Restrictiveness relative to notions of interpretation

Incurvati, L.; Löwe, B.

DOI
10.1017/S1755020316000058

Publication date
2016

Document Version
Final published version

Published in
Review of Symbolic Logic

THE REVIEW OF SYMBOLIC LOGIC
Volume 9, Number 2, June 2016

RESTRICTIVENESS RELATIVE TO NOTIONS OF INTERPRETATION

LUCA INCURVATI
Universiteit van Amsterdam

BENEDIKT LÖWE
Universiteit van Amsterdam & Universität Hamburg

Abstract. Maddy gave a semi-formal account of *restrictiveness* by defining a formal notion based on a class of interpretations and explaining how to handle false positives and false negatives. Recently, Hamkins pointed out some structural issues with Maddy's definition. We look at Maddy's formal definitions from the point of view of an abstract *interpretation* relation. We consider various candidates for this interpretation relation, including one that is close to Maddy's original notion, but fixes the issues raised by Hamkins. Our work brings to light additional structural issues that we also discuss.

§1. Motivation. There is a tradition of *naturalistic philosophy of set theory* in which philosophers observe that some theories are proposed by set theorists as reasonable contenders for foundations of mathematics and others are not. The naturalistic philosopher aims to understand the reasons for these decisions of the mathematical experts, preferably on the basis of mathematical understanding of the theories involved.

The most prominent approach for this is to rule out theories because they restrict the development potential of the foundations of mathematics, i.e., violate the maxim MAXIMIZE Maddy (1988a,1988b, 1998). Maddy (1998, 1997) gives a semi-formal account of *restrictiveness* by defining a corresponding formal notion based on a class of interpretations. In (Löwe, 2001, 2003), Maddy's notion of restrictiveness was discussed and the theory ZFG (i.e., ZF + 'Every uncountable cardinal is singular') was presented as a potential witness to the restrictiveness of ZFC. More recently, Hamkins has given more examples and pointed out some structural issues with Maddy's definition (Hamkins, 2013).

In this paper, we shall look at Maddy's definitions from the point of view of an abstract *interpretation* relation. We shall then consider various candidates for this interpretation relation, including one that is close to Maddy's original notion, but fixes the issues raised by Hamkins (2013). Our work brings to light additional structural issues that we also discuss.

§2. Interpretations and restrictiveness. In the following, we let $\trianglelefteq$ be an arbitrary interpretation relation between first-order theories. There is a rich body of literature on the space of theories with interpretations between them (de Bouvère, 1965; Ahlbrandt & Ziegler, 1986; Visser, 2006). An interpretation of T in S is a syntactic object describing for every model of S a subclass of that model that forms a model of T (for definitions, cf. § 2.2 below). In this literature, it is common to have interpretation relations that are reflexive (the identity is considered an interpretation) and transitive. We deal with relations that are

Received: May 20, 2014.

doi:10.1017/S1755020316000058

neither, and, therefore, we keep the discussion at a sufficient level of generality and make explicit all assumptions about $\trianglelefteq$ in the statement of theorems.

We shall consider three properties of interpretation relations: reflexivity and transitivity, as mentioned; and then a property we call monotonicity. The relation $\trianglelefteq$ is called *monotone* if it has the property that whenever $S \trianglelefteq T$ and $R \supseteq T$, then $S \trianglelefteq R$.

Let BST be a basic set theory, appropriately chosen. For all applications in what follows (with one exception), we let $\mathsf{BST} = \mathsf{ZF}$. For the purposes of this paper, we call a theory T a *set theory* if it is a finite extension of BST, so the word "set theories" refers to axiom systems in the language $\mathcal{L}_\in$ of the form $\mathsf{BST} + \mathsf{A}$ where A is an axiom (or, equivalently, a finite set of axioms).[1] If T and S are set theories, we say that T *extends* S ($S \subseteq T$) if the set of S-theorems is a subset of the set of T-theorems. If $S = \mathsf{BST} + \mathsf{A}$ and $T = \mathsf{BST} + \mathsf{B}$, we write $S \cup T$ for $\mathsf{BST} + \mathsf{A} \wedge \mathsf{B}$. Note that there is exactly one $+$ sign in the name of a set theory, so expressions of the form $\mathsf{BST} + \mathsf{A} \wedge \mathsf{B}$ or $\mathsf{BST} + \mathsf{A} \vee \mathsf{B}$ can only be parsed one way. As a consequence, we omit parentheses to improve readability.

2.1. Restrictiveness. As observed, Maddy's enterprise uses the maxim MAXIMIZE in order to rule out theories as *restrictive*. Here, we think of a theory T as restrictive if it restricts the possible lines of development of set theory, i.e., there is some other consistent desirable axiom that is incompatible with T. The following formal sequence of definitions follows Maddy's notions from (Maddy, 1997, § 6), except that we keep the notion of interpretation completely open for now and allow theories extending BST (whereas Maddy only allows theories extending ZFC).

As usual, we write $T \lhd S$ for $T \trianglelefteq S$ and $S \not\trianglelefteq T$. Clearly, this is irreflexive independently of what properties $\trianglelefteq$ has. Moreover, if $\trianglelefteq$ is transitive, then so is $\lhd$.

DEFINITION 2.1. *Let S and T be set theories.*

1. *We say that T* $\trianglelefteq$-recaptures *S if there is a consistent extension T^* of T such that $S \trianglelefteq T^*$.*
2. *We say that S* weakly $\trianglelefteq$-maximizes over *T if $T \lhd S$ and T does not $\trianglelefteq$-recapture S. (In symbols: $T <^{\trianglelefteq}_{\text{weak}} S$.)*
3. *We say that S* strongly $\trianglelefteq$-maximizes over *T if it weakly $\trianglelefteq$-maximizes over T and $S \cup T$ is inconsistent. (In symbols: $T <^{\trianglelefteq}_{\text{strong}} S$.)*
4. *We say that T is* (weakly / strongly) $\trianglelefteq$-restrictive *if there is a consistent set theory T^* that (weakly/strongly) $\trianglelefteq$-maximizes over T.*

Maddy does not have a notion of weak maximization and the corresponding notion of weak restrictiveness. We shall discuss the relation between weak and strong maximization in § 5.

We now note two facts that will be important later in the paper, viz. that the transitivity of $\trianglelefteq$ induces the transitivity of weak and strong $\trianglelefteq$-maximization.

PROPOSITION 2.2. *Suppose that $\trianglelefteq$ is transitive. If R, S, and T are set theories and*

$$R <^{\trianglelefteq}_{\text{weak}} S \trianglelefteq T,$$

then $R <^{\trianglelefteq}_{\text{weak}} T$. Therefore, $<^{\trianglelefteq}_{\text{weak}}$ is transitive.

[1] We use the term "set theory" for convenience, and the choice of terminology is not meant to suggest that infinite extensions of a basic set theory cannot qualify as genuine set theories.

Proof. By definition, $R <^{\trianglelefteq}_{\text{weak}} S$ implies $R \triangleleft S$, so by transitivity $R \trianglelefteq T$. If we had $T \trianglelefteq R$, then by transitivity $S \trianglelefteq R$ in contradiction to $R \triangleleft S$. Therefore, $R \triangleleft T$.

If there was $R^* \supseteq R$ such that $T \trianglelefteq R^*$, then by transitivity, $S \trianglelefteq R^*$ contradicting the fact that R does not $\trianglelefteq$-recapture S. □

PROPOSITION 2.3. *If $\trianglelefteq$ is transitive and monotone, then $<^{\trianglelefteq}_{\text{strong}}$ is transitive.*

Proof. Let $R <^{\trianglelefteq}_{\text{strong}} S <^{\trianglelefteq}_{\text{strong}} T$. It follows from Proposition 2.2 that $R <^{\trianglelefteq}_{\text{weak}} T$, so we only have to show that $R \cup T$ is inconsistent. Suppose not. Since $S \trianglelefteq T$, we have by monotonicity that $S \trianglelefteq R \cup T$. But $R \cup T$ is a consistent extension of R. Thus, R $\trianglelefteq$-recaptures S, which contradicts the assumption that S strongly $\trianglelefteq$-maximizes over R. □

2.2. Types of interpretations. Thus, the abstract notion of restrictiveness, whether weak or strong, is defined in terms of an abstract interpretation relation. The idea is that a restrictive theory is one that prevents us from *interpreting* some theories.

In our specific situation, a *translation* τ consists of an $\mathcal{L}_\in$-formula δ with one free variable and an $\mathcal{L}_\in$-formula ε with two free variables. A translation recursively defines a translation operation on all $\mathcal{L}_\in$-formulas as follows:

$$\begin{aligned} (x \in y)^\tau &:= \delta(x) \wedge \delta(y) \wedge \varepsilon(x, y), \\ (\varphi \wedge \psi)^\tau &:= \varphi^\tau \wedge \psi^\tau, \\ (\neg\varphi)^\tau &:= \neg\varphi^\tau, \\ (\exists x \varphi)^\tau &:= \exists x(\delta(x) \wedge \varphi^\tau). \end{aligned}$$

A translation $\tau = \langle \delta, \varepsilon \rangle$ is an *interpretation of T in S* if for every axiom φ of T, $S \vdash \varphi^\tau$ (Tarski, Mostowski, & Robinson, 1953, pp. 20ff.). Our interpretation relations $\trianglelefteq$ are usually defined in terms of a class of translations $\mathcal{C}$ where we say that $T \trianglelefteq S$ if there is a translation $\tau \in \mathcal{C}$ such that τ is an interpretation of T in S.

We say that the *interpretability strength* of T is at least as great as that of S if there is an interpretation of S in T. If T has greater interpretability strength than S, then T has greater consistency strength than S (Visser, 2006, § 1.2.2). The converse, however, is not true (Visser, 2006, § 1.1, fn. 1). The reason is that interpretations are designed to preserve certain important features of the interpreted theory. Specific types of interpretations can then be obtained by requiring that the interpretation should preserve further aspects of the original theory. Formally, this is done by requiring that the interpreting theory should prove that the relevant interpretation describes a model which agrees in certain respects with the larger universe. In a set-theoretic context, the following notions naturally suggest themselves:[2]

DEFINITION 2.4. *Let S and T be set theories and let $\tau = \langle \delta, \varepsilon \rangle$ be an interpretation of T in S.*

1. *We say that τ is an* $\in$-interpretation of T in S *if*

$$S \vdash \forall x \forall y((x \in y)^\tau \to x \in y).$$

[2] Many other natural notions are available, but will not be used in this paper. These include the notions of interpretations in ω-models or wellfounded models.

2. *We say that τ is a* transitive interpretation of T in S *if it is an* $\in$*-interpretation and*

$$S \vdash \forall x \forall y((x \in y \wedge \delta(y)) \rightarrow \delta(x)).$$

3. *We say that τ is an* inner model interpretation of T in S *if τ is a transitive interpretation of T in S and*

$$S \vdash \forall \alpha \delta(\alpha).$$[3]

§3. Fair interpretations.

3.1. A reference list of set theories. In this section, we present the set theories that we shall be concerned with. Most of these either come from Maddy's discussion of the original definitions, motivating the individual constituents (Maddy, 1997, pp. 218–231), or from Hamkins's discussion of "some new problems with Maddy's proposal" (Hamkins, 2013, § 2).

The basic additional axioms are large cardinal axioms; these are of the form $\exists\kappa\Phi(\kappa)$ where Φ is some large cardinal property. In the following, we shall consider:[4]

$$\begin{aligned} \mathsf{IC} \quad & \exists\kappa(\kappa \text{ is strongly inaccessible}),\\ \mathsf{Mahlo} \quad & \exists\kappa(\kappa \text{ is Mahlo}),\\ \mathsf{MC} \quad & \exists\kappa(\kappa \text{ is measurable}). \end{aligned}$$

Since inaccessible cardinals are very important for the set-up of the definitions, we introduce the abbreviation $\mathrm{Inacc}(\kappa)$ for "κ is strongly inaccessible".

For each large cardinal axiom $\mathsf{A} = \exists\kappa\Phi(\kappa)$, we can consider multiple cardinal versions such as $2\mathsf{A} = \exists\kappa < \lambda(\Phi(\kappa) \wedge \Phi(\lambda))$ and the unbounded version $\infty\mathsf{A} = \forall\alpha\exists\kappa > \alpha\Phi(\kappa)$. The latter produces axioms such as "there is a proper class of inaccessible cardinals", $\infty\mathsf{IC}$.

The axiom MC has greater consistency strength than Mahlo, which in turn has greater consistency strength than IC. Even more is true: MC has greater consistency strength than $\infty\mathsf{Mahlo}$, and Mahlo has greater consistency strength than $\infty\mathsf{IC}$. Furthermore, for each of our large cardinal axioms A, $\infty\mathsf{A}$ has strictly greater consistency strength than A. The nontrivial parts of these claims are all proved by truncating the universe at an appropriate inaccessible. As an example, consider MC and $\infty\mathsf{IC}$: by the ultrapower construction, if κ is measurable, there are unboundedly many Mahlo cardinals below κ (Kanamori, 2003, Proposition 5.15), but each Mahlo cardinal has unboundedly many inaccessibles below it. Let λ be the least Mahlo cardinal and consider $\mathbf{V}_\lambda$, the set of sets of Mirimanoff rank $< \lambda$. The structure $\langle \mathbf{V}_\lambda, \in\rangle$ has unboundedly many inaccessible cardinals and is a model of ZFC since λ was inaccessible (Kanamori, 2003, Proposition 1.2), so it is a set model of $\infty\mathsf{IC}$. This construction proves that $\mathsf{ZFC} + \mathsf{MC} \vdash \mathrm{Cons}(\mathsf{ZFC} + \infty\mathsf{IC})$. The operation of truncation used in this argument is crucial in § 3.2.

Now that we have a list of set theories, we can of course consider Boolean combinations of the axioms defining these, such as conjunctions, disjunctions, and negations, forming new set theories. In the case of negations of large cardinal axioms and their unbounded versions, we get axioms such as "there is no measurable cardinal", $\neg\mathsf{MC}$, or "there is at most a set of Mahlo cardinals", $\neg\infty\mathsf{Mahlo}$.

[3] Here, and throughout this paper, the variable α always ranges over ordinals, so $\forall\alpha\varphi$ is an abbreviation for $\forall\alpha(\alpha \in \mathrm{Ord} \rightarrow \varphi)$ and $\exists\alpha\varphi$ for $\exists\alpha(\alpha \in \mathrm{Ord} \wedge \varphi)$.

[4] For definitions, cf. (Kanamori, 2003, § 1&2). All of our large cardinal axioms will be used in the ZFC-context, so we are ignoring the subtleties that are required in the absence of AC.

Of course, a very important antilarge cardinal axiom is $\mathbf{V}{=}\mathbf{L}$, which implies $\neg\mathsf{MC}$ (Kanamori, 2003, § 3 & Corollary 5.5) but is consistent with other large cardinal axioms such as IC or $\infty\mathsf{IC}$. These axioms can be relativized to $\mathbf{L}$, and we write $\mathsf{A}^{\mathbf{L}}$ for these relativized axioms, e.g., $\mathsf{IC}^{\mathbf{L}}$ for "there is an inaccessible cardinal in $\mathbf{L}$".

Relativized large cardinal axioms are consistent with the negations of their full versions, e.g., $\mathsf{IC}^{\mathbf{L}} \wedge \neg\mathsf{IC}$ is true in the universe obtained by collapsing (by forcing) the only inaccessible of a ground model we are starting from. Constructions of this sort can be used to show that the negated large cardinal axioms we consider, which might seem to be restrictive in the pretheoretical sense, are not formally restrictive (cf. § 3.2).

Combinations like $\mathsf{IC}^{\mathbf{L}} \wedge \neg\mathsf{IC}$ clearly imply $\mathbf{V}{\neq}\mathbf{L}$, and so do statements such as $(\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$, which says "the true ω_1 is an inaccessible cardinal in $\mathbf{L}$".

3.2. Definitions. In § 3.1, we used the operation of truncation to deal with situations like $\mathsf{ZFC} + \mathsf{Mahlo}$ and $\mathsf{ZFC} + \infty\mathsf{IC}$. The consistency strength of the former is strictly greater than that of the latter, but the former does not imply the latter. Furthermore, it is consistent to have a model with a Mahlo cardinal without any inner models with unboundedly many inaccessible cardinals. Only by truncating such model at an inaccessible limit of inaccessible cardinals (e.g., at a Mahlo cardinal[5]) do we obtain a model of $\infty\mathsf{IC}$. Truncated inner model interpretations are needed to handle the operation of truncation.

DEFINITION 3.1. *Let S and T be set theories. We say that $\tau = \langle\delta, \varepsilon\rangle$ is a* truncated inner model interpretation of T in S *if τ is a transitive interpretation of T in S and* $S \vdash \exists\kappa(\mathrm{Inacc}(\kappa) \wedge \forall\alpha(\alpha < \kappa \leftrightarrow \delta(\alpha)))$.[6]

We can now define Maddy's notion of a fair interpretation.

DEFINITION 3.2. *Let S and T be set theories. We say that τ is a* fair interpretation of T in S *if τ is either an inner model interpretation of T in S or a truncated inner model interpretation of T in S. We write $T \trianglelefteq_{\text{fair}} S$ if there is such a τ.*

Thus, the definition of a fair interpretation involves a meta-disjunction: that a theory S fairly interprets a theory T implies that S either proves that the domain of the interpretation contains all ordinals or S proves that the domain of the interpretation contains all ordinals up to an inaccessible. This fact will be used in Hamkins's proof of Proposition 3.5, and will be discussed further in § 4.2.

Given an interpretation $\tau = \langle\delta, \varepsilon\rangle$ between set theories, we define $\mathrm{BR}(x, y)$ to be the formula $y \subseteq x^2$. For v, w such that $\delta(v)$ and $\delta(w)$, we write $\mathrm{op}^\tau(v, w)$ for the unique z such that $\delta(z)$ and τ interprets z as the ordered pair of v and w. If $\mathrm{BR}(x, y)^\tau$, we define two sets:

$$E_x^\tau := \{z \colon \varepsilon(z, x)\} \text{ and}$$
$$R_y^\tau := \{(v, w) \colon \varepsilon(\mathrm{op}^\tau(v, w), y)\}.$$

Then the formula $\exists x \exists X \subseteq x^2 \forall y \forall z(\mathrm{BR}(y, z)^\tau \rightarrow \langle x, X\rangle \ncong \langle E_y^\tau, R_z^\tau\rangle)$ says that there is some structure which is not isomorphic to any structure in δ. We abbreviate this formula as $\mathrm{NewIso}(\tau)$ and use it to characterize another type of interpretation.

[5] Note that one could also truncate at an inaccessible limit of inaccessible cardinals that is not Mahlo. (There must be many such cardinals below the Mahlo cardinal.)

[6] In (Maddy, 1998, p. 145) and (Maddy, 1997, p. 221), Maddy formulates truncation at inaccessible levels by requiring that $S \vdash \exists\kappa(\mathrm{Inacc}(\kappa) \wedge \forall\alpha(\alpha < \kappa \rightarrow \delta(\alpha)))$. But if truncation has to occur at an inaccessible cardinal, rather than at an inaccessible *or somewhere above*, we need to replace the conditional with a biconditional (Hamkins, 2013, § 2).

DEFINITION 3.3. *Let S and T be set theories and let $\tau = \langle \delta, \varepsilon \rangle$ be an interpretation of T in S. We say that τ is a* maximizing interpretation *of T in S if $S \vdash \mathrm{NewIso}(\tau)$.*[7]

The notion of interpretation Maddy settles on is that of a fair interpretation which is also maximizing.

DEFINITION 3.4. *Let S and T be set theories. We say that τ is a* Maddy interpretation of *T* in *S if τ is a fair and maximizing interpretation of T in S. We write $T \trianglelefteq_{\mathrm{Maddy}} S$ if there is such a τ.*

Whilst the relation $\trianglelefteq_{\mathrm{fair}}$ is reflexive, the relation $\trianglelefteq_{\mathrm{Maddy}}$ no longer is, since the requirement that interpretations should be maximizing rules out the trivial interpretation $\langle x = x, \in \rangle$. We shall say more about this requirement in §5.

Maddy's notion of restrictiveness, as used in (Maddy, 1997), is now just strong $\trianglelefteq_{\mathrm{Maddy}}$-restrictiveness. Her main result is to show that $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L}$ is strongly $\trianglelefteq_{\mathrm{Maddy}}$-restrictive.[8] Note, on the other hand, that the negated large cardinal axioms considered in this paper are not strongly $\trianglelefteq_{\mathrm{Maddy}}$-restrictive. For instance, although there is a Maddy interpretation of $\mathsf{ZFC} + \neg\mathsf{IC}$ in $\mathsf{ZFC} + \mathsf{IC}$, the former $\trianglelefteq_{\mathrm{Maddy}}$-recaptures the latter, since there is a Maddy interpretation of $\mathsf{ZFC} + \mathsf{IC}$ in $\mathsf{ZFC} + \neg\mathsf{IC} \wedge 2\mathsf{IC}^{\mathbf{L}}$.

3.3. Hamkins's counterexamples. Hamkins (2013) has noticed two structural problems with Maddy's definitions. We here reproduce the proofs using our notation, since this will facilitate the discussion in § 4.

The first problem is that disjunction is not an upper bound with respect to $\trianglelefteq_{\mathrm{fair}}$.

PROPOSITION 3.5 (Hamkins). *There is a set theory T and axioms A and B such that $T \trianglelefteq_{\mathrm{fair}} \mathsf{ZFC} + \mathsf{A}$, $T \trianglelefteq_{\mathrm{fair}} \mathsf{ZFC} + \mathsf{B}$, but $T \not\trianglelefteq_{\mathrm{fair}} \mathsf{ZFC} + \mathsf{A} \vee \mathsf{B}$.*

Proof. Let δ_{Mahlo} be the formula expressing "$x \in \mathbf{V}_\kappa$ where κ is the least Mahlo cardinal". It is clear that $\langle \delta_{\mathrm{Mahlo}}, \in \rangle$ is a fair interpretation of $\mathsf{ZFC}{+}\infty\mathsf{IC}$ in $\mathsf{ZFC}{+}\mathsf{Mahlo}$. And it is also clear that $\langle x \in \mathbf{L}, \in \rangle$ is a fair interpretation of $\mathsf{ZFC}{+}\infty\mathsf{IC}$ in $\mathsf{ZFC}{+}(\infty\mathsf{IC})^{\mathbf{L}}$. However, $\mathsf{ZFC} + \infty\mathsf{IC} \not\trianglelefteq_{\mathrm{fair}} \mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$.

To see the latter, note first that $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}} \nvdash \mathsf{IC}$, since from any model of $\mathsf{ZFC} + \infty\mathsf{IC}$ we can obtain by forcing a model of $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$ without inaccessibles (Hamkins, 2013, § 2). Thus, if $\tau = \langle \delta, \in \rangle$ is a fair interpretation of $\mathsf{ZFC} + \infty\mathsf{IC}$ in $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$, it must be because τ is an inner model interpretation of $\mathsf{ZFC} + \infty\mathsf{IC}$ in $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$, and so we must have that $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}} \vdash \forall\alpha\,\delta(\alpha)$.

---

[7] In her original presentation, Maddy (1997, p. 221) uses the formula $\exists x \exists X \subseteq x^2 \forall y \forall Y \subseteq y^2 (\delta(y) \wedge \delta(Y) \rightarrow \langle x, X \rangle \ncong \langle y, Y \rangle)$, but this expresses the new isomorphism type requirement only if one restricts attention to transitive interpretations. Note that if τ is a transitive interpretation (and given that S is a set theory and hence extends ZF), requiring that $S \vdash \mathrm{NewIso}(\tau)$ is the same as requiring that $S \vdash \exists x \neg\delta(x)$. For if there is a set A such that $\neg\delta(A)$, then the isomorphism type of the transitive closure of $\{A\}$ cannot be realized in δ, as this would force A itself into that structure by the Mostowski collapse. This was already pointed out by Maddy (1997, pp. 221–222, fn. 17) for her formulation of the new isomorphism type requirement.

[8] Maddy notices that her formal criterion of restrictiveness admits of false negatives and false positives. For this reason, she weakens it to a semi-formal criterion by demanding that the inner model described by the fair interpretation should be 'optimal' and that the witness to a theory's restrictiveness should not be a 'dud' theory (Maddy, 1997, pp. 225–231). Since we are here interested in the formal properties of the fair interpretation relation, we may ignore this feature of her proposal.

But now note that $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$ cannot prove $(\infty\mathsf{IC})^{\mathbf{L}}$ either, since by truncation we can get a model of $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$ with a Mahlo cardinal in $\mathbf{L}$ but no inaccessibles above. So if $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}} \vdash \forall\alpha\delta(\alpha)$, then $\mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}} \nvdash (\infty\mathsf{IC})^{\tau}$. Hence, $\mathsf{ZFC} + \infty\mathsf{IC} \ntrianglelefteq_{\text{fair}} \mathsf{ZFC} + \mathsf{Mahlo} \vee (\infty\mathsf{IC})^{\mathbf{L}}$. □

The second structural problem is that the fair interpretation relation is not transitive.

PROPOSITION 3.6 (Hamkins). *There are set theories S, T, and R such that $S \trianglelefteq_{\text{fair}} T$, $T \trianglelefteq_{\text{fair}} R$, but $S \ntrianglelefteq_{\text{fair}} R$.*

Proof. Let δ_{IC} be the formula expressing "$x \in \mathbf{V}_\kappa$ where κ is the least inaccessible cardinal". It is easy to see that $\langle \delta_{\mathsf{IC}}, \in\rangle$ is a fair interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \mathsf{IC}$. And it is also easy to see that $\langle x \in \mathbf{L}, \in\rangle$ is a fair interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \mathsf{IC}$ in $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$. However, $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC} \ntrianglelefteq_{\text{fair}} \mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$.

To see this, note first that $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}} \nvdash \mathsf{IC}$. So if $\tau = \langle \delta, \in\rangle$ is a fair interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$, it must be because τ is an inner model interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$, and so we must have that $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}} \vdash \forall\alpha\delta(\alpha)$. But then in order to have $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}} \vdash (\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC})^{\tau}$, δ would have to be '$x \in \mathbf{L}$'. Clearly, however, $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}} \vdash \mathsf{IC}^{\mathbf{L}}$, and so $\langle x \in \mathbf{L}, \in\rangle$ is not a fair interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$ after all. Thus, $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC} \ntrianglelefteq_{\text{fair}} \mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$. □

Hamkins (2013, § 4) proposes a very simple solution: replace the fair interpretation relation with the transitive interpretation relation. This relation is transitive, and disjunction is an upper bound with respect to it. The former fact is obvious; to establish the latter fact, we prove the more general Theorem 3.8 below, which we shall also make use of in § 4.2.

But although allowing all transitive interpretations avoids the problems of the Maddy interpretation relation $\trianglelefteq_{\text{Maddy}}$, it also has the consequence that $\mathbf{V}{=}\mathbf{L}$ is no longer restrictive: $\mathbf{L}$ and $\mathbf{V}$ have transitive models of exactly the same set theories (Hamkins, 2013, § 3).[9] In Hamkins's view, this reflects the fact that "the believer in $\mathbf{V}{=}\mathbf{L}$ seems fully able to converse meaningfully with any large cardinal set theorist" (Hamkins, 2013, § 4).

At the beginning of her discussion of fair interpretations, Maddy had already considered transitive countable models inside $\mathbf{L}$ of theories contradicting $\mathbf{V}{=}\mathbf{L}$ as a possible interpretation, but discarded them with the words: "[this] is a paltry interpretation …; to begin with, it's countable!" (Maddy, 1997, p. 219).

In the next section, we shall provide a solution to Hamkins's problems that is not as radical as using all transitive interpretations and is closer to the spirit of Maddy's proposal.

For Theorem 3.8, we need the following definitions.

DEFINITION 3.7. *In the following, let φ_0 and φ_1 be sentences of $\mathcal{L}_\in$, let $\tau_0 = \langle \delta_0, \varepsilon_0\rangle$ and $\tau_1 = \langle \delta_1, \varepsilon_1\rangle$ be translations, and let $\mathcal{C}$ be a class of translations.*

[9] This can be seen simply by observing that the assertion that a theory T has a transitive model has complexity Σ^1_2, and is therefore absolute between $\mathbf{L}$ and $\mathbf{V}$ by Shoenfield's absoluteness theorem (on the assumption that $T \in \mathbf{L}$, which is the case for the theories under consideration in this paper).

1. *We say that* φ_0 *and* φ_1 *are* mutually exclusive *if* $\mathsf{BST} \vdash \varphi_0 \to \neg\varphi_1$.
2. *If* φ_0 *and* φ_1 *are mutually exclusive, then* $\langle(\varphi_0 \to \delta_0) \wedge (\varphi_1 \to \delta_1), (\varphi_0 \to \varepsilon_0) \wedge (\varphi_1 \to \varepsilon_1)\rangle$ *is the* disjunction of τ_0 and τ_1 induced by φ_0 and φ_1. *(In symbols:* $\tau_0 \vee_{\varphi_0,\varphi_1} \tau_1$.*)*
3. *We say that* $\mathcal{C}$ *is* closed under disjunction *if for mutually exclusive formulas* φ_0 *and* φ_1*, if* τ_0 *and* τ_1 *are in* $\mathcal{C}$ *then so is* $\tau_0 \vee_{\varphi_0,\varphi_1} \tau_1$.

Given these definitions, we can now prove the following theorem.

THEOREM 3.8. *Let* $\mathcal{C}$ *be a class of translations closed under disjunctions and define* $T \trianglelefteq S$ *by "there is a translation of* T *in* S *in* $\mathcal{C}$*". If* $T \trianglelefteq \mathsf{BST} + \mathsf{A}$ *and* $T \trianglelefteq \mathsf{BST} + \mathsf{B}$*, then* $T \trianglelefteq \mathsf{BST} + \mathsf{A} \vee \mathsf{B}$.

Proof. Suppose $\tau \in \mathcal{C}$ is a witness to $T \trianglelefteq \mathsf{BST} + \mathsf{A}$ and $\upsilon \in \mathcal{C}$ is a witness to $T \trianglelefteq \mathsf{BST} + \mathsf{B}$. Clearly, A and $\mathsf{B} \wedge \neg\mathsf{A}$ are mutually exclusive. By our closure assumption on $\mathcal{C}$, we have that $\tau \vee_{\mathsf{A},\mathsf{B}\wedge\neg\mathsf{A}} \upsilon$ is also in $\mathcal{C}$. This translation witnesses $T \trianglelefteq \mathsf{BST} + \mathsf{A} \vee \mathsf{B}$. □

§4. Possibly truncated inner model interpretations.

4.1. Definitions. Hamkins's counterexamples highlight two problematic features of Maddy's definitions. First, the definition of a fair interpretation requires the interpreting theory either to prove that the interpretation's domain contains all ordinals or to prove that it contains all ordinals up to an inaccessible. In the proof of Proposition 3.5, this fact is exploited by constructing a case where there is an inner model interpretation of T in $\mathsf{ZFC} + \mathsf{A}$ and a truncated inner model interpretation of T in $\mathsf{ZFC} + \mathsf{B}$ but there is no transitive interpretation $\tau = \langle\delta, \in\rangle$ of T in $\mathsf{ZFC} + \mathsf{A} \vee \mathsf{B}$ such that $\mathsf{ZFC} + \mathsf{A} \vee \mathsf{B} \vdash \forall\alpha\delta(\alpha)$ or $\mathsf{ZFC} + \mathsf{A} \vee \mathsf{B} \vdash \exists\kappa(\mathrm{Inacc}(\kappa) \wedge \forall\alpha(\alpha < \kappa \leftrightarrow \delta(\alpha)))$. The definitions below deal with the problem by dispensing with the meta-disjunction involved in the definition of a fair interpretation and introducing the notion of a possibly truncated interpretation, which includes a disjunctive element in what the interpreting theory is required to prove.

Second, the definition of a truncated inner model interpretation allows us to truncate at a cardinal κ only if κ is inaccessible in the universe of the interpreting theory. In the example used by Hamkins to show that $\trianglelefteq_{\mathrm{fair}}$ is not transitive, however, we would like to interpret $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + (\mathrm{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$ by going to $\mathbf{L}$ and then truncating at a cardinal which is inaccessible there and not in the larger universe. The definitions below are designed to allow truncation at a cardinal which is only inaccessible in the inner model defined by the interpretation.[10]

We now proceed to give our definitions. First, we need some abbreviations. Let $\tau = \langle\delta, \varepsilon\rangle$ be a translation and rk be the Mirimanoff rank function. Then we write $\delta_\alpha(y)$ for $\delta(y) \wedge \mathrm{rk}(y) < \alpha$ and τ_α for $\langle\delta_\alpha, \varepsilon\rangle$.

[10] One might also want to allow for truncation at a cardinal which is not inaccessible but wordly in the inner model defined by the interpretation (where a cardinal κ is *worldly* if $\mathbf{V}_\kappa \models \mathsf{ZFC}$). For the purposes of this paper, we follow Maddy in using inaccessible cardinals to provide the points at which truncation can occur, but our definitions can be easily adapted to the case of worldy cardinals. We hope to return to the issue in future work when discussing Maddy's choice to focus on inner models and truncations thereof (cf.§ 5).

Let S and T be set theories, i.e., $T = \mathsf{BST} + \mathsf{A}$. Then a translation $\tau = \langle \delta, \varepsilon \rangle$ is a *possibly truncated interpretation of T in S* if

1. $S \vdash \varphi^\tau$ for every axiom φ of BST;
2. $S \vdash \mathsf{A}^\tau \vee (\exists\kappa\ \mathrm{Inacc}(\kappa)^\tau \wedge \mathsf{A}^{\tau_\kappa})$.

It should now be clear why we have taken set theories to be *finite* extensions of BST. If we had allowed infinite extensions of BST, then we would have had to require T to prove $\varphi^\tau \vee (\exists\kappa\ \mathrm{Inacc}(\kappa)^\tau \wedge \varphi^{\tau_\kappa})$ for every axiom φ of S. But then it could have been the case that some axioms of S held only in τ and others only in τ_κ.

Also, note that the notion of a possibly truncated interpretation is not a special case of the notion of an interpretation as defined in § 2.2. Nonetheless, the results in this paper that concern an abstract interpretation relation $\trianglelefteq$ also apply to possibly truncated interpretations.

DEFINITION 4.1. *Let S and T be set theories and let $\tau = \langle \delta, \varepsilon \rangle$ be a possibly truncated interpretation of T in S.*

1. *We say that τ is a* possibly truncated $\in$-interpretation of T in S *if* $S \vdash \forall x \forall y((x \in y)^\tau \to x \in y)$.
2. *We say that τ is a* possibly truncated transitive interpretation of T in S *if τ is a possibly truncated $\in$-interpretation of T in S and $S \vdash \forall x \forall y((x \in y \wedge \delta(y)) \to \delta(x))$.*
3. *We say that τ is a* possibly truncated inner model interpretation of T in S *if τ is a possibly truncated transitive interpretation of T in S and $S \vdash \forall\alpha\delta(\alpha)$. We write $T \trianglelefteq_{\mathrm{ptim}} S$ if there is such a τ.*

DEFINITION 4.2. *Let S and T be set theories, i.e., $T = \mathsf{BST} + \mathsf{A}$. Then a translation $\tau = \langle \delta, \varepsilon \rangle$ is a* maximizing possibly truncated interpretation of T in S *if*

1. *$S \vdash \varphi^\tau$ for every axiom φ of BST;*
2. $S \vdash (\mathsf{A}^\tau \wedge \mathrm{NewIso}(\tau) \vee (\exists\kappa\ \mathrm{Inacc}(\kappa)^\tau \wedge (\mathsf{A}^{\tau_\kappa} \wedge \mathrm{NewIso}(\tau_\kappa)))$.

DEFINITION 4.3. *Let S and T be set theories and let τ be a possibly truncated interpretation of T in S. We say that τ is a* Maddy* interpretation *if τ is a possibly truncated inner model interpretation and is a maximizing possibly truncated interpretation. We write $T \trianglelefteq_{\mathrm{Maddy}^\star} S$ if there is such a τ.*

4.2. The counterexamples defused. Propositions 2.2, 2.3, and Theorem 3.8 show that if $\trianglelefteq_{\mathrm{Maddy}^\star}$ is transitive and monotone and the class of Maddy* interpretations is closed under disjunctions, then we have resolved all of Hamkins's problems.

PROPOSITION 4.4. *Both $\trianglelefteq_{\mathrm{ptim}}$ and $\trianglelefteq_{\mathrm{Maddy}^\star}$ are monotone.*

Proof. The relations $T \trianglelefteq_{\mathrm{ptim}} S$ and $T \trianglelefteq_{\mathrm{Maddy}^\star} S$ are just defined in terms of a list of statements of the form $S \vdash \Phi$, so replacing S with $R \supseteq S$ preserves the property. □

PROPOSITION 4.5. *If $T \trianglelefteq_{\mathrm{ptim}} \mathsf{BST}+\mathsf{A}$ and $T \trianglelefteq_{\mathrm{ptim}} \mathsf{BST}+\mathsf{B}$, then $T \trianglelefteq_{\mathrm{ptim}} \mathsf{BST}+\mathsf{A}\vee\mathsf{B}$; similarly, if $T \trianglelefteq_{\mathrm{Maddy}^\star} \mathsf{BST}+\mathsf{A}$ and $T \trianglelefteq_{\mathrm{Maddy}^\star} \mathsf{BST}+\mathsf{B}$, then $T \trianglelefteq_{\mathrm{Maddy}^\star} \mathsf{BST}+\mathsf{A}\vee\mathsf{B}$.*

Proof. It is easy to see that both the class of possibly truncated inner model interpretations and the class of Maddy* interpretations are closed under disjunction. The result then follows immediately from Theorem 3.8. □

THEOREM 4.6. *If $S \trianglelefteq_{\text{ptim}} T$ and $T \trianglelefteq_{\text{ptim}} R$, then $S \trianglelefteq_{\text{ptim}} R$.*

Proof. Let $S = \mathsf{BST} + \mathsf{A}$ and $T = \mathsf{BST} + \mathsf{B}$ and suppose that $S \trianglelefteq_{\text{ptim}} T$ and $T \trianglelefteq_{\text{ptim}} R$. By definition, there is a $\tau = \langle \delta, \in \rangle$ such that τ is a possibly truncated inner model interpretation of S in T and a $\upsilon = \langle \xi, \in \rangle$ such that υ is a possibly truncated inner model interpretation of T in R. Let $\vartheta = \delta \circ \xi$. We claim that $\sigma = \langle \vartheta, \in \rangle$ is a possibly truncated inner model interpretation of S in R.

We begin by observing that $R \vdash \forall x \forall y((x \in y \wedge \vartheta(y)) \rightarrow \vartheta(x))$ and $R \vdash \forall \alpha \vartheta(\alpha)$. Moreover, for all φ in BST, $R \vdash \varphi^\sigma$ since by assumption $R \vdash \varphi^\upsilon$ for all φ in BST. It remains to be shown that $R \vdash A^\sigma \vee (\exists \nu \text{ Inacc}(\nu)^\sigma \wedge \mathsf{A}^{\sigma_\nu})$.

Since $R \vdash \mathsf{B}^\upsilon \vee (\exists \lambda \text{ Inacc}(\lambda)^\upsilon \wedge \mathsf{B}^{\upsilon_\lambda})$, then $R \vdash (\mathsf{A}^\tau \vee (\exists \kappa \text{ Inacc}(\kappa)^\tau \wedge \mathsf{A}^{\tau_\kappa}))^\upsilon \vee (\exists \lambda \text{ Inacc}(\lambda)^\upsilon \wedge (\mathsf{A}^\tau \vee (\exists \mu \text{ Inacc}(\mu)^\tau \wedge \mathsf{A}^{\tau_\mu}))^{\upsilon_\lambda})$.

We now note that for any α, ${\tau_\alpha}^\upsilon = \sigma_\alpha$ and $\tau^{\upsilon_\alpha} = \sigma_\alpha$. So we have that $R \vdash (\mathsf{A}^\sigma \vee (\exists \kappa \text{ Inacc}(\kappa)^\sigma \wedge \mathsf{A}^{\sigma_\kappa})) \vee (\exists \lambda \text{ Inacc}(\lambda)^\upsilon \wedge (\mathsf{A}^{\sigma_\lambda} \vee (\exists \mu \text{ Inacc}(\mu)^{\sigma_\lambda} \wedge \mathsf{A}^{\sigma_{\min(\lambda,\mu)}})))$.

By elementary reasoning, it follows that $R \vdash \mathsf{A}^\sigma \vee ((\exists \kappa \text{ Inacc}(\kappa)^\sigma \wedge \mathsf{A}^{\sigma_\kappa}) \vee (\exists \lambda \text{ Inacc}(\lambda)^\upsilon \wedge \mathsf{A}^{\sigma_\lambda}) \vee (\exists \lambda \text{ Inacc}(\lambda)^\upsilon \wedge \exists \mu \text{ Inacc}(\mu)^{\sigma_\lambda} \wedge \mathsf{A}^{\sigma_\mu}))$.

But for α and β such that $\alpha < \alpha^+ < \beta$, $V_\beta \models \text{Inacc}(\alpha)$ if and only if $V \models \text{Inacc}(\alpha)$. It follows that $R \vdash \mathsf{A}^\sigma \vee (\exists \nu \text{ Inacc}(\nu)^\sigma \wedge \mathsf{A}^{\sigma_\nu})$. □

It might be instructive to see how the definition of a possibly truncated inner model interpretation deals with the example Hamkins used to prove that $\trianglelefteq_{\text{fair}}$ is not transitive (cf. the proof of Proposition 3.6). We need a τ such that τ is a possibly truncated inner model interpretation of $\mathsf{ZFC} + \mathbf{V}{=}\mathbf{L} \wedge \neg\mathsf{IC}$ in $\mathsf{ZFC} + (\text{Inacc}(\omega_1^{\mathbf{V}}))^{\mathbf{L}}$. The required τ is $\langle x \in \mathbf{L}, \in \rangle$.

PROPOSITION 4.7. *If $S \trianglelefteq_{\text{Maddy}^\star} T$ and $T \trianglelefteq_{\text{Maddy}^\star} R$, then $S \trianglelefteq_{\text{Maddy}^\star} R$.*

Proof. Let $S = \mathsf{BST} + \mathsf{A}$ and $T = \mathsf{BST} + \mathsf{B}$ and suppose that $S \trianglelefteq_{\text{Maddy}^\star} T$ and $T \trianglelefteq_{\text{Maddy}^\star} R$. By definition, there is a $\tau = \langle \delta, \in \rangle$ such that τ is a Maddy* interpretation of S in T and a $\upsilon = \langle \xi, \in \rangle$ such that υ is a Maddy* interpretation of T in R. Let $\vartheta = \delta \circ \xi$. We claim that $\sigma = \langle \vartheta, \in \rangle$ is a Maddy* interpretation of S in R.

The proof of Theorem 4.6 tells us that σ witnesses that $S \trianglelefteq_{\text{ptim}} R$, so we only need to check that

$$R \vdash (\mathsf{A}^\sigma \wedge \text{NewIso}(\sigma)) \vee (\exists \nu \text{ Inacc}(\nu)^\sigma \wedge (\mathsf{A}^{\sigma_\nu} \wedge \text{NewIso}(\sigma_\nu))).$$

The same argument used in the proof of Theorem 4.6 establishes that $R \vdash (\mathsf{A}^\sigma \wedge \text{NewIso}(\upsilon)) \vee (\exists \nu \text{ Inacc}(\nu)^\sigma \wedge (\mathsf{A}^{\sigma_\nu} \wedge \text{NewIso}(\upsilon_\nu)))$. If $R \vdash \text{NewIso}(\upsilon)$, then $R \vdash \text{NewIso}(\sigma)$. If $R \vdash \exists \nu (\text{Inacc}(\nu)^\sigma \wedge \text{NewIso}(\upsilon_\nu))$, then $R \vdash \exists \nu (\text{Inacc}(\nu)^\sigma \wedge \text{NewIso}(\sigma_\nu))$. □

§5. Discussion of some design choices. In Section 4, we provided a notion of Maddy* interpretation $\trianglelefteq_{\text{Maddy}^\star}$ and its corresponding notion of strong $\trianglelefteq_{\text{Maddy}^\star}$-restrictiveness that are as similar to Maddy's original notion as possible while dealing with the formal issues observed by Hamkins. The main technical components of Maddy's definitions are

(DC1) the focus on interpretations in inner models or truncations of inner models,
(DC2) the notion of maximizing interpretations, resulting in an irreflexive interpretation relation, and
(DC3) the fact that the notion of restrictiveness is defined in terms of strongly maximizing theories (rather than weakly maximizing theories).

Obviously, among the three, (**DC1**) is the most fundamental. In (Löwe, 2001, pp. 352–353) and (Löwe, 2003, pp. 329–331), the focus on inner models was identified as the culprit when the theory ZFG $\trianglelefteq_{\text{Maddy}}$-strongly maximizes over ZFC. In (Löwe, 2003, p. 330), the second author proposed a "compromise option [using] a system of ranked naturalness"; this discussion deserves an in-depth analysis that will be the topic of future work and we shall not go into (**DC1**) any further here (but see the remark about AFA below).

Concerning (**DC2**), the main effect that the requirement of interpretations satisfying $\text{NewIso}(\tau)$ is that the interpretation relation becomes irreflexive. We already mentioned that this is a rather unusual feature since in the general theory of interpretations, interpretation relations tend to be reflexive and transitive. The following argument for this design choice can be extrapolated from Maddy's discussion (Maddy, 1997, pp. 216–222):

Consider the theory AFA consisting of ZFC with the Axiom of Foundation replaced by Aczel's Anti-Foundation Axiom (Aczel, 1988, Chapter 1). Clearly, ZFC or any consistent extension thereof cannot fairly interpret AFA, since any fair interpretation is also an $\in$-interpretation. But the von Neumann hierarchy $\mathbf{WF}$ is a fair interpretation of ZFC in AFA (Maddy, 1997, p. 221). So $\mathsf{ZFC} \vartriangleleft_{\text{fair}} \mathsf{AFA}$, and ZFC is $\trianglelefteq_{\text{fair}}$-restictive. However, as Maddy points out (Maddy, 1997, p. 218, fn. 9), $\mathsf{AFA} \nvdash \text{NewIso}(\langle \mathbf{WF}, \in \rangle)$ and indeed $\mathsf{AFA} \vdash \neg\text{NewIso}(\langle \mathbf{WF}, \in \rangle)$, and hence the von Neumann hierarchy is not a Maddy interpretation of ZFC in AFA. In fact, no such interpretation exists (Maddy, 1997, p. 222, fn. 18).

Note that the statement "$\mathsf{ZFC} \vartriangleleft_{\text{fair}} \mathsf{AFA}$" is the one instance mentioned in §2 where our definitions are applied to a set theory which is not a finite extension of ZF. However, this makes a renewed discussion of (**DC1**) necessary as the focus on $\in$-interpretations (which plays a crucial role in the argument) heavily depends on our choice of ZF as base theory and the moment one moves to weaker base theories, this choice has to be re-assessed. So, if the base theory (as in this case) is $\mathsf{ZF} - \mathsf{Foundation}$, then we might want to deal with interpretations that are not $\in$-interpretations.

Concerning (**DC3**), we should like to present the following abstract result.

PROPOSITION 5.1. *Let S and T be set theories. Suppose that $\trianglelefteq$ is reflexive and monotone.*[11] *Then $S <^{\trianglelefteq}_{\text{weak}} T$ if and only if $S <^{\trianglelefteq}_{\text{strong}} T$.*

Proof. The right-to-left direction is immediate from the definitions. For the left-to-right direction, suppose $S <^{\trianglelefteq}_{\text{weak}} T$ but not $S <^{\trianglelefteq}_{\text{strong}} T$. Then

(i) $S \vartriangleleft T$;
(ii) S does not $\trianglelefteq$-recapture T;
(iii) $T \cup S$ is consistent.

By assumption, $\trianglelefteq$ is reflexive and monotone and so $T \trianglelefteq T \cup S$. Since, by (iii), $T \cup S$ is consistent, S $\trianglelefteq$-recaptures T, contradicting (ii). □

Proposition 5.1 connects (**DC2**) and (**DC3**) since all of the interpretations in this paper are monotone, and it is only the requirement "$\text{NewIso}(\tau)$" that is an obstacle for reflexivity. We would like to propose the idea of using a reflexive notion of interpretation $\trianglelefteq$ (such as $\trianglelefteq_{\text{ptim}}$) and then using weak $\trianglelefteq$-restrictiveness as a criterion for foundational theories. Arguments for using a notion such as weak $\trianglelefteq_{\text{ptim}}$-restrictiveness would be: (a) as a reflexive notion of interpretation, $\trianglelefteq$ fits better with the standard literature on interpretations; (b) by removing the conditions of "$\text{NewIso}(\tau)$" and "$S \cup T$ is inconsistent", we simplify the notion considerably; (c) the two conditions we are removing are the ones that are hardest

[11] For the proof, it is enough to have that $T \supseteq S$ implies $S \trianglelefteq T$ which immediately follows from reflexivity and monotonicity.

to argue for,[12] as they do not naturally arise from the informal explanation of the notion of restrictiveness.

§6. Acknowledgements. The first author acknowledges the support of Magdalene College, Cambridge during the time that this paper was written. The second author acknowledges the support of the Isaac Newton Institute for Mathematical Sciences during the programme *Semantics & Syntax* in 2012 and the hospitality of Corpus Christi College, Cambridge during his stay in Cambridge and subsequent visits.

BIBLIOGRAPHY

Aczel, P. (1988). *Non-Well-Founded Sets*. CSLI Lecture Notes, Vol. 14. Stanford, CA: CSLI.

Ahlbrandt, G., & Ziegler, M. (1986). Quasi-finitely axiomatizable totally categorical theories. *Annals of Pure and Applied Logic*, **30**(1), 63–82.

de Bouvère, K. L. (1965). Logical synonymity. *Indagationes Mathematicae*, **27**, 622–629.

Hamkins, J. D. (2013). A multiverse perspective on the axiom of constructibility. In Chong, C., Feng, Q., Slaman, T. A., and Woodin, W. H., editors. *Infinity and Truth*. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, Vol. 25. Singapore: World Scientific, pp. 25–45.

Kanamori, A. (2003). *The Higher Infinite, Large cardinals in set theory from their beginnings* (second edition). Springer Monographs in Mathematics. Berlin: Springer-Verlag.

Löwe, B. (2001). A first glance at non-restrictiveness. *Philosophia Mathematica*, **9**(3), 347–354.

Löwe, B. (2003). A second glance at non-restrictiveness. *Philosophia Mathematica*, **11**(3), 323–331.

Maddy, P. (1988a). Believing the axioms. I. *Journal of Symbolic Logic*, **53**(2), 481–511.

Maddy, P. (1988b). Believing the axioms. II. *Journal of Symbolic Logic*, **53**(3), 736–764.

Maddy, P. (1997). *Naturalism in mathematics*. New York: The Clarendon Press Oxford University Press.

Maddy, P. (1998). $V = L$ and MAXIMIZE. In Makowsky, J. A. and Ravve, E. V., editors. *Logic Colloquium '95, Papers from the colloquium held as part of the European Summer Meeting of the Association for Symbolic Logic in Haifa, August 9–18, 1995*. *Lecture Notes Logic*, Vol. 11. Berlin: Springer, pp. 134–152.

Tarski, A., Mostowski, A., & Robinson, A. (1953). *Undecidable Theories*. Amsterdam: North-Holland.

[12] What is Maddy's argument for the condition "$S \cup T$ is inconsistent" in (Maddy, 1997, pp. 222–223)? She considers ZFC and $\mathsf{ZFC} + \mathsf{IC}$, observes that $\mathsf{ZFC} \lhd_{\text{Maddy}} \mathsf{ZFC} + \mathsf{IC}$, and concludes "[b]ut, it seems wrong to say that ZFC is restrictive simply because it doesn't assert the existence of an inaccessible cardinal" (Maddy, 1997, p. 222). She proposes that "for a theory T to be genuinely restrictive, the maximizing theory T must actually contradict T" (Maddy, 1997, pp. 222–223). However, we can easily see that $\mathsf{ZFC} \trianglelefteq_{\text{Maddy}}$-recaptures $\mathsf{ZFC} + \mathsf{IC}$ (just take $\mathsf{ZFC} + 2\mathsf{IC}$), so the extra condition of inconsistency is not needed to resolve this potential issue. This type of argument can be generalized to any theory T such that $T + \exists\kappa(\text{Inacc}(\kappa) \wedge \mathbf{V}_\kappa \models T)$ is consistent, which is a reasonable assumption for all $T = \mathsf{ZFC} + \mathsf{A}$ where A is a large cardinal axiom.

Visser, A. (2006). Categories of theories and interpretations. In Enayat, A., Kalantari, I., and Moniri, M., editors. *Logic in Tehran, Proceedings of the Workshop and Conference on Logic, Algebra and Arithmetic held in Tehran, October 18–22, 2003*. Lecture Notes in Logic, Vol. 26. La Jolla, CA: Association for Symbolic Logic, pp. 284–341.

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION
UNIVERSITEIT VAN AMSTERDAM
POSTBUS 94242, 1090 GE AMSTERDAM, THE NETHERLANDS
E-mail: {l.incurvati,b.loewe}@uva.nl

AFDELING WIJSBEGEERTE
UNIVERSITEIT VAN AMSTERDAM
OUDE TURFMARKT 141-147, 1012 GC AMSTERDAM, THE NETHERLANDS

FACHBEREICH MATHEMATIK
UNIVERSITÄT HAMBURG
BUNDESSTRASSE 55, 20146 HAMBURG, GERMANY