On Reflection Principles [★]

Peter Koellner

Harvard University, Massachusetts, USA, 02138, (617) 495-3970

Abstract

Gödel initiated the program of finding and justifying axioms that effect a significant reduction in incompleteness and he drew a fundamental distinction between intrinsic and extrinsic justifications. Reflection principles are the most promising candidates for new axioms that are intrinsically justified. Taking as our starting point Tait's work on general reflection principles, we prove a series of limitative results concerning this approach. These results collectively show that general reflection principles are either weak (in that they are consistent relative to the Erdös cardinal $\kappa(\omega)$) or inconsistent. The philosophical significance of these results is discussed.

Key words: Set Theory, Incompleteness, Large Cardinal Axioms, Reflection Principles
1991 MSC: 03E55

The incompleteness phenomenon in set theory provides us with natural statements of mathematics that cannot be settled on the basis of the standard axioms of set theory, ZFC. Two classic examples of such statements are PU (the statement that all projective sets admit of a projective uniformization) and CH (Cantor's continuum hypothesis). This leads to the program of seeking and justifying new axioms which settle the undecided statements. This program has both a mathematical component and a philosophical component. On the mathematical side, one must find axioms that are sufficiently strong to do the work. On the philosophical side, one must determine, first, what would count as a justification and, second, whether the axioms in question are justified. In this paper I will investigate these two aspects of one promising approach to justifying new axioms—the approach based on *reflection principles.* [1]

[★] I would like to thank Bill Tait for sparking my interest in the topic and for fruitful correspondence.

Email address: `koellner@fas.harvard.edu` (Peter Koellner).

[1] The result in Section 4 was proved in my dissertation (Koellner (2003)). The remaining results were proved in the Spring and Summer of 2007.

1 Introduction

The question motivating this work is: Can intrinsic justifications secure reflection principles that are sufficiently strong to effect a significant reduction in incompleteness? To render this question more precise I will discuss the notions of an "intrinsic justification" and a "significant reduction in incompleteness" (in the present section) and the notion of a "reflection principle" (in the next section).

Intrinsic versus Extrinsic Justifications. In his classic paper on the continuum problem (Gödel (1947) and Gödel (1964)) Gödel drew a fundamental distinction between intrinsic and extrinsic justifications. The discussion pertains to the iterative concept of set, that is, the concept of set "according to which a set is something obtainable from the integers (or some other well-defined objects) by iterated application of the operation "set of" " (Gödel (1964), p. 259). Gödel maintains that the "axioms of set theory [ZFC] by no means form a system closed in itself, but, quite on the contrary, the very concept of set on which they are based suggests their extension by new axioms which assert the existence of still further iterations of the operation "set of" " (260). He mentions as examples the axioms asserting the existence of inaccessible and Mahlo cardinals and maintains that "[t]hese axioms show clearly, not only that the axiomatic system of set theory as used today is incomplete, but also that it can be supplemented without arbitrariness by new axioms which only *unfold the content of the concept of set* as explained above" (260–261, my emphasis). Since Gödel later refers to such axioms as having "intrinsic necessary" I shall accordingly speak of such axioms being *intrinsically justified on the basis of the iterative concept of set.*[2]

The notion of an intrinsic justification on the basis of the iterative conception of set begs for sharpening. It appears that Gödel took it to be a fundamental form of justification, one that cannot be explained in more primitive terms. Nevertheless, one can explicate the idea of that Gödel appears to have in mind by comparing and contrasting it with other notions and by pointing to examples. One such point of contrast is that of an *extrinsic* justification, which Gödel introduces as follows: "[E]ven disregarding the intrinsic necessity of some new axiom, and even in case it has no intrinsic necessity at all, a probable decision about its truth is possible also in another way, namely, inductively by studying its "success". (261) Here by "success" Gödel means "fruitfulness in consequences, in particular "verifiable" consequences". In a famous passage he

[2] I shall also employ the more neutral notion of the iterative *conception* of set rather that the iterative *concept* of set since there is nothing in this discussion that rests on a robust form of conceptual realism such as that of Gödel. For discussions of Gödel's conceptual realism see Parsons (1995) and Martin (2005).

says: “There might exist axioms so abundant in their verifiable consequences, shedding so much light upon a whole field, and yielding such powerful methods for solving problems (and even solving them constructively, as far as that is possible) that, no matter whether or not they are intrinsically necessary, they would have to be accepted at least in the same sense as any well-established physical theory” (261).

Let us now consider some examples. Consider first the conception of natural number that underlies the system of Peano Arithmetic (PA). This conception of natural number not only justifies mathematical induction for the language of PA but for any extension of the language of PA that is meaningful. For example, if we extend the language of PA by adding the Tarski truth predicate and we extend the axioms of PA by adding the Tarski truth axioms, then, on the basis of the conception of natural number, we are justified in accepting instances of mathematical induction involving the truth predicate. In the resulting system one can prove Con(PA). This process can then be iterated. Moreover, there are other examples of axioms that are intrinsically justified on the basis of the conception of natural number; for example, the proof-theoretic reflection principles.[3] In contrast, the Π^0_1 statement $\mathrm{Con}(\mathrm{ZF}+\mathrm{AD})$ is undoubtedly *not* intrinsically justified on the basis of the conception of natural number; rather its justification flows from an intricate network of theorems in contemporary set theory.[4]

Consider next the iterative conception of set. As in the case of arithmetic this conception (arguably) intrinsically justifies instances of Replacement and Comprehension for certain extensions of the language of set theory. But there are richer principles that are (arguably) intrinsically justified on the basis of this conception, namely, the set-theoretic reflection principles. These principles assert (roughly) that any property that holds of V holds of some initial segment V_α. These principles yield inaccessible and Mahlo cardinals (and more) and quite likely underly Gödel’s claims in the above passage. I shall return to them in the next section.

One can also gain a sharper understanding of the notion of intrinsic justification by pointing to some of its properties. First, an intrinsically justified statement need not be self-evident, in part because the justification may be quite involved (for example, in the case of arithmetic, this would be the case with reflection principles at the level of some large ordinal approaching Γ_0), in part because it is possible that the underlying conception is problematic (as, for example, was the case with the Fregean conception of extension). On the other hand, the notion of intrinsic justification is intended to be more secure than mere “intrinsic plausibility” (in the sense of Parsons (2000)), in

[3] For more on this subject see Feferman (1991) and the references therein.
[4] See Section 3 of Koellner (2006) for more on this.

that whereas the latter merely adds credence, the former is intended to be definitive (modulo the tenability of the conception).

The question of how far intrinsic justifications can take us in securing new axioms is important for a number of reasons. First, intrinsic justifications would seem to be more secure than extrinsic justifications. Second, intrinsic justifications are more in line with traditional conceptions of mathematics. Indeed a number of people reject extrinsic justifications. This appears to be true of Gödel during his early development (as suggested in Section 1 of Koellner (2006)). And it is certainly true of a number of more recent thinkers. For example, in a discussion of extrinsic justifications, Tait writes:

> It is difficult to reconcile this with the iterative conception of the universe of sets we are discussing here. On the latter conception, the "intrinsic necessity" of an axiom arises from the fact that it expresses that some property possessed by the totality of ordinals is possessed by some ordinal. To introduce a new axiom as "true" on this conception because of its "success" would have no more justification than introducing in the study of Euclidean space points and lines at infinity because of their success. ... A "probable decision" about the truth of a proposition from the point of view of the iterative conception can only be a probable decision about its derivability from that conception. Otherwise, how can we know that a probable decision on the basis of success might not lead us to negate what we otherwise take to be an intrinsically necessary truth? (Tait (2001), reprinted in Tait (2005b), p. 284) [5]

In addition to being interesting because of its rejection of extrinsic justifications this passage is of interest since in it Tait takes intrinsic justifications (on the basis of the iterative conception of set) to be *exhausted* by reflection principles.

Now, I would not wish to defend this idea. But I do think that reflection principles are the best current candidates for axioms that admit such an intrinsic justification. In any case, reflection principles shall be my focus here (though in the final section of the paper I shall consider some alternatives). Our question then is whether intrinsic justifications (on the basis of the iterative conception of set) can secure reflection principles that effect a significant reduction in incompleteness.

[5] I am not convinced of this claim, in part because the analogous claim concerning the conception of natural number seems to be false. For example, the justification of the statement Con(ZF + AD) is not "derivable from" the conception of natural number and yet I do not think that we should be worried that it might "lead us to negate what we otherwise take to be an intrinsically necessary truth" with respect to the conception of natural number.

Significant Reduction in Incompleteness. I will be using the large cardinal hierarchy as a yardstick to measure the strength of reflection principles. As one climbs this hierarchy there is an increasing reduction in incompleteness. To render our question precise it is useful to fix on a specific target. A natural candidate is the axiom "for all sets X, $X^{\#}$ exists". One reason is that at this stage the basic results of descriptive set theory which can be established in ZFC lift to the next level. For example, this axiom implies $\underset{\sim}{\Pi}^1_1$-determinacy. A more theoretical reason is connected with absoluteness. Shoenfield showed (in ZFC) that the $\underset{\sim}{\Sigma}^1_2$-theory is "frozen" in that it cannot be altered by set forcing. This provides one with a method of converting consistency (established via forcing) into truth since if one shows a $\underset{\sim}{\Sigma}^1_2$-statement to be consistent by forcing then it must, by Shoenfield absoluteness, be true in V. Now one should like to be in this situation for the next pointclass, $\underset{\sim}{\Sigma}^1_3$. The axiom "for all sets X, $X^{\#}$ exists" is precisely the level at which this happens, in the following sense: By results of Woodin, Martin and Solovay, for axioms A which are invariant under set forcing, the theory ZFC + A will be generically absolute for $\underset{\sim}{\Sigma}^1_3$ iff it proves that for every X, $X^{\#}$ exists.[6] Since, without loss of generality, we will be able to assume that our axioms (reflection principles) have this feature of generic invariance, this is a reasonable target in terms of a reduction in incompleteness.

Our question then is whether reflection principles can effect a significant reduction in this sense. The first step would be to show that reflection principles imply $0^{\#}$. Now it is often maintained that large cardinals *in general* are justified in terms of reflection principles. Gödel appears to have held such a view:

> Generally I believe that, in the last analysis, every axiom of infinity should be derivable from the (extremely plausible) principle that V is indefinable, where definability is to be taken in [a] more and more generalized and idealized sense. (Wang (1977), p. 325; Wang (1996), p. 285)

Since the most natural way to assert that V is undefinable is via reflection principles and since to assert this in a "more and more generalized and idealized sense" is to move to languages of higher-order with higher-order parameters, Gödel is (arguably) espousing the view that higher-order reflection principles imply all large cardinal axioms. Others appear to say this directly.[7]

[6] See Woodin (1982).

[7] For example, in Martin and Steel (1989), the authors write: "We know of one proper extension of ZFC which is as well justified as ZFC itself, namely ZFC + 'ZFC is consistent'. Extrapolating wildly, we are led to *strong reflection principles*, also known as *large cardinal axioms* (One can fill in some intermediate steps.) These principles assert that certain properties of the universe V of all sets are shared by, or "reflect to", initial segments V_α of the cumulative hierarchy of sets" (72). However, although the authors appear to speak of reflection principles in our sense, they may have in mind the principles of Reinhardt, which, as will be discussed in the final

It is of interest then to determine whether such a view can be sharpened and upheld. We shall do this by taking Tait's work on general reflection principles (Tait (1990), Tait (1998a), Tait (1998b), and Tait (2005a)) as our starting point. Tait (2005a) says that his bottom-up approach may have the resources to lead beyond the $V=L$ barrier (p. 135). As we shall see, it follows from the limitative results below that current reflection principles do not imply $0^\#$ and hence cannot lead to a significant reduction in incompleteness (in the sense indicated above).

2 Reflection Principles

Reflection principles aim to articulate the informal idea that the height of the universe is "absolutely infinite" and hence cannot be "characterized from below". These principles assert that any statement true in V is true in some smaller V_α. Thus, for any φ one cannot define V as the collection which satisfies φ since there will be a proper initial segment V_α of V that satisfies φ. More formally, we shall write this as

$$V \models \varphi(A) \;\rightarrow\; \exists\alpha\; V_\alpha \models \varphi^\alpha(A^\alpha)$$

where $\varphi^\alpha(\,\cdot\,)$ is the result of relativizing the quantifiers of $\varphi(\,\cdot\,)$ to V_α and A^α is the result of relativizing the A to V_α. This schematic characterization of a reflection principle will be filled in as we proceed by (1) specifying the language and (2) specifying the nature of relativization.

For the time being our language will be the language of set theory extended with variables of all finite orders. We shall use $x, y, z, \ldots$ as variables of the first order and, for $m > 2$, $X^{(m)}$, $Y^{(m)}$, $Z^{(m)}$, $\ldots$ as variables of the m^{th} order.[8] Relative to V_α the first-order variables are interpreted to range over the elements of V_α and, for $m > 1$, the m^{th}-order variables are interpreted to range over the elements of an isomorphic version of $V_{\alpha+(m-1)}$. The reason for using an isomorphic copy of $V_{\alpha+(m-1)}$ and not $V_{\alpha+(m-1)}$ itself is that we wish to keep track of the set/class distinction in cases where a set and class have the same extension. For definiteness, in the case of $m = 2$ we shall use $\alpha \times V_{\alpha+1}$ and we shall interpret class membership "$y \in (\alpha, x)$" as $y \in x$. The cases where $m > 2$ are handled similarly.

We now turn to the nature of relativization. If $A^{(2)}$ is a second-order parameter over V_α, then the relativization of $A^{(2)}$ to V_β, written $A^{(2),\beta}$, is $A \cap V_\beta$.[9]

section, have a different form.

[8] When the order of a variable or parameter is clear from context we shall often drop the superscript for notational simplicity.

[9] More precisely, given our coding apparatus, $A^{(2),\beta}$ is really $\{x \in V_\beta \mid (\alpha, x) \in$

This is how $A^{(2)}$ looks from "the point of view of V_β". The relativization of higher-order parameters is defined inductively in the natural way: For $m > 1$, $A^{(m+1),\beta} = \{B^{(m),\beta} \mid B^{(m)} \in A^{(m+1)}\}$. The relativization of a higher-order formula φ to the level V_β is obtained by interpreting the first-order variables to range over V_β and, for $m > 1$, m^{th}-order variables to range over the elements of $V_{\beta+(m-1)}$.

With these specifications we now have a hierarchy of reflection principles. For reasons which will become apparent in the next section, we shall restrict ourselves for the time being to parameters of *second order*. Let us recall some basic facts: Let T be the theory ZFC $-$ Infinity $-$ Replacement. If one supplements T with the scheme for second-order reflection then the resulting theory implies Infinity and Replacement. Moreover, in the second-order language, one can express the statement "Ω is (strongly) inaccessible" (where 'Ω' designates the class of ordinals) and so, assuming second-order reflection, there exists κ such that κ is inaccessible (and thus $V_\kappa \models$ ZFC). We can then reflect the statement "Ω is an inaccessible greater than κ" to get an inaccessible above κ and, continuing in this manner, we obtain a proper class of inaccessibles. Thus, Ω is an inaccessible limit of inaccessibles and hence, by reflection, there exists κ which is an inaccessible limit of inaccessibles. Continuing in this manner one obtains the various orders of inaccessibles and Mahlos. One then obtains weakly compact cardinals and, moving up through the higher-order languages, one obtains the higher-order indescribable cardinals.

Before proceeding further let us examine some philosophical difficulties with the claim that higher-order reflection principles are intrinsically justified on the basis of the iterative conception of set.

The most basic difficulty involves the interpretation of higher-order quantification and turns on how one conceives of the "totality of sets". There are two conceptions of the "totality of sets"—the *actualist* conception and the *potentialist* conception. The actualist maintains that the totality of sets is a "completed totality", while the potentialist denies this. These two viewpoints face complementary difficulties in providing an intrinsic justification of higher-order reflection principles. On the actualist view one can refer to the totality of sets and thus one can articulate the idea that this totality cannot be described from below and hence satisfies the reflection principles. However, since there are no sets beyond this totality it is hard on this view to make sense of full higher-order quantification over the universe of sets. [10] On the potentialist

$A^{(2)}\}$, but we shall suppress such fine points in the future.

[10] One can, of course, simulate the construction of L over the universe for any ordinal that one can make sense of "internally" with the help of bootstrapping, but the resulting principles are quite weak. Now, there are some actualists who think that there is no problem in having full second-order quantification over the universe of sets. I am thinking here of advocates of the plural interpretation of second-order

view the closest one can come to speaking of the totality of sets is through speaking of some V_α. One can certainly make sense of higher-order quantification over V_α but now the difficulty lies in motivating and justifying reflection principles.

I know of two attempts to get around this difficulty. The first is the *theory of legitimate candidates* of Reinhardt. The second is the *bottom up approach* of Tait. I will say something about the former in the final section of this paper. Here I will concentrate on the latter.

Tait's approach takes its starting point in what he calls the *Cantorian Principle*, namely, the principle which asserts that if and initial segment $A \subseteq \Omega$ is a set then it has a strict upper bound $S(A) \in \Omega$. This is the principle which Cantor used to introduce (in a highly impredicative fashion) the totality of ordinals Ω. It follows (from the well-foundedness of the ordinals) that Ω is not a set. The problem with this principle is that it involves reference to the notion of "set" (in contrast to the notion of "class" or "inconsistent multiplicity") and this notion (and distinction) is far from clear. For this reason Tait replaces the principle with a hierarchy of *Relativized Cantorian Principles*. For a given condition C (called an *existence condition*) such a principle asserts that if an initial segment $A \subseteq \Omega_C$ satisfies C then it has a strict upper bound $S(A) \in \Omega_C$. It follows (from the well-foundedness of the ordinals) that Ω_C does not satisfy the condition C.

One can now obtain reflection principles by the appropriate choice of an existence condition. For example, suppose we wish to construct an ordinal Ω_C such that for a given second-order formula $\varphi(X)$, V_{Ω_C} satisfies φ-reflection, that is,

$$\forall X^{(2)} \subseteq V_{\Omega_C} \left(V_{\Omega_C} \models \varphi(X^{(2)}) \rightarrow \exists \alpha\, \varphi^{\alpha}(X^{(2),\alpha}) \right).$$

We simply take the condition C on initial segments $A \subseteq \Omega_C$ to be

$$\exists X^{(2)} \subseteq V_A \left(V_A \models \varphi(X^{(2)}) \wedge \forall \alpha \in A \, \neg \varphi^{\alpha}(X^{(2),\alpha}) \right).$$

Applying the Relativized Cantorian Principle to this condition and appealing again to the well-foundedness of the ordinals, we have that Ω_C does not satisfy C, that is, V_{Ω_C} satisfies φ-reflection. The trouble is that this method is too general. For example, suppose we wish to introduce $\Omega_{C'}$ such that $V_{\Omega_{C'}}$ satisfies that there is a φ-cardinal, where 'φ' could be anything, such as 'supercompact' or 'Reinhardt'. Let C' be the following condition on A: $V_A \not\models$ "There is a φ-cardinal". Applying the Relativized Cantorian Principle to this condition and appealing to the well-foundedness of the ordinals, we have that Ω_C does not satisfy C', that is, $V_{\Omega_C} \models$ "There is a φ-cardinal".

quantification. It would take us too far afield to consider this view in detail. In any case, the resulting principles fall under the limitative results to be presented below.

In short, whether the Relativized Cantorian Principle is intrinsically justified on the basis of the iterative conception of set will turn on the particular choice of the condition C that one considers. One would need to argue that in the case of conditions C that give rise to reflection principles the resulting instances of the Relativized Cantorian Principle are intrinsically justified. This is far from immediate. Furthermore, as we shall see, such a case cannot be made for a broad class of the instances that Tait considers since the resulting principles are inconsistent. Moreover, it is hard to see how one could draw the line in a principled fashion.

These philosophical difficulties show that at the moment we do not have a strong intrinsic justification of higher-order reflection principles. One would like, however, to say something more definitive. In the remainder of this paper I will prove a number of limitative results which collectively show that higher-order reflection principles are either weak or inconsistent.

3 Strong Reflection Principles

The next step is to allow parameters of third- and higher-order. Unfortunately, when one does this the resulting reflection principles are inconsistent, as noted by Reinhardt (1974). To see this let

$$A^{(3)} = \Big\{\{\xi \mid \xi < \alpha\}^{(2)} \mid \alpha \in \Omega\Big\}^{(3)}$$

and let $\varphi(A^{(3)})$ be the statement that each element of $A^{(3)}$ is bounded. This statement is true over V but for each $\alpha \in \Omega$ the reflected version of the statement, $\varphi^{\alpha}(A^{(3),\alpha})$, is false since $\{\xi \mid \xi < \alpha\}^{(2)} \in A^{(3),\alpha}$ is unbounded.

This counter-example and related counter-examples force one to forgo negative statements of the form $X^{(m)} \notin Y^{(m+1)}$ and $X^{(m)} \neq Y^{(m)}$ when $m \geq 2$. This leads to the following notions, due to Tait.

Definition 1 *A formula in the language of finite orders is positive iff it is build up by means of the operations* $\vee$, $\wedge$, $\forall$ *and* $\exists$ *from atoms of the form* $x = y$, $x \neq y$, $x \in y$, $x \notin y$, $x \in Y^{(2)}$, $x \notin Y^{(2)}$ *and* $X^{(m)} = X'^{(m)}$ *and* $X^{(m)} \in Y^{(m+1)}$, *where* $m \geq 2$.

Surprisingly, even when one restricts the language in this way, there are reflection principles which have significant strength.

Definition 2 *For* $0 < n < \omega$, $\Gamma^{(2)}_n$ *is the class of formulas of the form*

$$\forall X^{(2)}_1 \exists Y^{(k_1)}_1 \cdots \forall X^{(2)}_n \exists Y^{(k_n)}_n \, \varphi(X^{(2)}_1, Y^{(k_1)}_1, \ldots, X^{(2)}_n, Y^{(k_n)}_n, A^{(l_1)}, \ldots, A^{(l_{n'})})$$

where φ does not have quantifiers of second- or higher-order and $k_1, \ldots, k_n$, $l_1, \ldots, l_{n'}$ are natural numbers.

Definition 3 *For $0 < n < \omega$, $\Gamma^{(2)}_n$-reflection is the schema asserting that for each sentence $\varphi \in \Gamma^{(2)}_n$, if $V \models \varphi$ then there is a $\delta \in \Omega$ such that $V_\delta \models \varphi^\delta$*

Definition 4 (Baumgartner) *For $0 < n < \omega$, κ is n-ineffable iff for any $\langle K_{\alpha_1,\ldots,\alpha_n} \mid \alpha_1 < \cdots < \alpha_n < \kappa\rangle$ with $K_{\alpha_1,\ldots,\alpha_n} \subseteq \alpha_1$ for $\alpha_1 < \cdots < \alpha_n < \kappa$, there is an $X \subseteq \kappa$ and an S stationary in κ such that for $\beta_1 < \cdots < \beta_n$, all in S, $X \cap \beta_1 = K_{\beta_1,\ldots,\beta_n}$.*

Theorem 5 (Tait) *Suppose $n < \omega$ and $V_\kappa \models \Gamma^{(2)}_n$-reflection. Then κ is n-ineffable.*

Theorem 6 (Tait) *Suppose κ is a measurable cardinal. Then, for each $n < \omega$, $V_\kappa \models \Gamma^{(2)}_n$-reflection*

Two questions remain: (1) How strong is $\Gamma^{(2)}_n$-reflection? (2) Can one allow universal quantifiers of order greater than 2?

4 Consistency

Definition 7 *For $\alpha \geq \omega$ the Erdös cardinal $\kappa(\alpha)$ is the least κ such that $\kappa \to (\alpha)^{<\omega}_2$, that is, such that for each partition $P : [\kappa]^{<\omega} \to 2$ there is an $X \in [\kappa]^\alpha$ such that $\mathrm{Card}(P``[X]^n) = 1$ for all $n < \omega$, where $P``Y = \{P(a) \mid a \in Y\}$.*

Lemma 8 (Silver) *Assume $\alpha \geq \omega$ is a limit ordinal. Then the following are equivalent:*

(1) *$\kappa \to (\alpha)^{<\omega}_2$.*
(2) *For all structures M such that*
  (a) *$\mathrm{Card}(\mathscr{L}(\mathrm{M})) = \omega$ and*
  (b) *$\kappa \subseteq |M|$*
  there is an $X \in [\kappa]^\alpha$ which is a set of indiscernibles for M.

Theorem 9 *Assume $\kappa = \kappa(\omega)$ exists. Then there is a $\delta < \kappa$ such that V_δ satisfies $\Gamma^{(2)}_n$-reflection for all $n < \omega$.*

Proof. Our strategy is to use the Erdös cardinal to obtain a countable structure M and a non-trivial elementary embedding $j : M \to M$. We will then examine the critical point and show that it has the necessary reflection properties.

STEP 1: Consider the structure $N = \langle V_\kappa, \in, < \rangle$ where $<$ is a well-ordering of V_κ. Let $I' = \{\iota'_k\}$ be the indiscernibles of N given by Silver's lemma. Let $\mathrm{Hull}^N(I')$ be the Skolem hull of these indiscernibles and let

$$\pi : M \to \mathrm{Hull}^N(I') \subseteq V_k$$

be the inverse of the transitive collapse map. Let I be the image of I' under the transitive collapse. Notice that I is a set of indiscernibles for M and that by including the well-ordering $<$ in the structure we have ensured that these indiscernibles, which we will enumerate as $\{\iota_k\}$, obey the key properties (with respect to M) obeyed by the Silver indiscernibles (with respect to L).

Now let $\rho : I \to I$ be an order preserving map which moves the first indiscernible ι_0. This map uniquely extends to an elementary embedding $j : M \to M$ with $\mathrm{crit}(j) = \iota_0$. We aim to show that $V^M_{\iota_0}$ satisfies $\Gamma^{(2)}_n$-reflection for all n.

As motivation consider a formula $\varphi(A_1, \ldots, A_m) \in \Gamma_n$ and assume

$$V^M_{\iota_0} \models \varphi(A_1, \ldots, A_m).$$

This is (equivalent to) a first-order statement in M about the parameters $\iota_0, A_1, \ldots, A_m$. In what follows we will implicitly appeal to such equivalences in arguments using the elementarity of j to "shift" various facts. We would like to show that

$$V^M_{j(\iota_0)} \models \exists \alpha < j(\iota_0) \left(\varphi^\alpha(j(A_1)^\alpha, \ldots, j(A_m)^\alpha) \right)$$

since, by the elementarity of j (applied to the corresponding first-order statement), this would imply

$$V^M_{\iota_0} \models \exists \alpha < \iota_0 \left(\varphi^\alpha(A^\alpha_1, \ldots, A^\alpha_m) \right)$$

and we would be done. Now we have that

$$V^M_{j(\iota_0)} \models \varphi^{\iota_0}(A_1, \ldots, A_m).$$

So we would be done if $j(A)^{\iota_0} = A$. Unfortunately, this is not always true. For example, consider

$$A^{(3)} = \left\{ \{\xi \mid \xi < \alpha\}^{(2)} \mid \alpha < \iota_0 \right\}^{(3)}$$

and notice that $j(A^{(3)})^{\iota_0} \neq A^{(3)}$ since the former picks up $\{\xi \mid \xi < \iota_0\}^{(2)}$.

STEP 2: Fortunately, the following suffices:

Lemma 10 *Suppose $\varphi(A_1,\ldots,A_m)\in\Gamma^{(2)}_n$. If $V^M_{\iota_0}\models\varphi(A_1,\ldots,A_m)$ then $V^M_{\iota_0}\models\varphi(j(A_1)^{\iota_0},\ldots,j(A_m)^{\iota_0})$.*

Proof. Base Case: It will be convenient to separate the second-order variables from the higher order variables, so let us write $\varphi(A_1,\ldots,A_m)$ as

$$\varphi\Big(A^{(2)}_1,\ldots,A^{(2)}_j,B^{(n_{j+1})}_{j+1},\ldots,B^{(n_m)}_m\Big).$$

Suppose that $V^M_{\iota_0}\models\varphi\Big(A^{(2)}_1,\ldots,A^{(2)}_j,B^{(n_{j+1})}_{j+1},\ldots,B^{(n_m)}_m\Big)$. Since ι_0 is inaccessible in M,

$$C=\Big\{\alpha<\iota_0\;\Big|\;\Big\langle V^M_\alpha,\in,A^{(2),\alpha}_1,\ldots,A^{(2),\alpha}_j\Big\rangle\prec\Big\langle V^M_{\iota_0},\in,A^{(2)}_1,\ldots,A^{(2)}_j\Big\rangle\Big\}$$

is club in ι_0. We claim that for $\alpha\in C$,

$$V^M_\alpha\models\varphi\Big(A^{(2)}_1,\ldots,A^{(2)}_j,B^{(n_{j+1})}_{j+1},\ldots,B^{(n_m)}_m\Big).$$

The key point is this: Suppose that $A^{(2)}\in B^{(3)}$ or $B^{(k)}\in B^{(k+1)}$ are constituents of φ. If such a constituent is false (evaluated at $V^M_{\iota_0}$) then, since it occurs positively in φ, it does not contribute to the truth of φ (evaluated at $V^M_{\iota_0}$). If such a constituent is true (evaluated at $V^M_{\iota_0}$) then its reflected version to the α^{th}-level will be true (evaluated at V_α) for every $\alpha<\iota_0$.

Now the statement that φ reflects to each α in C is a first-order statement of M about the parameters $\iota_0,C,A_1,\ldots,A_j,B_{j+1},\ldots,B_m$. Thus, by the elementarity of j, the corresponding statement holds with respect to the image of these parameters, that is, the statement

$$\varphi(j(A_1),\ldots,j(A_j),j(B_{j+1}),\ldots,j(B_m))$$

reflects to the club of points $j(C)$ below $j(\iota_0)$. Since $j(C)\cap C=C$ and since C is unbounded in ι_0 and $j(C)$ is club, it follows that $\iota_0\in j(C)$, that is, the statement $\varphi(j(A_1),\ldots,j(A_j),\ldots,j(B_{j+1}),\ldots,j(B_m))$ reflects to ι_0.

Induction Step: Assume the lemma is true for $\psi\in\Gamma^{(2)}_n$. Our aim is to show that it is true for $\forall X^{(2)}\exists Y^{(k)}\psi(X^{(2)},Y^{(k)},\vec{A})$:

$$\begin{aligned}
&V^M_{\iota_0}\models\forall X^{(2)}\exists Y^{(k)}\,\psi(X^{(2)},Y^{(k)},\vec{A})\\
&\quad\leftrightarrow\forall B\subseteq V^M_{\iota_0}\Big[V^M_{\iota_0}\models\psi(B^{(2)},f(B)^{(k)},\vec{A})\Big]\\
&\quad\to\forall B\subseteq V^M_{\iota_0}\Big[V^M_{\iota_0}\models\psi(j(B)^{(2),\iota_0},j(f(B))^{(k),\iota_0},j(\vec{A})^{\iota_0})\Big]\\
&\quad\to\forall B\subseteq V^M_{\iota_0}\Big[V^M_{\iota_0}\models\psi(B^{(2)},f'(B)^{(k)},j(\vec{A})^{\iota_0})\Big]\\
&\quad\leftrightarrow V^M_{\iota_0}\models\forall X^{(2)}\exists Y^{(k)}\,\psi(X^{(2)},Y^{(k)},j(\vec{A})^{\iota_0}).
\end{aligned}$$

In the first equivalence $f : \mathscr{P}(V^M_{\iota_0}) \to \mathscr{P}^k(V^M_{\iota_0})$ is a Skolem function. The second implication holds by the induction hypothesis. The final equivalence is immediate. The third implication requires further comment: The first line of the implication provides us with a map

$$j(B)^{(2),\iota_0} \mapsto j(f(B))^{(k),\iota_0}$$

that is defined for each $B \subseteq V^M_{\iota_0}$. We would like to extract from this map a Skolem function for the quantifier alternation $\forall X^{(2)} \exists Y^{(k)}$. Fortunately, (and this is the key point), for each $B \subseteq V^M_{\iota_0}$, $j(B)^{(2),\iota_0} = B$. Thus,

$$\begin{aligned} f' : \mathscr{P}(V^M_{\iota_0}) &\to \mathscr{P}^k(V^M_{\iota_0}) \\ B &\mapsto j(f(B))^{(k),\iota_0} \end{aligned}$$

is the desired Skolem function. □

Step 3: We can now show that $V^M_{\iota_0} \models \Gamma^{(2)}_n$-reflection, for all $n < \omega$. Assume $V^M_{\iota_0} \models \varphi(A_1, \ldots, A_n)$. Then

$$V^M_{\iota_0} \models \varphi(j(A_1)^{\iota_0}, \ldots, j(A_n)^{\iota_0})$$

by the lemma. So

$$V^M_{j(\iota_0)} \models \exists \alpha < j(\iota_0) \left(\varphi^\alpha(j(A_1)^\alpha, \ldots, j(A_n)^\alpha) \right)$$

as witnessed by $\alpha = \iota_0$. Finally, $V^M_{\iota_0} \models \exists \alpha < \iota_0 \left(\varphi^\alpha(A^\alpha_1, \ldots, A^\alpha_n) \right)$, by the elementarity of j.

Finally, let $\delta = \pi(\iota_0)$. Applying π we have that $V_\delta \models \Gamma^{(2)}_n$-reflection for all $n < \omega$, which completes the proof. □

It is important to note that in the above proof we make key use of the fact that the higher-order universal quantifiers in a $\Gamma^{(2)}_n$-formula are second-order. The proof does not generalize to establish the consistency of $\Gamma^{(m)}_n$-reflection for $m > 2$ (contrary to what is suggested at the end of Tait (2005a)). The key step in which $m = 2$ is used is the step where we derive the choice function f' from f using j. Here it is crucial that the domain of the choice function be second-order since it is only for second-order parameters B that we can be guaranteed that $j(B)^{\iota_0} = B$ and hence that the derived choice function is total. To be more specific, the third implication in the induction step of the key lemma fails for $m = 3$ since in this case the domain of the derived choice function is $\{j(B)^{(3),\iota_0} \mid B \subseteq \mathscr{P}(V^M_{\iota_0})\}$ which is a proper subset of $\mathscr{P}^2(V^M_{\iota_0})$.

The question remains whether $\Gamma^{(m)}_n$-reflection for $m > 2$ is consistent relative to large cardinal axioms. There is a high-level reason for thinking that any reflection principle which is consistent relative to large cardinals is consistent relative to $\kappa(\omega)$. To see this recall that a canonical class of large cardinal axioms assert the existence of a nontrivial elementary embedding $j : V \to M$, where M is a transitive proper class and that as one increases the agreement between M and V the reflection properties of the critical point increase. The limiting case in which $M = V$ was shown to be inconsistent (with AC) by Kunen. If one drops AC and takes the embedding $j : V \to V$ one is in a situation that closely resembles our situation with $j : M \to M$. The difference, of course, is that we are dealing with a countable model M and not the entire universe. However, from the point of view of a consistency proof it would appear that whatever reflection is provable from $j : V \to V$ should also be provable from $j : M \to M$. Since reflection would appear to be an entirely internal matter, this is a reason for thinking that any conceivable reflection principle must have consistency strength below that of $\kappa(\omega)$.

5 Inconsistency

It turns out that the above consistency proof is optimal in that $\Gamma_1^{(3)}$-reflection is inconsistent (using a fourth-order parameter). The counterexample is best thought of in terms of a combinatorial consequence of $\Gamma_1^{(3)}$-reflection.

Definition 11 *Suppose that κ is an uncountable regular cardinal. Let $m \geq 2$. A 1-sequence$^{(m)}$ is a function $K : \kappa \to V_\kappa$ such that for all $\alpha < \kappa$, $K(\alpha) \subseteq V_{\alpha+(m-2)}$.*

Definition 12 *Suppose that κ is an uncountable regular cardinal, K is a 1-sequence$^{(m)}$, and $X^{(m)}$ is an m^{th} order class over V_κ. Then*

$$[K, X] = \{\alpha < \kappa \mid K(\alpha) = X^\alpha\}.$$

This is the set of points at which X "correctly guesses" K.

Definition 13 *Suppose κ is an uncountable regular cardinal and $m \geq 2$. Let $D \subseteq \kappa$.*

(1) *D is 0-stationary$^{(m)}$ iff D is stationary.*
(2) *D is $(n+1)$-stationary$^{(m)}$ iff for all 1-sequences$^{(m)}$ K there exists $X^{(m)}$ such that $[K, X] \cap D$ is n-stationary$^{(m)}$.*

We verify that (the relevant case of) one of Tait's results generalizes from the second-order to the third-order context.

Theorem 14 (Tait) *Suppose that V_κ satisfies $\Gamma_1^{(3)}$-reflection. Then κ is 1-stationary$^{(3)}$.*

Proof. Suppose, for contradiction, that κ is not 1-stationary$^{(3)}$. We claim that there exists $\varphi \in \Gamma_1^{(3)}$ and a fourth-order parameter $T^{(4)}$ such that $\varphi(T^{(4)})$ does not reflect, that is,

(1) $V_\kappa \models \varphi(T^{(4)})$ and
(2) for all $\beta < \kappa$, $V_\beta \not\models \varphi(T^{(4),\beta})$.

Let $K : \kappa \to V_\kappa$ be a 1-sequence$^{(3)}$ which is a counterexample to the 1-stationarity$^{(3)}$ of κ. For each $X^{(3)} \subseteq V_{\kappa+1}$ let C_X be a club such that

$$[K, X] \cap C_X = \varnothing.$$

Let

$$T^{(4)} = \{(K^{(2)}, X^{(3)}, C_X^{(2)}) \mid X^{(3)} \subseteq V_{\kappa+1}\}.$$

Let

$$\varphi(T^{(4)}) = \forall X^{(3)} \exists K^{(2)} \exists C^{(2)} \Big((K, X, C) \in T^{(4)} \wedge C \text{ is unbounded}\Big).$$

Notice that this is a $\Gamma_1^{(3)}$-statement about a fourth-order parameter. (We are implicitly using coding devices to collapse the existential quantifiers and code the heterogeneous relation $T^{(4)}$ as a fourth-order class $T^{*,(4)}$ in such a way that for all $\alpha \in \mathrm{Lim}$, $T^{(4),\alpha}$ is coded (in the same way) by $T^{*,(4),\alpha}$. See Tait (2005a) for details.)

Claim 15 *For each $X^{(3)} \subseteq V_{\kappa+1}$,*

(i) $V_\kappa \models$ "*C_X is unbounded*" *and*
(ii) *for all $\beta \in [K, X]$, $V_\beta \not\models$ "C_X^β is unbounded".*

Proof. (i) This is immediate since C_X is club in κ. (ii) Suppose, for contradiction, that $\beta \in [K, X]$ is such that $V_\beta \models$ "C_X^β is unbounded". Since C_X is club in κ this implies $\beta \in C_X \cap [K, X]$, which contradicts the fact that C_X was chosen to be such that $C_X \cap [K, X] = \varnothing$. □

It follows that $V_\kappa \models \varphi(T^{(4)})$, since for each $X^{(3)} \subseteq V_{\kappa+1}$ our fixed K and chosen C_X are witnesses. So we have proved (1).

It remains to prove (2), namely, $V_\beta \not\models \varphi^\beta(T^{(4),\beta})$, for all $\beta < \kappa$. Suppose, for contradiction, that $\beta < \kappa$ is such that $V_\beta \models \varphi^\beta(T^{(4),\beta})$, that is,

$$V_\beta \models \forall X^{(3)} \exists K^{(2)} \exists C^{(2)} \Big((K, X, C) \in T^{(4),\beta} \wedge C \text{ is unbounded}\Big).$$

For the particular choice $X = K(\beta)$, let $K_0^{(2)}$ and $C_0^{(2)}$ be such that

(*) $V_\beta \models (K_0^{(2)}, X, C_0^{(2)}) \in T^{(4),\beta} \wedge C_0^{(2)}$ is unbounded.

Since $(K_0, K(\beta), C_0) \in T^{(4),\beta}$ we have

(a) $K_0 = K^\beta$,
(b) $K(\beta) = X'^\beta$ for some $X' \subseteq V_{\kappa+1}$, and
(c) $C_0 = C'^\beta_{X'}$ (where this is our canonical choice for X'),

where $(K, X', C_{X'}) \in T^{(4)}$.

Now, in defining $T^{(4)}$ we chose C_X to be such that when $(K, X, C_X) \in T^{(4)}$ we have

$$[K, X] \cap C_X = \varnothing.$$

Thus, in particular, $[K, X'] \cap C_{X'} = \varnothing$, that is, for all $\alpha \in C_{X'}$, $K(\alpha) \neq X'^\alpha$. Now, by the Claim,

(d) $V_\kappa \models$ "$C_{X'}$ is unbounded" and
(e) $\forall \beta \in [K, X']\ V_\beta \not\models$ "$C'^\beta_{X'}$ is unbounded."

However, by (b), $\beta \in [K, X']$ and so, by (e),

$$V_\beta \not\models \text{"}C'^\beta_{X'} \text{ is unbounded."}$$

But by (c), $C_0 = C'^\beta_{X'}$, so

$$V_\beta \not\models \text{"}C_0 \text{ is unbounded,"}$$

which contradicts (*). □

Theorem 16 *$\Gamma_1^{(3)}$-reflection is inconsistent.*

Proof. Suppose, for contradiction, that V_κ satisfies $\Gamma_1^{(3)}$-reflection. Then, by the above theorem, κ is 1-stationary$^{(3)}$. We arrive at a counterexample by constructing a 1-sequence$^{(3)}$ which cannot be stationarily guessed. For $\alpha \in$ Lim, let

$$A_\alpha = \Big\{\{\xi \mid \xi < \gamma\} \mid \gamma < \alpha\Big\}.$$

Let $K_A : \kappa \to V_\kappa$ be such that $K(\alpha) = A_\alpha$ if $\alpha \in$ Lim and $K(\alpha) = \varnothing$ otherwise. We claim that for each $X^{(3)} \subseteq V_{\kappa+1}$, $[K_A, X] \cap$Lim contains at most one point. Suppose $X^\alpha = K_A(\alpha)$, where $\alpha \in$ Lim. If $\alpha' \in$ Lim is such that $\alpha' > \alpha$ and $X^{\alpha'} = K_A(\alpha')$ then there exists $Y^{(2)} \in X$ such that $Y^{(2),\alpha'} = \{\xi \mid \xi < \alpha\}$, in which case $Y^{(2),\alpha} = \{\xi \mid \xi < \alpha\} \in K_A(\alpha)$, which is a contradiction. Similarly, since $X^\alpha = K_A(\alpha)$, for each $\bar\alpha \in \text{Lim} \cap \alpha$ there exists $Y^{(2)} \in X$ such that $Y^{(2),a} = \{\xi \mid \xi < \bar\alpha\}$, in which case $Y^{(2),\bar\alpha} = \{\xi \mid \xi < \bar\alpha\} \in K_A(\bar\alpha)$ and hence $X^{\bar\alpha} \neq K_A(\bar\alpha)$. Hence $[K_A, X] \cap$ Lim contains at most one point.

It follows that $X^{(3)}$ cannot stationarily guess K_A since if $[K_A, X]$ is stationary then $[K, X] \cap \mathrm{Lim}$ is stationary, which is clearly impossible since it contains at most one point. $\square$

6 Dichotomy

The results of the previous two sections show that the reflection principles we have considered can be divided into two classes:

(1) Weak: $\Gamma^{(2)}_n$-reflection, for $n < \omega$.
(2) Inconsistent: $\Gamma^{(m)}_n$-reflection, for $m > 2$ and $n \geq 1$.

Since $\Gamma^{(3)}_1$ comes directly after $\bigcup_{n<\omega} \Gamma^{(2)}_n$, this classification is exhaustive and we have a *dichotomy theorem*: Reflection principles are either weak or inconsistent.

One response to this is that although we have a dichotomy with respect to our coarse classification there is still the possibility that a finer classification leads to reflection principles which are strong (and so fall outside the scope of our consistency theorem) and manage to skirt inconsistency. Indeed a finer classification can be readily obtained by looking not at full universal third- and higher-order quantification but various restricted forms of these. Setting aside the problem of motivating such a restriction in a principled way, in this section we prove a much sharper dichotomy theorem.

To isolate the necessary restriction on the domain of third-order universal quantification we begin by looking at a series of counter-examples and responses.

First Modification. Notice that although K_A (from the proof of the inconsistency theorem (Theorem 16)) cannot be stationarily guessed by any $X^{(3)} \subseteq V_{\kappa+1}$ it *can* be stationarily guessed by some $X^{(2)} \subseteq V_\kappa$; in fact, it is guessed *everywhere* by $\kappa^{(2)}$. This suggests modifying the notion of 1-stationarity$^{(3)}$ to allow guesses of *either* the form $X^{(2)} \subseteq V_\kappa$ or $X^{(3)} \subseteq V_{\kappa+1}$. However, there is a counterexample to this as well. For $\alpha \in \mathrm{Lim}$, let

$$B_\alpha = \Big\{ \{\xi \mid \gamma > \xi \leq \alpha\} \mid \gamma < \alpha \Big\}.$$

Let $K_B : \kappa \to V_\kappa$ be such that $K(\alpha) = B_\alpha$ if $\alpha \in \mathrm{Lim}$ and $K(\alpha) = \varnothing$ otherwise. For $X^{(2)} \subseteq V_\kappa$, $[K_B, X] \cap \mathrm{Lim} = \varnothing$. For $X^{(3)} \subseteq V_{\kappa+1}$, $[K_B, X] \cap \mathrm{Lim}$ can contain at most one point.

Second Modification. One response to the counterexample K_B is to restrict our attention to "full" 1-sequences$^{(3)}$, that is, 1-sequences$^{(3)}$ such that for each $\alpha \in \mathrm{Lim}$, $\min\{\mathrm{rank}(Y) \mid Y \in K(\alpha)\} < \alpha$. However, this also has a counterexample. For $\alpha \in \mathrm{Lim}$, let

$$C_\alpha = \Big\{\alpha - \{\xi\} \mid \xi < \alpha\Big\}.$$

Let $K_B : \kappa \to V_\kappa$ be such that $K(\alpha) = C_\alpha$ if $\alpha \in \mathrm{Lim}$ and $K(\alpha) = \varnothing$ otherwise. This is a full 1-sequence$^{(3)}$ such that for each $X^{(2)} \subseteq V_\kappa$, $[K_C, X] \cap \mathrm{Lim} = \varnothing$ and for each $X^{(3)} \subseteq V_{\kappa+1}$, $[K_C, X] \cap \mathrm{Lim}$ contains at most one point.

Third Modification. One response to the counterexample K_C is that although it is full and cannot be stationarily guessed by any $X^{(2)} \subseteq V_\kappa$ or any $X^{(3)} \subseteq V_\kappa$, it can be *recast* as a 1-sequence$^{(2)}$ which can be guessed by some $X^{(2)} \subseteq V_\kappa$. More generally, suppose $K : \kappa \to V_\kappa$ is a full 1-sequence$^{(3)}$ which is "narrow" in the sense that for each $\alpha \in \mathrm{Lim}$, $|K(\alpha)| \leq |\alpha|$. Relative to a fixed well-ordering let $\langle K(\alpha)_\xi \mid \xi < \alpha\rangle$ enumerate $K(\alpha)$. Now define the derived sequence $K' : \kappa \to V_\kappa$ to be such that $K'(\alpha) = \{\langle \xi, x\rangle \mid \xi < \alpha \wedge x \in K(\alpha)_\xi\}$ if $\alpha \in \mathrm{Lim}$ and $K'(\alpha) = \varnothing$ otherwise. The sequence K' codes K and can be stationarily guessed by some $X^{(2)} \subseteq V_\kappa$. So, to obtain strength, the above counter-examples suggest restricting attention to 1-sequences K which are "full" and "wide", that is, such that (1) for all $Y \in K(\alpha)$, $Y \subseteq \alpha \times V_\alpha$ and $\mathrm{dom}(Y) = \alpha$ and (2) $|K(\alpha)| > |\alpha|$. However, although the restriction to full and wide 1-sequences$^{(3)}$ rules out counterexamples like K_A, K_B, and K_C it does not rule out a simple "wide" version of K_C: For $\alpha \in \mathrm{Lim}$, $\xi < \alpha$, and $Y \subseteq \alpha$, let

$$Y_\xi = \Big\{\langle \gamma, i, j\rangle \mid (\gamma < \alpha) \wedge (i = 0 \leftrightarrow \gamma \neq \xi) \wedge (j = 0 \leftrightarrow \gamma \in Y)\Big\},$$

where i and j range over $\{0, 1\}$. (The role of the second coordinate is to code $\alpha - \{\xi\}$ and the role of the third coordinate is to code Y. Thus we are "tagging" $\alpha - \{\xi\}$ with Y.) For $\alpha \in \mathrm{Lim}$, let

$$C^*_\alpha = \{Y_\xi \mid \xi < \alpha \wedge Y \subseteq \alpha\}.$$

Finally, let $K_{C^*} : \kappa \to V_\kappa$ be such that $K_{C^*} = C^*_\alpha$ if $\alpha \in \mathrm{Lim}$ and $K_{C^*} = \varnothing$ otherwise. The idea is that $K_{C^*}(\alpha)$ has 2^α-many subcollections, each corresponding to a given $Y \subseteq \alpha$, each of size $|\alpha|$, and each such that something specific happens unboundedly often. This is a 1-full-wide sequence which cannot be stationarily guessed.

What all of the above counter-examples have in common is that they involve collections which lack a certain "closure". To make this precise we concentrate on third-order classes consisting of second-order class of ordinals. Over V_α such a class B is canonically coded by a collection of branches through 2^α, each branch being the characteristic function of a second-order class of ordinals. To say that a third-order class of second-order classes of ordinals is *closed* is simply to say that the associated collection of branches is closed in the standard topology (where basic open intervals have the form $O_s = \{t \in 2^\alpha \mid s \subseteq t\}$, where $s \in 2^{<\alpha}$.)

The above counter-examples all involve $K(\alpha)$ (for $\alpha \in \mathrm{Lim}$) which are not closed. In each case $K(\alpha)$ is uniformly defined for $\alpha \in \mathrm{Lim}$. Moreover, for each $\alpha \in \mathrm{Lim}$, $K(\alpha)$ is missing a continuity point that is inevitably added by the relativization of any $X^{(3)}$ which correctly guesses $K(\alpha')$ for some $\alpha' \in \mathrm{Lim}$ such that $\alpha' > \alpha$.

The above counter-examples all involve restrictions of the combinatorial notion of 1-stationarity$^{(3)}$ but our main interest is in restrictions of the notion of $\Gamma_1^{(3)}$-reflection. It will therefore be of importance to investigate the connection between the two.

Since we are interested in reflection principles extending second-order reflection principles we may assume that the height of the universe V_κ is Mahlo (or more) and that in any reflection argument we reflect to a level V_α where α is also Mahlo (or more). Since in this case $|V_\alpha| = |\alpha|$, we may, without loss of generality, concentrate on third-order quantifiers which range over $\mathscr{P}(\mathscr{P}(\alpha))$.

Let $X^{\mathrm{c}(3)}$ range over the closed third-order classes of second-order classes of ordinals. This range of quantification is uniformly defined with respect to V_α for each $\alpha \in \mathrm{Lim}$. More generally, let $X^{\Gamma(3)}$ range over the $\Gamma(3)$-third-order classes of second-order classes of ordinals, where $\Gamma(3)$ is a pointclass which is uniformly defined with respect to V_α for each $\alpha \in \mathrm{Lim}$. The pointclasses $\Gamma(3)$ that will be of particular importance for us are those in the *generalized Borel hierarchy* which starts with the closed third-order classes over 2^α and proceeds by iterating the operations of α-union and complementation.

Definition 17 *Suppose κ is regular and uncountable. Suppose $\Gamma(3)$ is a third-order pointclass as above. A 1-sequence$^{\Gamma(3)}$ is a function $K : \kappa \to V_\kappa$ such that there is a club C_K in κ such that for all $\alpha \in C_K$, $K(\alpha) \subseteq 2^\alpha$ is in $\Gamma(3)$. A cardinal κ is 1-stationary$^{\Gamma(3)}$ iff for all 1-sequences$^{\Gamma(3)}$ K there exists $X^{(3)}$ such that $[K, X]$ is stationary.*

Definition 18 *Suppose $\Gamma(3)$ is a third-order pointclass as above. The collection of $\Gamma_n^{\Gamma(3)}$-formulas is defined exactly as before except that now all third-order universal quantifiers are replaced with $\forall X^{\Gamma(3)}$ and interpreted to range over the pointclass $\Gamma(3)$.*

Theorem 19 *Suppose V_κ satisfies $\Gamma_1^{\Gamma(3)}$-reflection. Then κ is 1-stationary$^{\Gamma(3)}$.*

Proof. The proof is a modification of the proof of Theorem 14. Suppose, for contradiction, that κ is not 1-stationary$^{\Gamma(3)}$. Let $K : \kappa \to V_\kappa$ be a 1-sequence$^{\Gamma(3)}$ which is a counter-example to the 1-stationarity$^{\Gamma(3)}$ of κ. For each $X^{\Gamma(3)} \in \mathscr{P}(\mathscr{P}(\kappa))$ let C_X be a club such that $[K, X] \cap C_X = \varnothing$. Let C_K be the club from the definition of a 1-sequence$^{\Gamma(3)}$. Let

$$T^{(4)} = \{(K^{(2)}, X^{\Gamma(3)}, C_X^{(2)}, C_K^{(2)}) \mid X^{\Gamma(3)} \in \mathscr{P}(\mathscr{P}(\kappa))\}.$$

Let

$$\varphi(T^{(4)}) = \forall X^{\Gamma(3)} \exists K^{(2)} \exists C^{(2)} \exists C'^{(2)} \Big((K, X, C, C') \in T^{(4)} \wedge \\ C_1 \text{ and } C_2 \text{ are unbounded}\Big).$$

This is a $\Gamma_1^{\Gamma(3)}$-formula. As before, for each $X^{\Gamma(3)} \in \mathscr{P}(\mathscr{P}(\kappa))$,

(1) $V_\kappa \models$ "C_X is unbounded" and
(2) for all $\beta \in [K, X]$, $V_\beta \not\models$ "C_X^β is unbounded".

It follows that

$$V_\kappa \models \varphi(T^{(4)}),$$

since for each $X^{\Gamma(3)} \in \mathscr{P}(\mathscr{P}(\kappa))$ our fixed K, C_K and chosen C_X are witnesses. It remains to prove that

$$V_\beta \not\models \varphi^\beta(T^{(4),\beta}),$$

for all $\beta < \kappa$. Suppose, for contradiction, that $\beta < \kappa$ is such that

$$V_\beta \models \varphi^\beta(T^{(4),\beta}),$$

that is,

$$V_\beta \models \forall X^{\Gamma(3)} \exists K^{(2)} \exists C^{(2)} \exists C'^{(2)} \Big((K, X, C, C') \in T^{(4),\beta} \wedge \\ C \text{ and } C' \text{ are unbounded}\Big).$$

Notice that $C_1 = C_K^\beta$. Moreover, since C_K^β is unbounded in β and since C_K is club, it follows that $\beta \in C_K$. Thus, $X = K(\beta)$ is in $\Gamma(3)$ and hence is a legitimate substituent for the universal quantifier $\forall X^{\Gamma(3)}$ in the formula displayed above. The rest of the proof is as before. □

Theorem 20 *Suppose* $\Gamma(3)$ *is a pointclass in the generalized Borel hierarchy that properly extends the closed classes. Then* $\Gamma_1^{\Gamma(3)}$*-reflection is inconsistent.*

Proof. This follows from the previous theorem in conjunction with the earlier counter-examples. □

Because of this one is essentially forced to pare the third-order quantifiers down to the closed sets. The question remains whether doing so leads to consistent reflection principles.

Theorem 21 *Assume $\kappa = \kappa(\omega)$ exists. Then there is a $\delta < \kappa$ such that V_δ satisfies $\Gamma^{\mathrm{c}(3)}_n$-reflection for all $n < \omega$.*

Proof. The proof is as before. The key point is that for $B^{\mathrm{c}(3)}$,

$$j(B)^{\mathrm{c}(3),\iota_0} = B^{\mathrm{c}(3)}.$$

Thus, we can extract the derived Skolem function in the third implication of the induction step as before. □

It remains to consider quantifiers of order beyond third-order. For the reasons noted earlier we may, without loss of generality, concentrate on n^{th}-order quantifiers which range over $\mathscr{P}^{n-1}(\alpha)$ when interpreted over V_α.

The earlier counter-examples easily generalize to higher-orders and enable us to isolate the appropriate notion of closure needed to avoid them. For illustrative purposes we concentrate on the fourth-order.

For $\alpha \in \mathrm{Lim}$ and for $\gamma < \alpha$ let T^α_γ be the tree consisting of the single branch $b \in 2^\alpha$ such that for all $\xi < \gamma + 1$, $b(\xi) = 1$ and for all $\xi \geq \gamma + 1$, $b(\xi) = 0$. For $\alpha \in \mathrm{Lim}$, let

$$D_\alpha = \{T^\alpha_\gamma \mid \gamma < \alpha\}.$$

Let $K_D : \kappa \to V_\kappa$ be such that $K(\alpha) = D_\alpha$ if $\alpha \in \mathrm{Lim}$ and $K_D(\alpha) = \varnothing$ otherwise. This is a 1-sequence$^{(4)}$ such that for each $X^{(4)} \subseteq V_{\kappa+2}$, $[K_D, X] \cap \mathrm{Lim}$ contains at most one point.

To rule out such counter-examples we must restrict to fourth-order classes that are "closed" in the following sense: Suppose $\langle T_\gamma \mid \gamma < \alpha \rangle$ is a sequence of trees such that each $T_\gamma \subseteq 2^{<\alpha}$. We say that the sequence is *increasing* if for each $\xi < \alpha$ there is an ordinal $f(\xi) < \alpha$ such that for all $\gamma_1, \gamma_2 \geq \eta$, we have $T_{\gamma_1} \restriction \xi = T_{\gamma_2} \restriction \xi$. In such a situation we say that the sequence *converges* to the *limit tree* $T = \bigcup_{\xi<\alpha} T_{f(\xi)}$. A fourth-order class $X^{(4)}$ over V_α is said to be *closed* iff it consists of closed third-order classes and is such that it contains the limit trees of every convergent subsequence of length α. Let $X^{\mathrm{c}(4)}$ range over the closed fourth-order classes. This is exactly the notion of closure which is needed to rule out the counter-examples. Moreover, the counter-example easily generalizes to higher-orders. We let $X^{\mathrm{c}(m)}$ range over the closed sets of m^{th}-order. As before there is a corresponding generalized Borel hierarchy at each level and the proof of Theorem 19 generalizes to show that for any level $\Gamma(m)$ of this hierarchy beyond $\mathrm{c}(m)$, $\Gamma_1^{\Gamma(m)}$-reflection is inconsistent for all $m > 2$.

It remains to see that for $m > 2$ and $n < \omega$, $\Gamma_n^{\mathrm{c}(m)}$-reflection is weak. Work over V_κ. The key point is that we can code trees $T \subsetneq 2^{<\kappa}$ with $b(T) \in 2^k$ in such a way that

$$\langle\, T_\alpha \mid \alpha < \kappa \,\rangle \text{ converges to } T \ \text{ iff } \ \langle\, b(T_\alpha) \mid \alpha < \kappa \,\rangle \text{ converges to } b(T).$$

It follows that each closed fourth-order class can be coded by a tree $T \subsetneq 2^{<\kappa}$. This, of course, generalized to higher-orders.

Theorem 22 *Assume $\kappa = \kappa(\omega)$ exists. Then there is a $\delta < \kappa$ such that V_δ satisfies $\Gamma_n^{\mathrm{c}(m)}$-reflection for all $m, n < \omega$.*

Proof. The proof is a modification of that of Theorem 9. The key change is in the inductive step.

Suppose $B^{\mathrm{c}(m)}$ is a closed m^{th}-order class over $V_{\iota_0}^M$ where $m > 2$. Let $T_B \subsetneq 2^{\iota_0}$ be a closed tree coding $B^{\mathrm{c}(m)}$. Let C be the closure of

$$\{\alpha < \iota_0 \mid M \models \text{“}\alpha \text{ is strongly inaccessible”}\}.$$

We may assume that the coding has been done in such a way that for all $\alpha \in C$, T_B^α codes B^α. Since T_B is closed,

$$j(T_B)^{\iota_0} = T_B.$$

By elementarity, for each $\alpha \in j(C)$,

$$j(T_B)^\alpha \text{ codes } j(B)^\alpha.$$

However, since $j(C)$ is club and since $j(C) \cap \iota_0 = C$ is unbounded in ι_0, $\iota_0 \in j(C)$. Thus,

$$j(T_B)^{\iota_0} = T_B \text{ codes } j(B)^{\iota_0},$$

which means that

$$j(B)^{\iota_0} = B.$$

This ensures that we can extract the derived choice function as before. □

Thus we have the following sharper dichotomy:

(1) Weak: $\Gamma_n^{\mathrm{c}(m)}$-reflection, for all $m > 2$ and $n < \omega$.
(2) Inconsistent: $\Gamma_n^{\Gamma(m)}$-reflection, for all $m > 2$ and $n \geq 1$ and any pointclass $\Gamma(m)$ containing the first level of the generalized Borel hierarchy beyond the closed sets.

7 Discussion

In the above results we have concentrated on quantifiers and parameters of finite order. However, one can make sense of quantifiers and parameters of transfinite orders and, for each ordinal α, one can define the notion of $\Gamma^{(\alpha)}_n$-reflection in the natural way. The results generalize to show (i) that to avoid inconsistency one must impose a closure constraint on the universal quantifiers of third- and higher-order and (ii) that resulting reflection principles are bounded below $\kappa(\omega)$.

One would like to conclude from this that "reflection principles" in general are weak. But absent a precise characterization of the notion of a reflection principle one cannot state, let alone prove, a limitative result to this effect. Moreover, it is hard to see how one could give an adequate precise characterization of the informal notion of a reflection principle since the notion appears to be inherently schematic and "indefinitely extendible" in the sense that any attempted precisification can be transcended by reflecting on reflection. However, our main limitative result is also schematic and the proof would appear to be able to track any degree of reflecting on reflection—the Erdös cardinal $\kappa(\omega)$ appears to be an impassable barrier as far as reflection is concerned. This is not a precise statement. But it leads to the following challenge: Formulate a strong reflection principle which is intrinsically justified on the iterative conception of set and which breaks the $\kappa(\omega)$ barrier.

It is natural at this point to think of the classic discussion of the justification of new axioms in Reinhardt (1974). It is important to note, however, that this discussion involves a very difference conception of set which has its roots in Reinhardt's dissertation (Reinhardt (1967)). This conception involves supplementing the iterative conception with what one might call the *theory of legitimate candidates*. On this view there are a number of "possible alternative interpretations of V", each of which has the form V_α. Let $V_{\alpha_0}, V_{\alpha_1}$ be a pair of such candidates, where $V_{\alpha_0} \subsetneq V_{\alpha_1}$. Reinhardt's basic method for obtaining strong principles is "to exploit the principle which says that mathematical truths should be necessary truths" and "[a]ccording to this principle, if the notion of possibility we have introduced is a good one, something true in one interpretation of V should be necessarily true, that is, true in all possible alternative interpretations of V" (Reinhardt (1967), p. 76). In particular, taking the language to be first-order with parameters from V_{α_0} one should have that for each φ and for each parameter $a \in V_{\alpha_0}$, $V_{\alpha_0} \models \varphi[a]$ iff $V_{\alpha_1} \models \varphi[a]$, that is, $V_{\alpha_0} \prec V_{\alpha_1}$. The next step is to enrich the language to second-order and allow second-order parameters. Reinhardt assumes that for each class $X \subset V_{\alpha_0}$ one can "reinterpret" the class over V_{α_1} as $j(X) \subset V_{\alpha_1}$ in such a way that $(V_{\alpha_0}, X) \prec (V_{\alpha_1}, j(X))$. Letting the interpretation function j be constant on elements of V_{α_1} this is equivalent to asserting that there is an elementary embedding $j : V_{\alpha_0+1} \to V_{\alpha_1+1}$, with critical point α_0, that is, it is equivalent to asserting that α_0 is a 1-extendible.

There are a number of difficulties with this approach—there are problems with the underlying conception and problems with the derivation of strong principles. One problem with the underlying conception is that the theory of potential candidates is difficult to defend. For example, since candidates which come later in the sequence will have greater closure properties than candidates which come earlier it is hard to defend the idea that they are both equally legitimate interpretations of V. Another difficulty is that underlying notion of mathematical modality would require considerable clarification and defense (especially in light of the fact that mathematics is traditionally thought to concern objects that necessarily exist).

But even if the underlying conception can be clarified in a satisfactory way, there are two problems with the derivation of strong principles. The first problem is what one might call the *problem of tracking*: In reinterpreting the class $X \subset V_{\alpha_0}$ as $j(X) \subset V_{\alpha_1}$ there must be some intensional notion at play. Now, one can certainly track *definable* classes by using their definitions. But Reinhardt wishes to shift *every* subset of V_{α_0} and for this he requires an exceedingly rich collection of intensional notions. Moreover, these intensional notions must be of a very special sort. For example, it would not do to associate to each set the concept of being that set since Reinhardt needs to "stretch" the classes. It is unclear that such a collection of concepts exists. Moreover, even if it did it would be a further step to assume that it gave rise to an elementary embedding of the required sort.

The second problem is what one might call the *problem of extendibility to inconsistency*: Even if one could provide and defend a theory of the required intensional objects it would appear that the theory would generalize and lead to inconsistency. In his dissertation Reinhardt did indeed think that the theory generalized: "[I]n order to extend [the above schema] to allow parameters of arbitrary (in the sense of V_2) order over V_0 we simply remove the restriction $X \subseteq V_0$." (Reinhardt (1967), p. 79). Here V_0 is our V_{α_1} and V_2 is some legitimate candidate V_{α_2} beyond V_{α_0} and V_{α_1}. The trouble is that when one generalizes in this way the result is a non-trivial elementary embedding $j : V \to V$ which Kunen showed to be inconsistent (with AC). (In fact, Kunen showed that even the existence of a non-trivial embedding $j : V_{\lambda+2} \to V_{\lambda+2}$ is inconsistent.) By the time he wrote his 1974 paper Reinhardt knew of this result. The point, however, is that the case Reinhardt makes for 1-extendibles appears to extend to a case for the inconsistent axiom. Hence, unless one can give principled reasons for blocking the extension, the case falters.

One can overcome both the problem of tracking and the problem of extendibility to consistency by restricting to classes which are definable with parameters.

Iterating this through the constructible universe built over a given legitimate candidate V_{α_0} one can make a case for the following axiom, which is more appropriately called an *extension principle*: For all γ there exist α_0 and α_1 and an elementary embedding $j : L(V_{\alpha_0}) \to L(V_{\alpha_1})$ where $\gamma < \alpha_0 < \alpha_1$. This axiom is consistent (relative to mild large cardinal assumptions). So by overcoming the problem of tracking in this way one also overcomes the problem of extendibility to inconsistency. Furthermore, the resulting axiom implies that $X^{\#}$ exists for all X and so freezes $\underset{\sim}{\Sigma}^1_3$. This still leaves us with the problem of defending the underlying conception. In Koellner (2003) I examined this conception and concluded on a skeptical note. In any case, although such an axiom might be intrinsically plausible on such an alternative conception, it is hard to see how it could be intrinsically justified solely on the basis of the iterative conception of set that we have been discussing.

There is another way in which one might try to justify strong principles resembling reflection principles. Up until now we have concentrated on principles which say that the *height* of the universe cannot be approximated from *below*. One might consider related principles which articulate the idea that the *width* of the universe cannot be approximated from *within*. On this approach to say that the universe cannot be approximated from within is to say that there is no "L-like model" which "approximates" or "covers" the universe. This general principle—*the principle of width reflection*—would then be rendered precise in terms of the various models occurring in inner model theory and their corresponding covering properties. For example, at the first stage one would simply take Gödel's constructible universe L as the approximating universe and as the notion of approximation one would take the notion involved in Jensen's original covering lemma, that is, to say that L covers V is to say that for every uncountable set of ordinals X there is an $Y \in L$ such that $|Y| = |X|$ and $X \subseteq Y$. The statement that L does not cover V implies, by the covering lemma, that $0^{\#}$ exists. A second application of width reflection would then lead to the existence of $0^{\#\#}$. In this way we proceed through the "sharp hierarchy" (using the same covering property in each application of with reflection) until we reach the Dodd-Jensen core model K. One more application of width reflection yields an inner model with a measurable cardinal. From this stage onward the covering property used in the applications of width reflection is necessarily weaker, by a result of Prikry. A basic consequence of current inner model theory (in particular, the core model induction) is that successive applications of *width reflection* ultimately imply PD and $\mathrm{AD}^{L(\mathbb{R})}$ and so definitely lead to a significant reduction in incompleteness. However, although the principle of width reflection may be intrinsically plausible it is hard to defend the idea that it is intrinsically justified on the basis of the iterative conception of set.

There may be other ways of intrinsically justifying principles which lead to a significant reduction in incompleteness. Gödel certainly believed that a more profound analysis of the concept of set (following the lines of Husserl's phenomenology) would lead to such principles. I am not optimistic but I do not wish to make a stronger claim than that.

Let me close by discussing some applications of the above limitative results. The first concerns inner model theory. In the approach discussed above one approximates the hypothesis that "there exists a measurable cardinal" from within via width reflection. One would like to approximate $0^{\#}$ from below in a similar fashion and the most natural way to do this is through height reflection. However, the results of this paper make this approach seem doubtful since $\kappa(\omega)$ appears to be out of reach of reflection. The second concerns intrinsic justifications. The inconsistency result shows that serious problems can arise even when one is embarked on the project of unfolding the content of a conception. It should give us pause in placing too much confidence in the security of intrinsic justifications. Third, the consistency result shows that intrinsic justifications, insofar as they are exhausted by the general reflection principles discussed above, will not take us very far. Finally, these results can be used to provide a rational reconstruction of Gödel's early view to the effect that $V = L$, PU, and CH are "absolutely undecidable". The idea is that if one has a conception of set theory which admits only intrinsic justifications and if one thinks that these are exhausted by reflection principles then the above results make a case for the claim that these statements really are "absolutely undecidable".[11] Fortunately, extrinsic justifications go a long way and I think that one can make a strong extrinsic case for $V \neq L$ and PU.[12] Whether CH is "absolutely undecidable" is, of course, a more delicate question.[13]

References

Feferman, S., 1991. Reflecting on incompleteness. Journal of Symbolic Logic 56, 1–49.

Gödel, K., 1947. What is Cantor's continuum problem? In: Gödel (1990). Oxford University Press, pp. 176–187.

Gödel, K., 1964. What is Cantor's continuum problem? In: Gödel (1990). Oxford University Press, pp. 254–270.

Gödel, K., 1990. Collected Works, Volume II: Publications 1938–1974. Oxford University Press, New York and Oxford.

Koellner, P., 2003. The search for new axioms. Ph.D. thesis, MIT.

Koellner, P., 2006. On the question of absolute undecidability. Philosophia Mathematica 14 (2), 153–188.

[11] See Sections 1 and 2 of Koellner (2006).

[12] See Section 3 of Koellner (2006) and the references therein.

[13] See Sections 4 and 5 of Koellner (2006) and Koellner and Woodin (2008) for more on this topic.

Koellner, P., Woodin, H., 2008. Incompatible Ω-complete theories. To appear in the Journal of Symbolic Logic.

Martin, D. A., 2005. Gödel's conceptual realism. Bull. Symbolic Logic 11 (2), 207–224.

Martin, D. A., Steel, J. R., January 1989. A proof of projective determinacy. Journal of the American Mathematical Society 2 (1), 71–125.

Parsons, C., 1995. Platonism and mathematical intuition in Kurt Gödel's thought. Bulletin of Symbolic Logic 1 (1), 44–74.

Parsons, C., 2000. Reason and intuition. Synthese 125, 299–315.

Reinhardt, W., 1967. Topics in the metamathematics of set theory. Ph.D. thesis, University of California, Berkeley.

Reinhardt, W., 1974. Remarks on reflection principles, large cardinals, and elementary embeddings. In: Proceedings of Symposia in Pure Mathematics. Vol. 10. pp. 189–205.

Tait, W. W., January 1990. The iterative hierarchy of sets. Iyyun 39, 65–79.

Tait, W. W., 1998a. Foundations of set theory. In: Dales, H., G., O. (Eds.), Truth in Mathematics. Oxford University Press, pp. 273–290.

Tait, W. W., 1998b. Zermelo on the concept of set and reflection principles. In: Schirn, M. (Ed.), Philosophy of Mathematics Today. Oxford: Clarendon Press, pp. 469–483.

Tait, W. W., 2001. Gödel's unpublished papers on foundations of mathematics. Philosophia Mathematica 9, 87–126.

Tait, W. W., 2005a. Constructing cardinals from below. In: Tait (2005b). Oxford University Press, pp. 133–154.

Tait, W. W., 2005b. The Provenance of Pure Reason: Essays in the Philosophy of Mathematics and Its History. Oxford University Press.

Wang, H., 1977. Large sets. In: Butts, Hintikka (Eds.), Logic, Foundations of Mathematics, and Computability Theory. D. Reidel Publishing Company, Dordrecht-Holland, pp. 309–333.

Wang, H., 1996. A Logical Journey: From Gödel to Philosophy. MIT Press.

Woodin, W. H., 1982. On the consistency strength of projective uniformization. In: Stern, J. (Ed.), Proceedings of the Herbrand Symposium. Logic Colloquium '81. North-Holland, Amsterdam, pp. 365–383.