Interpretability logics and large cardinals in set theory

Joost J. Joosten

March 12, 2010

Abstract

This paper is a first exploration how the modal logic of interpretability can play a role in the study of large cardinals in set theory. First we reason that such a role is likely to be fruitful. Next, for two particular readings of the binary modality $\rhd$ we study its repercussions for the corresponding logic.

The first reading of $A \rhd B$ is "Inside any V_κ where A holds we can find a $V_{\kappa'}$ where B holds". Here κ and κ' range over inaccessible cardinals. We show how this reading can be related and reduced to previous results by Solovay.

The second reading of $A \rhd B$ is "Inside any V_κ where A holds we can find a V_λ where B holds". Here κ ranges over one particular large (at least inaccessible) cardinal notion and λ ranges over a different (but also at least inaccessible) large cardinal. We see that in this case a bi-modal logic suffices. Moreover, we obtain that for an ample class of cardinal notions the corresponding logic will not include the basic modal logic **IL**.

1 Introduction

This paper is a first exploration how the modal logic of interpretability can play a role in the study of large cardinals in set theory. In this paper we shall introduce the important notions in set theory and sketch/reason where and why modal logics are likely to play a fruitful role.

1.1 Independence in set theory

With the foundational crisis of mathematics at the turn of the 19th century mathematicians/logicians craved for more mathematical rigor. In mathematical practice this resulted in that more and more parts of mathematics got formalized to the very detail. Set theory grew to be the *golden standard* in this tendency to formalizing and large parts of mathematics obtained their rigor from their formalization within set theory.

As a consequence one particular axiomatization of set theory itself, what we now call *Zermelo Fraenkel* set theory, got well developed and studied in great detail. In the course of this study it was found that many elementary and natural questions can not be settled within set theory itself. The most prominent such example is probably the question of the Continuum Hypothesis, CH, stated boldly as "Is there a subset of the reals that is strictly larger[1] than the natural numbers $\mathbb{N}$ but strictly smaller than $\mathbb{R}$, the set of reals itself?" CH claims that there is no such subset of the reals. This question was the first of Hilbert's 23 questions that he considered most urgent at the turn of the 19th century.

In 1940 Gödel showed ([5]) that CH cannot be refuted by the axioms of set theory by providing a model of ZF (actually of ZFC) where CH holds. In addition, Paul Cohen ([3], [4]) showed in his seminal work from 1963 that CH is also not provable from set theory by introducing his celebrated technique of forcing. These two results showed that CH is actually independent from set theory.

Various other natural questions in set theory also turned out to (likely) be independent from set theory. Among those are the existence assertions of so-called *large cardinals*. For the purpose of this paper very little actual set theoretical detail is needed and for the moment it suffices to superficially describe the large cardinals. Large cardinals are (infinite) cardinal numbers with some specific very strong closure properties. A *large cardinal axiom* states the existence of such large cardinals.

1.2 Modal logics and mathematics

Numerous versions of large cardinal axioms are around and most of them are likely to be independent from set theory. The natural question is how these large cardinal axioms relate to each other and which of them are natural from which point of view without even entering the discussion of Platonism/Realism in mathematics. In this enterprise, the author thinks, modal logics can play a guiding role.

Modal logic as a calculus

For our purposes, modal logics are calculi developed to reason about a particular part of mathematics. The most useful modal logics are formulated in an easy -often propositional- language and are decidable. They typically contain one or more modalities to capture a particular feature of mathematics. A classical example is Solovay's study of provability ([14]) where the $\Box$ modality is used to study the notion of "provable in, say, Peano Arithmetic".

The beauty and power of modal logics lies in the following. Large parts of mathematical reasoning can be captured in a single axiom scheme or rule. For example, in provability logic the axiom scheme $\Box A \to \Box\Box A$ reflects the theorem of so-called Provable Σ^0_1 Completeness: any true Σ^0_1 sentence is actually provable. Thus, as the axiom schemata and rules in a modal logic correspond to whole blocks of reasoning and entire theorems in mathematics, modal logics provide a very high level reason tool par excellence.

[1] In the sense of cardinality.

The situation is very much comparable to a calculus for, say, integration. This calculus comprises a set of rules like $\int x^n dx = \frac{1}{n+1}x^{n+1}$. But for example this particular rule represents a whole block of reasoning including upper and lower Riemann approximations of the area under the function x^n that can be analyzed in their limit behavior importing other results like for example the binomial expansion theorem etcetera.

The difference with a modal calculus is that this calculus is formulated in a logic framework and thus lends itself naturally to a study of the limits of the calculus. To conclude, modal logics provide a natural and high-level framework to analyze and study mathematical phenomena. As such they can play the role as an excellent viewpoint on the naturality of certain independent principles. In the context of set theory modal logic is already used for this purpose.

Modal logics and set theory

In [7], Hamkins used a modal framework where the $\Box$ was used to model "truth in all forcing extensions". On a high level he could formulate a natural principle inspired by C. Chalons. This principle states that any set-theoretical statement φ is true whenever φ is such that once true in a certain forcing extension $\mathbb{P}$, it stays true in *any* subsequent forcing extension of $\mathbb{P}$. Or as Hamkins phrases it in his plain language slogan: "*anything forceable and not subsequently unforceable is true*". In modal language the scheme is represented by $\Diamond\Box\varphi \to \varphi$ which in this setting is actually equivalent to $\Diamond\Box\varphi \to \Box\varphi$ which is more familiar to modal logicians and is known as the Euclidean Axiom.

It turns out that this principle is equi-consistent with set theory. However, once the principle is allowed to have real valued parameters in φ various interesting statements about large cardinals and in particular various large cardinal axioms follow. Thus these particular large cardinals are natural in the viewpoint provided by this modal setting that allowed so elegantly to formulate a general principle.

This paper can be seen as a first exploration to extending an endeavor as in Hamkins' [7] to a modal language of a richer signature. Such a logic of richer signature is interpretability logic. This logic has apart from the unary modality $\Box$ a binary modality $\rhd$. This modality is used so that $U \rhd V$ is understood to model that a theory U interprets a theory V or alternatively, $A \rhd B$ is understood to model that for some basic theory T, the theory $T + A$ interprets the theory $T + B$. In the paper we shall focus on possible set-theoretical readings of $\rhd$. In looking for such readings, we are guided by two criteria, Criterium **S** and Criterium **I**.

- **S** The set-theoretical content of the translation should be interesting.
- **I** The translation of the $\rhd$-modality should somehow capture the most typical interpretabilty features. In other words, the modal logic should be interesting.

In interpretability logics we often use a unary modality $\Box$ and a binary modality $\rhd$. The $\Box$ modality can actually be dispensed with as we want $\Box A \leftrightarrow \neg A \rhd \bot$ to be a valid principle.

Modal logics, heuristics and predictive power

We believe that a good modal framework provides good heuristics and a rich viewpoint for analyzing and exploring certain parts of mathematics. And in this believe and in particular that interpretability logics can play the role of a good viewpoint we are sustained by historical evidence. Recently, in [6] a new mathematical theorem is obtained from pure modal desiderata. It was observed that for a certain modal logic –which was known to be arithmetically sound– to be modally complete, a particular modal principle had to hold. As it turned out, the mathematical content of this modal principle actually turned out to be a mathematical theorem.

This practice is somewhat similar as observed in other sciences. When the language and the formalism is good, valid predictions are bound to follow. For example, Albert Einstein's theory of General Relativity on a purely theoretical basis predicted certain phenomena like the perihelion precession of Mercury's orbit, and gravitational redshift of light which were later empirically confirmed. Likewise, the existence of certain elementary particles that go by the name of mesons were first predicted in 1935 by Hideki Yukawa on purely theoretical grounds and were later confirmed by experiments. Of course, in the realm of set theory such an empirical counterpart is unlikely to be present.

1.3 Basic notions in set theory

As mentioned before, for the purpose of this paper, we need very little detail on the actual results in set theory. We refer the reader to [11] and [12] for details. In particular in this paper we work with the notions of inaccessible cardinals, forcing extensions, and he cumulative hierarchy of the set theoretical universe. In particular we shall use that when κ is an inaccessible cardinal, then V_κ, the "κ"th level of the cumulative hierarchy, is a model of set theory.

2 Interpretability logics

Interpretability logics are an extension of provability logic. The modal language contains apart from the usual unary $\Box$ modality a binary $\rhd$ modality that should capture the notion of interpretability.

A short word on reading conventions in interpretability logics is due here. For the boolean connectives we have the usual conventions. The $\Box$ and $\Diamond$ have the same syntactical treatment and preference as the negation. The $\rhd$-modality is binding weaker than all the previously but stronger than $\rightarrow$.

If we want a set-theoretical reading of the $\rhd$-modality to be reminiscent of the notion of interpretability so that we have access to its rich structure and well-developed theory, we should require that some minimal interpretability logic should be sound under the translation. This minimal part is the logic **IL**.

Definition 2.1. The logic **IL** is the modal logic generated by all propositional tautologies, the rule of modus ponens and the rule of necessitation ($\vdash A / \vdash \Box A$) together with the following axiom schemes.

$$\begin{array}{ll}
\mathsf{L}_1: & \Box(C \to D) \to (\Box C \to \Box D) \\
\mathsf{L}_2: & \Box A \to \Box\Box A \\
\mathsf{L}_3: & \Box(\Box A \to A) \to \Box A \\
\mathsf{J}_1: & \Box(C \to D) \to C \rhd D \\
\mathsf{J}_2: & (C \rhd D) \wedge (D \rhd E) \to C \rhd E \\
\mathsf{J}_3: & (C \rhd E) \wedge (D \rhd E) \to C \vee D \rhd E \\
\mathsf{J}_4: & C \rhd D \to (\Diamond C \to \Diamond D) \\
\mathsf{J}_5: & \Diamond A \rhd A
\end{array}$$

Two other prominent principles are

$$\mathsf{M} \ : \ A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$$

and

$$\mathsf{P} \ : \ A \rhd B \to \Box(A \rhd B).$$

The logic that arises by adding more axiom schemes to **IL** is denoted by **IL** with the names of the principles postfixed to it.

Often we consider the part of **IL** that only contains the $\Box$-modality. This is the logic GL.

Definition 2.2. The logic GL is the logic in a modal language only containing the $\Box$-modality that is closed under modus ponens, necessitation and the axiom schemes L_1, L_2, and L_3. The logic K4 is just GL with omission of L_3.

Let us do a small exercise in proofs in **IL** that will serve us well later on.

Lemma 2.3. **IL** *proves all of the following formulas.*

1. $A \rhd A$

2. $\Box A \leftrightarrow \neg A \rhd \bot$

3. $\Diamond \bot \leftrightarrow \bot$

Proof.

Ad 1 $A \to A$ is a tautology, whence by necessitation we obtain $\Box(A \to A)$. By J_1 we obtain $A \rhd A$.

Ad 2 By basic modal reasoning we get $\Box A \to \Box(\neg A \to \bot)$. But from $\Box(\neg A \to \bot)$ we arrive at $\neg A \rhd \bot$ by an application of J_1.

Ad 3 The implication $\bot \to \Diamond\bot$ is obvious. And $\Diamond\bot \to \bot$ is equivalent to $\Box\top$ which is theorem of GL as we saw in 1.

$\dashv$

For **IL** we have good modal semantics that provide intuition and guidance in reasoning about the logic. The semantics is an extension of the well known semantics for GL.

Definition 2.4. An **IL** frame is a triple $\langle W, R, \{S_x \mid x \in W\}\rangle$. Here W is a set of *worlds*. The R is a binary relation on W that models the $\Box$ modality. For each $x \in W$, the S_x is a binary relation on W that models the $\rhd$ modality. The requirements on the R and S_x relations are as follows.

1. R is transitive;
2. R is conversely well-founded, that is, there is no infinite chain $x_0Rx_1Rx_2R\ldots$;
3. For each $x \in W$, the S_x is a binary relation on $x{\uparrow}$ where $x{\uparrow} := \{y \in W \mid xRy\}$;
4. For each $x \in W$, the S_x is reflexive;
5. For each $x \in W$, the S_x is transitive;
6. $R \upharpoonright (x{\uparrow}) \subseteq S_x$ in other words/symbols $xRyRz \to yS_xz$.

Definition 2.5. An **IL** model is a quadruple $\langle W, R, \{S_x \mid x \in W\}, \Vdash\rangle$ where $\langle W, R, \{S_x \mid x \in W\}\rangle$ is an **IL** model and $\Vdash$ is a so-called forcing relation that assigns to each propositional variable p a set of worlds in W where p holds (we also say where p is forced). The forcing relation $\Vdash$ is extended to the set of all modal formulas by stipulating that

+ $x \Vdash A \wedge B \quad\Leftrightarrow\quad [x \Vdash A$ and $x \Vdash B]$ and likewise for other Boolean connectives;

+ $x \Vdash \Box A \quad\Leftrightarrow\quad \forall y[xRy \Rightarrow y \Vdash A]$;

+ $x \Vdash A \rhd B \quad\Leftrightarrow\quad \forall y[xRy \wedge y \Vdash A \Rightarrow \exists z(yS_xz \wedge z \Vdash B)]$.

It is well known that **IL** is sound and complete with respect to its modal semantics.

3 Inaccessible cardinals

It was Solovay who first studied some modal logics of set theory. Some of his findings he did not publish but were included in the book by Boolos ([2]). In particular, Solovay gave a characterization of the modal formulas that hold if the $\Box$ is taken to denote a formalization of "True in all set theory models V_κ where κ is an inaccessible cardinal". In order to formulate his result, we first need a definition.

Definition 3.1. The logic **J** is is obtained by adding the linearity axiom schema $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ to GL. Here $\boxdot B$ is an abbreviation of $B \wedge \Box B$.

Solovay's theorem now reads as follows. (See [2], Theorem 2 of Chapter 13.)

Theorem 3.2. $\mathbf{J} \vdash A \Leftrightarrow \forall * \ \text{ZFC} \vdash A^*$

In this theorem, the $*$ is a so-called *realization* that sends propositional variables to sentences in the language of ZFC. The $*$ is extended to sentences in the obvious way, by, as mentioned before, translating the $\Box$ to "true in all V_κ". As always, realizations are understood to commute with the Boolean connectives in that $(A \wedge B)^* := A^* \wedge B^*$ etcetera.

3.1 Interpretability and inaccessible cardinals

Inspired by Solovay's result, Theorem 3.2, a first idea is to read the $\rhd$ as follows.

$$(A \rhd B)^* := \quad \text{"Inside every } V_\kappa \text{ where } A^* \text{ holds, we can find a } V_{\kappa'} \text{ where } B^* \text{ holds."}$$

Here κ and κ' are inaccessibles. As $\mathbf{IL} \vdash A \rhd A$, the reading should be slightly adapted.

$$(A \rhd B)^* := \quad \text{"Inside every } V_\kappa \text{ where } A^* \text{ holds, either } B^* \text{ holds, or we can find a } V_{\kappa'} \text{ where } B^* \text{ holds."} \qquad (\dagger)$$

With κ and κ' again inaccessibles. Clearly, under this reading we have that $A \rhd A$ holds for any A. A more technically precise but otherwise equivalent formulation of this interpretation is given by

$$(A \rhd B)^* := \quad \forall \kappa\, [V_\kappa \models A^* \Rightarrow (V_\kappa \models B^* \vee V\kappa \models \text{"}\exists \kappa' V_{\kappa'} \models B^*\text{"})] \qquad (\dagger)$$

As mentioned before, in **IL**, the $\rhd$ modality is tightly bound to the regular $\Box$ modality by the fact that $\mathbf{IL} \vdash \Box A \leftrightarrow \neg A \rhd \bot$. If we wish to preserve this duality between the modalities under our $(\dagger)$ interpretation, we are left with no choice for the interpretation of the $\Box$ modality as expressed in the following lemma.

Lemma 3.3. *If the equivalence $\Box A \leftrightarrow \neg A \rhd \bot$ is to hold under the $(\dagger)$ interpretation, the $\Box$-modality should be translated as "true in all V_κ".*

Proof. We only use properties of the compositionallity of both the Tarski truth definition and of the definition of a set-theoretical realization.

$$\begin{aligned}
(\Box A)^* &\Leftrightarrow (\neg A \rhd \bot)^* \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models (\neg A)^* \Rightarrow (V_\kappa \models (\bot)^* \vee V\kappa \models \text{"}\exists \kappa' V_{\kappa'} \models (\bot)^*\text{"})] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models \neg A^* \Rightarrow (V_\kappa \models 0 = 1 \vee V\kappa \models \text{"}\exists \kappa' V_{\kappa'} \models 0 = 1\text{"})] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models \neg A^* \Rightarrow (\bot \vee V\kappa \models \text{"}\exists \kappa' V_{\kappa'} \models 0 = 1\text{"})] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models \neg A^* \Rightarrow (V\kappa \models \text{"}\exists \kappa' V_{\kappa'} \models 0 = 1\text{"})] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models \neg A^* \Rightarrow (V\kappa \models 0 = 1)] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models \neg A^* \to 0 = 1] \\
&\Leftrightarrow \forall \kappa\, [V_\kappa \models A^*]
\end{aligned}$$

$\dashv$

We remark that various steps in the above proof have their modal counterpart in Lemma 2.3. With similar reasoning, so that we allow ourselves some shortcuts, we see that under the (†) translation using inaccessibles we get the following theorem. The theorem states that by and large the $\rhd$ modality can be reduced to the $\Box$ modality.

Theorem 3.4. *When translating $\rhd$ with the scheme as in (†) and the $\Box$ modality as in Lemma 3.3, the following is a sound priciple:*

$$A \rhd B \leftrightarrow \Box(A \to B \vee \Diamond B). \quad (i)$$

Proof.

$$\begin{aligned}
(A \rhd B)^* \quad &\Leftrightarrow \quad \forall V_\kappa\, [V_\kappa \models A^* \Rightarrow (V_\kappa \models B^* \vee V_\kappa \models \text{“}\exists V_{\kappa'} V_{\kappa'} \models B^*\text{”})] \\
&\Leftrightarrow \quad \forall V_\kappa\, [V_\kappa \models A^* \to (B^* \vee \text{“}\exists V_{\kappa'} V_{\kappa'} \models B^*\text{”})] \\
&\Leftrightarrow \quad \forall V_\kappa\, [V_\kappa \models A^* \to (B^* \vee \Diamond B^*)] \\
&\Leftrightarrow \quad \Box(A^* \to (B^* \vee \Diamond B^*)) \\
&\Leftrightarrow \quad \Box(A^* \to B^* \vee \Diamond B^*)
\end{aligned}$$

$\dashv$

As a corollary of this theorem, we see that the soundness of **IL** under the (†) translation becomes just an exercise in GL or actually in K4.

Lemma 3.5. *Each of the axioms $\mathsf{J}_1 \ldots \mathsf{J}_5$ is provable in* K4 *when each occurrence of $A \rhd B$ is replaced by $\Box(A \to B \vee \Diamond B)$.*

Proof. This is an easy exercise in K4. The most involved axiom is J_2. We shall briefly comment on J_4. Thus, we need to show that the translation of $A \rhd B \to (\Diamond A \to \Diamond B)$ is provable in K4. To this extent, we translate $A \rhd B$ to $\Box(A \to B \vee \Diamond B)$ and reason in K4. An elementary theorem of K4 tells us that $\Box(C \to D)$ implies $\Diamond C \to \Diamond D$ whence we obtain $\Diamond A \to \Diamond(B \vee \Diamond B)$. But, $\Diamond(B \vee \Diamond B)$ is actually equivalent to $\Diamond B \vee \Diamond\Diamond B$ and by an application of the L_2 axiom we see that the latter is equivalent to just $\Diamond B$. Quad erat demonstrandum. $\dashv$

Theorem 3.6. *The logic* **IL** *is sound under translating*

- *$A \rhd B$ to “Inside each V_{kappa} where A holds: either B holds or we can find a $V_{\kappa'}$ where B holds”;*
- *$\Box A$ to “A holds in each V_κ”.*

Proof. By 3.4 we see that under our translation, we can replace any occurrence of $A \rhd B$ by $\Box(A \to B \vee \Diamond B)$. But then, each proof in **IL** translates by 3.5 to a proof in K4. As K4 $\subset$ GL we get the soundness by Theorem 3.2. $\dashv$

We thus conclude that under (†) the logic **IL** is sound. The next subsection deals with completeness. As we shall see, for completeness some additional principle needs to be added.

3.2 Characterizations of the interpretability logic for inaccessible cardinals

We can give a complete characterization of all the principles under the above reading of $\rhd$. First some definitions.

Definition 3.7. The logic $\mathbf{J}'$ is the modal logic with a unary modality $\Box$ and a binary modality $\rhd$. It has the same rules and axioms as $\mathbf{J}$. In addition it has the axiom scheme $A \rhd B \leftrightarrow \Box(A \to B \vee \Diamond B)$.

Lemma 3.8. *The logic* $\mathbf{J}'$ *is conservative over* $\mathbf{J}$*.*

Proof. The logic $\mathbf{J}'$ is just an extension by definition of $\mathbf{J}$ and we can thus obtain our result by a simple definition on the proofs in $\mathbf{J}'$. $\dashv$

Corollary 3.9 (First characterization)**.** $\mathbf{J}' \vdash A \Leftrightarrow \forall +\ \mathrm{ZFC} \vdash A^+$

Recall, in this corollary, the $+$ is a *realization* that sends propositional variables to sentences in the language of ZFC. The $+$ is extended to sentences in the obvious way, translating the $\Box$ to "true in all V_κ" and $A \rhd B$ to "In every V_κ where A^+ holds, either B^+ holds, or we can find a $V_{\kappa'}$ where B^+ holds." We use the $+$ symbol here to distinguish from the $*$ in Theorem 3.2.

Proof. We define a translation on modal formulas as follows. It will be the identity translation except for $\rhd$. In that case we define

$$(A \rhd B)^{\mathsf{tr}} := \Box(A^{\mathsf{tr}} \to (B^{\mathsf{tr}} \vee \Diamond B^{\mathsf{tr}})).$$

Clearly we have for any formula A that $\mathbf{J}' \vdash A \leftrightarrow A^{\mathsf{tr}}$. Thus $\mathbf{J}' \vdash A \Leftrightarrow \mathbf{J}' \vdash A^{\mathsf{tr}}$. By Lemma 3.8 we see that the latter is equivalent to $\mathbf{J} \vdash A^{\mathsf{tr}}$. By Theorem 3.2 we see that this is equivalent to $\forall *\ \mathrm{ZFC} \vdash (A^{\mathsf{tr}})^*$. Clearly, for every $+$ we can find a $*$ such that $A^+ = (A^{\mathsf{tr}})^*$ and also the other way round. And thus we get the required equivalence, that is, $\mathbf{J}' \vdash A \Leftrightarrow \forall +\ \mathrm{ZFC} \vdash A^+$. $\dashv$

This characterization is not very satisfactory as it is not (essentially) formulated in the language using $\rhd$.

In Theorem 3.6 we saw that all of **IL** is provable in $\mathbf{J}'$. It is actually not hard to see that the principles M and P are also provable in $\mathbf{J}'$. For those familiar in interpretability logics, this is known to be a sign that the logic is of little interest as **IL**P and **IL**M normally characterize different sort of logics. In other words, if a logic proves both M and P this logic is probably not very informative.

We can reformulate our first characterization in a setting with a stronger interpretability flavor.

Definition 3.10. The logic **IL**LW is obtained by adding the linearity axiom schema: $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ to **IL**W. Here W is the axiom scheme $A \rhd B \to A \rhd B \wedge \Box\neg A$.

Theorem 3.11 (Second characterization)**.** $\mathbf{IL}\mathsf{LW} \vdash A \Leftrightarrow \forall +\ \mathrm{ZFC} \vdash A^{+}$.

The proof of this theorem was actually first encountered in the realm of provability logics with restricted substitutions as dealt with in [10] and in [9].

Proof. We again consider the translation as defined in the proof of Corollary 3.9. We will see that

$$\begin{array}{cc} \mathbf{IL}\mathsf{LW} \vdash \varphi \Leftrightarrow \mathbf{J} \vdash \varphi^{\mathsf{tr}} & (*) \\ \text{and} & \\ \mathbf{IL}\mathsf{LW} \vdash \varphi \leftrightarrow \varphi^{\mathsf{tr}}. & (**) \end{array}$$

We first see that we have $(**)$. It is sufficient to show that $\mathbf{IL}\mathsf{LW} \vdash p \rhd q \rightarrow \Box(p \rightarrow (q \vee \Diamond q))$. We reason in $\mathbf{IL}\mathsf{LW}$. An instantiation of the linearity axiom gives us $\Box(\Box\neg q \rightarrow (\neg p \vee q)) \vee \Box((\neg p \vee q) \wedge \Box(\neg p \vee q) \rightarrow \neg q)$. The first disjunct immediately yields $\Box(p \rightarrow (q \vee \Diamond q))$.

In case of the second disjunct we get by propositional logic $\Box(q \rightarrow \Diamond(p \wedge \neg q))$ and thus also $\Box(q \rightarrow \Diamond p)$. Now we assume $p \rhd q$. By W we get $p \rhd q \wedge \Box\neg p$. Together with $\Box(q \rightarrow \Diamond p)$, this gives us $p \rhd \bot$, that is $\Box\neg p$. Consequently we have $\Box(p \rightarrow (q \vee \Diamond q))$.

We now prove $(*)$. By induction on $\mathbf{IL}\mathsf{LW} \vdash A$ we see that $\mathbf{J} \vdash A^{\mathsf{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in GL. The only axioms where the $\Box A \rightarrow \Box\Box A$ axiom scheme is really used is in J_2 and J_4. To prove the translation of W we also need L_3.

If $\mathbf{J} \vdash A^{\mathsf{tr}}$ then certainly $\mathbf{IL}\mathsf{LW} \vdash A^{\mathsf{tr}}$ and by $(**)$, $\mathbf{IL}\mathsf{LW} \vdash \varphi$.

We now invoke Theorem 3.2 and combine it with $(*)$ to see that $\mathbf{IL}\mathsf{LW} \vdash A \leftrightarrow \forall *\ \mathrm{ZFC} \vdash (A^{\mathsf{tr}})^{*}$. Again we realize that for every $+$ we can find a $*$ such that $A^{+} = (A^{\mathsf{tr}})^{*}$ ans also the other way round. Thus we obtain $\mathbf{IL}\mathsf{LW} \vdash A \leftrightarrow \forall +\ \mathrm{ZFC} \vdash (A^{\mathsf{tr}})^{+}$.

⊣

We note that we could have defined $\mathbf{IL}\mathsf{LW}$ also starting with $\mathbf{IL}\mathsf{M}$ or $\mathbf{IL}\mathsf{P}$. We have chosen to use $\mathbf{IL}\mathsf{W}$ as the W-principle is the minimal principle for which our argument seems to work.

Let us summarize what we have obtained now. We have chosen a set-theoretical reading of $\rhd$ using inaccessibles. We have found a completeness result à la Solovay. Actually we have found two different characterizations. Are we satisfied now? Not really.

Especially the the first characterization tells us that our set-theoretical reading of $\rhd$ does not fully employ the richness of the interpretability logic. The second characterization tries to hide this a bit. But a second moment of thought again reveals the simple nature of the logic.

We therefore consider the modal semantics of $\mathbf{IL}$. If the linearity axiom is to hold on a frame, we are bounded to linear frames. The S_x-relations (the semantical counterpart of the $\rhd$ modality) can do nothing on linear frames that can not be done by the R-relation.

Thus Criterium **I** actually rules out the reading we have studied so far. But this should come hardly as a surprise. The very way we defined the interpretation of $\rhd$ in (†) is expressible in the language of modal logic with just the $\Box$ modality. An interesting reading should capture somehow the feature that the $\Box$ modality is really stronger than the $\rhd$ modality in the sense of Axiom J_1. We can adopt thus a new desideratum. We do not want that

$$A \rhd B \to \Box(A \to B \vee \Diamond B) \quad (\mathsf{Tr})$$

is a valid principle for the set-theoretical reading of $\rhd$ and $\Box$. If for the $\Box$ we use an inaccessable (or any large (at least inaccessible) cardinal, as Lemma 4.4 tells us) we get the linearity axiom into our logic. It is easy to see that (Tr) holds on any linear frame. We thus come to the following natural and relevant question.

Question 3.12. Is $\mathbf{IL}\mathsf{L}$ a complete logic? In particular, do we have $\mathbf{IL}\mathsf{L} \vdash A \rhd B \to \Box(A \to B \vee \Diamond B)$?

4 Pairs of large cardinals and bimodal logics

In finding a suitable set-theoretical reading of the $\Box$ and $\rhd$ modality we want to capture the notion that $A \rhd B$ is somewhat weaker than $\Box(A \to B \vee \Diamond B)$. One idea would be to read $A \rhd B$ as follows.

$A \rhd B \quad :=$ "In every model of V_κ where A holds, either B holds, or there is a model V_λ of B"

Here and in the sequel of this section when referring to a pair κ-λ, κ is a large (at least inaccessible) cardinal of type 1 and λ a large (at least inaccessible) cardinal of type 0.

We see that, similar to Theorem 3.4, we have the following principle to be sound under this reading.

$$A \rhd B \leftrightarrow [1](A \to B \vee \langle\ \rangle B)$$

Here the [1]-modality is to be read as "truth in all V_κ" with κ of type 1. The [ ]-modality is to be read as "truth in all V_λ" with λ of type 0. Thus actually $\rhd$ gets translated to a statement in a bimodal logic.

As a bimodal logic, this is interesting to study and it tells us something about how the κ and the λ behave with respect to each other. The first interesting pair of large cardinals already yields an interesting question.

Question 4.1. What is the bimodal logic where the [1]-modality is read as "true in all V_κ" with κ mahlo and [ ] is read as "true in all V_λ" with λ inaccessible.

As we shall see, in various cases it is not very likely though, that the new reading of the $\rhd$-modality will serve our purposes. We conclude this from some "empirical" facts from set-theory.

Empirical Fact 4.2. Somehow, all large cardinal axioms seem to be linearly ordered along their consistency strength. To be more precise, for each pair of cardinal axioms A_κ, A_λ we have that one of the following three situations hold.

(i) ZFC proves [$\mathrm{ZFC} + A_\kappa$ is consistent if and only if $\mathrm{ZFC} + A_\lambda$ is consistent]

(ii) $\mathrm{ZFC} + A_\kappa$ proves the consistency of $\mathrm{ZFC} + A_\lambda$

(iii) $\mathrm{ZFC} + A_\lambda$ proves the consistency of $\mathrm{ZFC} + A_\lambda$

This empirical fact, which so far has never been violated, has no direct bearing on the modalities introduced in this section. It would be expressible in modal logic if $\Diamond A$ were to be interpreted as "A holds in some model of set theory". However, in this section we restrict ourselves to the models of set theory of the form V_α.

Moreover, the ordering in consistency strength of two large cardinal axioms does not imply anything on the ordering of the actual ordinals that are asserted to exist by the respective axioms. For example ([13]), the existence of a *huge* cardinal is much stronger than the existence of a *supercompact* cardinal in terms of consistency strength. However, if both existed, then huge cardinals are smaller than supercompact ones in terms of the ordinal ordering.

Empirical Fact 4.3. For various pairs of large cardinals we have that they are comparable in our bimodal logic setting in the sense that $[1]\langle\ \rangle\top$ or $[\ ]\langle 1\rangle\top$.

As Benedikt Löwe pointed out te me this holds at least for pairs of cardinals that are taken from the following list of large cardinals:

supercompact
Woodin
measurable
Mahlo
inaccessible

This second empirical fact, in combination with a side assumption, will be seen to be problematic in combination with the $\mathsf{J_4}$-axiom:

$$A \rhd B \to (\langle 1\rangle A \to \langle 1\rangle B).$$

First we remark the following.

Lemma 4.4. GL *is sound for any translation of the* $\Box$*-modality into "truth in every* V_κ*", where* κ *ranges over some large (at least inaccessable) cardinal.*

Proof. The proof is a straight-forward generalization of the proof of this statement for inaccessibles. ⊣

Definition 4.5. A κ-λ *interpretability principle* of ZFC is a sentence A in the modal language with $\rhd$ and $\Box$ for which we have

$$\forall\dagger\ \mathrm{ZFC} \vdash A^\dagger.$$

Here † is a mapping that sends propositional variables to set-theoretical sentences. This mapping is extended to all modal sentences in the canonical way by translating $\Box$ to "true in all V_κ" and $C \rhd D$ to "in all V_κ where C holds, either D holds, or we can find a V_λ where D holds". Here κ ranges over cardinals of type 1, and λ over cardinals of type 0.

Theorem 4.6. *Let κ and λ come from two different natural large cardinal classes for which our Empirical Fact 4.3 holds. If T is a complete axiomatization of the κ-λ interpretability principles of* ZFC*, and if moreover, $T \nvdash [1][1]\bot$ then T does not contain* **IL**.

Proof. Suppose T is a complete axiomatization of the κ-λ interpretability principles of ZFC and suppose that T does contain **IL**. By Lemma 4.4, T should contain all the theorems of GL for the $[\,]$-modality. By the Empirical Fact 4.3 we have either $[1]\langle\,\rangle\top$ or $[\,]\langle 1\rangle\top$. In the first case we get by Lemma 4.7 that $[1]\bot$. In the second case we get by Lemma 4.8 that $[1][1]\bot$. Bot are in contradiction with our side assumption that $T \nvdash [1][1]\bot$. ⊣

The side assumption that $T \nvdash [1][1]\bot$ is really needed. In particular, for most large cardinals the axiom stating that there exist at least two such cardinals is strictly stronger than the axiom that merely asserts the existence of just at least one. However, it seems unlikely that there are mathematicians/logicians that do believe in the existence[2] of one particular cardinal of kind X but reject that there would be more of that kind of cardinal around too. We conclude this section by providing the two technical lemmata needed in the proof of Theorem 4.6.

Lemma 4.7. *Let T be a logic with one binary modality $\rhd$ and two unary modalities $[1]$ and $[\,]$ such that T contains all the axioms of* **IL** *(formulated with $\rhd$ and $[1]$) and is closed under the rules of modus ponens and necessitation for the $[1]$-modality. Furthermore, T is supposed to prove all the theorems of* GL *for the $[\,]$-modality. If now $T \vdash [1]\langle\,\rangle\top$, then $T \vdash [1]\bot$.*

Proof. So, let T satisfy the conditions of the theorem. We have that $T \vdash \langle\,\rangle\top \vee [\,]\bot$. By L_3 for the $[\,]$-modality, $\langle\,\rangle\top$ is actually equivalent to $\langle\,\rangle[\,]\bot$. Thus, $T \vdash \langle\,\rangle[\,]\bot \vee [\,]\bot$. By [1]-necessitation, $T \vdash [1](\langle\,\rangle[\,]\bot \vee [\,]\bot)$ and thus certainly $T \vdash [1](\top \rightarrow (\langle\,\rangle[\,]\bot \vee [\,]\bot))$. By J_1 this yields $\top \rhd [\,]\bot$. J_4 now gives us $\langle 1\rangle\top \rightarrow \langle 1\rangle[\,]\top$. Together with our assumption that $T \vdash [1]\langle\,\rangle\top$, we thus get $T \vdash [1]\bot$. ⊣

Lemma 4.8. *Let T be a logic with one binary modality $\rhd$ and two unary modalities $[1]$ and $[\,]$ such that T contains all the axioms of* **IL** *(formulated with $\rhd$ and $[1]$) and is closed under the rules of modus ponens and necessitation for the $[1]$-modality. Furthermore, T is supposed to prove all the theorems of* GL *for the $[\,]$-modality. If now $T \vdash [\,]\langle 1\rangle\top$, and $T \vdash A \rhd B \rightarrow [1](A \rightarrow B \vee \Diamond B)$, then $T \vdash [1][1]\bot$.*

[2] Or "the consistency of the existence" for that matter.

Proof. By standard **IL**-reasoning we see that $T \vdash \top \rhd [1]\bot$. As by assumption $T \vdash A \rhd B \to [1](A \to B \vee \Diamond B)$, we get that

$$T \vdash [1]([1]\bot \vee \langle\ \rangle[1]\bot). \quad (+)$$

Also by assumption $T \vdash [\ ]\langle 1\rangle\top$, and thus also $T \vdash [1][\ ]\langle 1\rangle\top$. Combining the latter with (+) we get $T \vdash [1][1]\bot$. ⊣

Note that we can sharpen Lemma 4.8 and see that $T \vdash [1]\bot$ if we also have that $T \vdash [\ ]A \to [1]A$. The formula $[\ ]A \to [1]A$ is commonly added as an axiom in the realm of graded provability logics where the [1] modality stands for a stronger provability notion than the [ ] modality. See for example [1].

5 Further research

Really, what we have seen in the above sections is but a first exploration in the application of interpretability logics in the realm of set theory. And in a sense the results presented are of a negative flavor: the set-theoretical interpretations proposed do not interact well with interpretability logics. However a bimodal logic for the κ-λ interpretability principles could be an interesting structure.

In Section 3 we concluded that the $\rhd$ modality did not essentially add any expressivity. In Section 4 we saw that the $\rhd$ modality could be replaced by introducing an additional unary modality thus working with a bi-modal logic. Both versions do not employ yet the full strength and expressibility of interpretability logics. In Appendix D of [15] definitions are provided on how to interpret the $\rhd$ and the $\Box$ in models of arithmetic. One could consider proceeding along these lines in the realm of set theory. In this final section we mention some other directions that can be pursued.

5.1 Using transitive models of set theory

This would be a generalization of Solovay's second set-theoretical result mentioned in [2], Chapter 13, Theorem 1. Let us rephrase the theorem here. To this end we first give two definitions.

Definition 5.1. A finite prewellordering is a frame $\langle W, R\rangle$ where W is finite and R is a transitive and irreflexive relation such that for every w, x, y in W, if wRx, then either wRy or yRx.

Definition 5.2. The logic **I** is defined by adding to GL the following axiom scheme.

$$\Box(\Box A \to \Box B) \vee \Box(\Box B \to \boxdot A)$$

Solovay's theorem on transitive models in set theory now reads as follows.

Theorem 5.3. *Let A be a modal sentence. Let $*$ be a realization that maps $(\Box A)^*$ to the sentence of the language of set theory that formalizes "A^* holds in all transitive models of* ZF*". Then (A), (B), and (C) are equivalent.*

(A) For all $*$, $\mathsf{ZF} \vdash A^*$.

(B) A is valid in all finite prewellorderings.

(C) $\mathbf{I} \vdash A$.

We remark that the S_x-relations really give us more expressive power on finite pre-wellorderings. We could also consider a combination of this with the ideas explored in this paper, like in the following translation.

$A \rhd B :=$ "for all V_κ where A holds, either B holds, or
we can find a transitive model of B''

5.2 Using large cardinals and forcing extensions

The next natural thing to consider would be to read $A \rhd B$ as follows.

$A \rhd B \to$ "In every V_κ where A holds, there is a forcing extension where B holds"

The spirit of forcing comes a lot closer to that of interpretability than the candidate notions we have considered before. To get something new, one really should combine forcing with some other notion as suggested here. This is because the logic of forcing is determined in [7] and [8].

5.3 Large cardinals and arithmetized interpretability statements

We conclude the paper with a question that is somewhat out of the scope of this paper. However it relates questions in interpretability to large cardinals in another way and is interesting on its own.

Question 5.4. Can we find a natural large cardinal notion and two natural (preferably arithmetical) theories whose arithmetized interpretability statements are dependent on the existence of these large cardinals?

Acknowledgement

I would like to thank Benedikt Löwe for useful comments and answers to questions on set-theoretical issues. Notwithstanding his help, any possible inadequacies or errors in the current document are the fault of the author alone.

References

[1] L.D. Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys* 60(2): 197-268 2005.

[2] G. Boolos. The Logic of Provability. Cambridge University Press, ISBN 0-521-43342-8, 1993.

[3] P. J.Cohen. The Independence of the Continuum Hypothesis. *Proc. Nat. Acad. Sci. U. S. A.* 50, 1143-1148, 1963.

[4] P. J.Cohen. The Independence of the Continuum Hypothesis. II. *Proc. Nat. Acad. Sci. U. S. A.* 51, 105-110, 1964.

[5] K. Gödel. The Consistency of the Continuum-Hypothesis. Princeton, NJ: Princeton University Press, 1940.

[6] E. Goris, and J. J. Joosten. A new principle in the interpretability logic of all reasonable arithmetical theories. *Logic Journal of the IGPL*, Advanced Access published December 24, 2009, doi:10.1093/jigpal/jzp082, 2009.

[7] J. Hamkins. A simple maximality principle. *Journal of Symbolic Logic* 68(2):527–550, 2003.

[8] J. Hamkins and B. Löwe. The modal logic of forcing. *Transactions of the american mathematical society.* 360(4):1793–1817, 2008.

[9] J. J. Joosten and T. Icard. Provability and Interpretability Logics with Restricted Realizations. Submitted to *Notre Dame Journal of Formal Logic*, 2010.

[10] J.J. Joosten. *Intepretability Formalized*, Ph.D. thesis, Department of Philosophy, University of Utrecht, 2004.

[11] A. Kanamori. The Higher Infinite: Large Cardinals in Set Theory from their Beginnings. Second edition. Springer Monographs in Mathematics, Springer-Verlag Berlin Heidelberg. 2003

[12] K. Kunen. Set Theory: An Introduction to Independence Proofs. ISBN 978-0-444-85401-8. Amsterdam: North-Holland, 1980.

[13] C. F. Morgenstern. On the ordering of certain large cardinals. *Journal of Symbolic Logic* 44(4): 563-565, 1979.

[14] R. M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 28: 33-71, 1976.

[15] A. Visser. An overview of interpretability logic. *Mathematical Logic, Proceedings of the 1988 Heyting Conference*, Plenum Press: 307-359, 1997.