provided by the Austrian Science Fund (Grant No. P33708), whose support we gratefully acknowledge. We're also grateful to audiences at uAnalytiCon-2022 at Ural Federal University, the Perspectives on Categoricity workshop at the University of Vienna, the Convention in Logic and Language conference at the University of Haifa, the 2022 FilMat Conference at the University of Pavia, the 1st EuPhiLo Conference at the University of Padua, the University of Salzburg, the 11th European Congress for Analytic Philosophy at the University of Vienna, the 10 Years of *Modal Logic as Metaphysics* conference at the University of Hamburg, and the (In)determinacy in Mathematics conference at the National University of Singapore for helpful discussion of the ideas presented here.

References

Bonnay, Denis and Dag Westerståhl (2016). "Compositionality solves Carnap's Problem". In: *Erkenntnis* 81.4, pp. 721–739. DOI: https://doi.org/10.1007/s10670-015-9764-8.

Brîncuş, Constantin C. (2024a). "Categorical quantification". In: *Bulletin of Symbolic Logic*. Advance online publication. DOI: https://doi.org/10.1017/bsl.2024.3.

— (2024b). "Inferential quantification and the ω-rule". In: *Perspectives on Deduction: Contemporary Studies in the Philosophy, History and Formal Theories of Deduction*. Ed. by Antonio Piccolomini d'Aragona. Springer, pp. 345–372. DOI: https://doi.org/10.1007/978-3-031-51406-7.

Carnap, Rudolf (1943). *Formalization of Logic*. Harvard University Press.

del Valle-Inclán, Pedro (2024). "Carnap's Problem, definability and compositionality". In: *Journal of Philosophical Logic*. Advance online publication. DOI: https://doi.org/10.1007/s10992-024-09767-2.

Garson, James (2013). *What Logics Mean: From Proof Theory to Model-Theoretic Semantics*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139856461.

McGee, Vann (2000). "'Everything'". In: *Between Logic and Intuition: Essays in Honor of Charles Parsons*. Ed. by Gila Sher and Richard Tieszen. Cambridge University Press, pp. 54–78. DOI: https://doi.org/10.1017/CBO9780511570681.

Murzi, Julien and Ole Thomassen Hjortland (2009). "Inferentialism and the categoricity problem: reply to Raatikainen". In: *Analysis* 69.3, pp. 480–488. DOI: https://doi.org/10.1093/analys/anp071.

Murzi, Julien and Brett Topey (2021). "Categoricity by convention". In: *Philosophical Studies* 178.10, pp. 3391–3420. DOI: https://doi.org/10.1007/s11098-021-01606-3.

Raatikainen, Panu (2008). "On rules of inference and the meanings of logical constants". In: *Analysis* 68.4, pp. 273–280. DOI: https://doi.org/10.1093/analys/68.4.282.

Rumfitt, Ian (2000). ""Yes" and "No"". In: *Mind* 109.436, pp. 781–823. DOI: https://doi.org/10.1093/mind/109.436.781.

Smiley, Timothy (1996). "Rejection". In: *Analysis* 56.1, pp. 1–9. DOI: https://doi.org/10.1093/analys/56.1.1.

Tennant, Neil (1999). "Negation, absurdity and contrariety". In: *What Is Negation?* Ed. by Dov M. Gabbay and Heinrich Wansing. Kluwer, pp. 199–222. DOI: https://doi.org/10.1007/978-94-015-9309-0.