THE REVIEW OF SYMBOLIC LOGIC
Volume 10, Number 4, December 2017

A STRONG REFLECTION PRINCIPLE

SAM ROBERTS

Abstract. This article introduces a new reflection principle. It is based on the idea that whatever is true in all entities of some kind is also true in a set-sized collection of them. Unlike standard reflection principles, it does not re-interpret parameters or predicates. This allows it to be both consistent in all higher-order languages and remarkably strong. For example, I show that in the language of second-order set theory with predicates for a satisfaction relation, it is consistent relative to the existence of a 2-extendible cardinal (Theorem 7.12) and implies the existence of a proper class of 1-extendible cardinals (Theorem 7.9).

§1. A new reflection principle. In this article, I introduce a new reflection principle. It is based on a very simple idea: whatever is true in all entities of some kind is also true in a set-sized collection of them.[1] More precisely:

$$\varphi \rightarrow \exists \mathcal{C} \varphi^{\mathcal{C}}, \tag{R}$$

where $\mathcal{C}$ is a set-sized collection of entities of some kind, φ only contains variables x, y, z etc. ranging over all entities of that kind, and $\varphi^{\mathcal{C}}$ is the result of replacing occurrences of quantifiers binding those variables—$\exists\mathsf{x}$, $\exists\mathsf{y}$, $\exists\mathsf{z}$ etc.— with quantifiers restricted to $\mathcal{C}$—$\exists\mathsf{x} \in \mathcal{C}$, $\exists\mathsf{y} \in \mathcal{C}$, $\exists\mathsf{z} \in \mathcal{C}$ etc. When φ contains variables ranging over multiple kinds of entities, there will be multiple set-sized collections in the consequent, one for each kind.

So, R merely restricts the ranges of quantifiers: it does not re-interpret them as ranging over entities of some other kind, nor does it re-interpret parameters or predicates.[2] In particular, $(\exists\mathsf{x} P(\mathsf{x}, \mathsf{y}))^{\mathcal{C}}$ is just $\exists\mathsf{x} \in \mathcal{C} P(\mathsf{x}, \mathsf{y})$. As we will see, this departure from standard reflection principles allows for instances of R that are both consistent in all higher-order languages and remarkably strong.

Here's the plan. In §2, I precisify R for the language of second-order set theory. In §3, I outline the well-known second-order reflection principle introduced by Bernays (1976) and isolate two implicit assumptions underlying it. I propose a generalisation of these assumptions for R, and show that imposing them yields a principle, R_2, equivalent to Bernays' (Theorem 7.3). In §4, I show that in an extension of the language of second-order set theory with predicates for a satisfaction relation, R_2 is consistent relative to the

Received: June 24, 2016.
2010 *Mathematics Subject Classification*: 03A05, 03E55, 03E65, 00A30.
Key words and phrases: set theory, reflection principles, large cardinals, intrinsic justification.

[1] I'm intentionally using the vague term "entity" here because I intend the idea to be as general as possible. For example, I take it to hold for objects like tables and sets, but also for pluralities and Fregean concepts. See §5 for discussion.

[2] In particular, it does not require that φ's parameters are contained in $\mathcal{C}$. Nonetheless, it is easy to see that this requirement is redundant. For, suppose that φ. Then, trivially, $\varphi \wedge \exists\vec{\mathsf{y}}(\vec{\mathsf{y}} = \vec{\mathsf{x}})$, where φ's free variables are among $\vec{\mathsf{x}}$. So, by R, $\exists\mathcal{C}(\varphi^{\mathcal{C}} \wedge \exists\vec{\mathsf{y}} \in \mathcal{C}(\vec{\mathsf{y}} = \vec{\mathsf{x}}))$ and thus $\exists\mathcal{C}(\varphi^{\mathcal{C}} \wedge \vec{\mathsf{x}} \in \mathcal{C})$.

doi:10.1017/S1755020317000223

existence of a 2-extendible cardinal (Theorem 7.12) and implies the existence of a proper class of 1-extendible cardinals (Theorem 7.9). The corresponding extension of Bernays' principle, in contrast, is inconsistent. In §5, I outline the main virtues of my principle, and in §6, I look at whether it is *intrinsically* justified. I argue that on our current understanding, it is at least as intrinsically justified as Bernays' principle. §7 is a technical appendix.

§2. The language of second-order set theory.

To obtain a completely precise principle from R, we need to specify (i) a class of formulas for which it is to hold, and (ii) exactly what set-sized collections of the relevant entities are and what it means for those entities to be elements of such collections—that is, we need to say what $\exists\mathcal{C}$ and $\mathsf{x} \in \mathcal{C}$ mean.

Consider the language of second-order set theory, $\mathcal{L}^2_\in$, in which there are first-order variables $x, y, z, \ldots$ ranging over sets and second-order variables $X, Y, Z, \ldots$, where $x = y$, $x \in y$, $x \in X$, and $X = Y$ are all taken to be well-formed. For readability, I will refer to whatever the second-order variables range over as *classes*. Moreover, I will assume that classes are extensional and obey a comprehension schema which says that any condition determines a class.[3] Formally:

$$\forall X, Y(\forall x(x \in X \leftrightarrow x \in Y) \rightarrow X = Y) \tag{ext}$$

$$\exists X \forall x(x \in X \leftrightarrow \varphi) \tag{comp}$$

for $\varphi \in \mathcal{L}^2_\in$ without X free. In this language, it is natural to take a set-sized collection of sets to simply be a set, and a set-sized collection of classes to be coded by a class of ordered pairs whose domain is a set. More precisely:

DEFINITION 2.1. *Say that a class X codes a set-sized collection of classes if there is a set x such that $\mathsf{dom}(X)$ is co-extensive with x (that is, $\forall y(y \in \mathsf{dom}(X) \leftrightarrow y \in x)$), where $\mathsf{dom}(X) = \{y : \exists z(\langle y, z\rangle \in X)\}$. Abbreviated: $\mathsf{sm}(X)$. Say that a class Y is in X if $\exists x \in \mathsf{dom}(X)(Y = X_x)$, where $X_x = \{y : \langle x, y\rangle \in X\}$.*[4] *Abbreviated: $Y \in X$.*

For the language of second-order set theory, then, we can precisify R as:

$$\varphi \rightarrow \exists x, X(\mathsf{sm}(X) \wedge \varphi^{x,X}), \tag{R^*_2}$$

where $\varphi \in \mathcal{L}^2_\in$, and $\varphi^{x,X}$ is the result of replacing occurrences of first-order quantifiers $\exists y$ in φ with $\exists y \in x$ and occurrences of second-order quantifiers $\exists Y$ with $\exists Y \in X$.

Unfortunately, R^*_2 is very weak. In particular, together with the axioms of second-order ZFC (ZFC2),[5] it is consistent relative to the existence of a strongly inaccessible cardinal (Theorem 7.1). Nonetheless, I will now show that it can be supplemented in a natural way to yield a remarkably strong principle.

[3] See §5 for discussion of these assumptions.

[4] So, X_x is the empty class when $x \notin \mathsf{dom}(X)$.

[5] I will take ZFC2 to be the theory in $\mathcal{L}^2_\in$ consisting of Extensionality, Infinity, Pairing, Union, Powerset, Foundation, Separation, Choice, ext, comp, the second-order axiom of Separation:

$$\forall X \forall x \exists y \forall z(z \in y \leftrightarrow z \in x \wedge z \in X)$$

and the second-order axiom of Replacement (which is stated similarly, in the obvious way). Given comp, the schemas of Separation and Replacement in $\mathcal{L}^2_\in$ follow from these axioms. ZFC is just ZFC2 without ext and comp, and with the axioms of Separation and Replacement swapped for their schemas in the language of first-order set theory.

§3. Bernays' reflection principle.

Bernays (1976) introduced what is now considered the paradigm of a second-order reflection principle.[6] It says that whatever is true in the sets and classes is true in some V_α and its subsets. Formally:

$$\varphi \rightarrow \exists\alpha\varphi^\alpha, \qquad (\mathsf{BR}_2)$$

where $\varphi \in \mathcal{L}^2_\in$, and φ^α is the result of replacing occurrences of first-order quantifiers $\exists x$ in φ with $\exists x \in V_\alpha$, second-order quantifiers $\exists X\psi(X)$ with $\exists y \subseteq V_\alpha\psi(y)$, and free second-order variables X with $X \cap V_\alpha$.

BR_2 is quite strong. Over $\mathsf{ZFC2}$, it implies that there are strongly inaccessible, Mahlo, weakly compact, and Π^1_n-indescribable cardinals.[7] It is thus much stronger than R^*_2. There are, however, two implicit assumptions BR_2 makes that R^*_2 does not. First, it assumes that the first-order domain of the reflecting structure is a V_α, rather than merely a set. Second, it assumes that the second-order domain of the reflecting structure contains *all* subsets of the first-order domain, rather than merely some subsets.[8]

Why are these assumptions permissible? It is natural to see them as claiming that certain fundamental features of the sets and classes are instantiated in the reflecting structure.[9] For the sets, the relevant feature is that they have the form of a V_α: that $V = \bigcup V_\alpha$. For the classes, the relevant feature is that any condition determines a class: that comp is true. Bernays opts to instantiate the second feature by requiring that the classes of the reflecting structure are exactly the subsets of its sets.[10] But there is a more general, and perhaps more natural, way to instantiate the second feature: namely, by requiring that any collection of sets in the reflecting structure determines a class. More precisely:

Definition 3.1. *Say that a class X is standard for a set x if for all subsets y of x there is some $z \in \mathsf{dom}(X)$ such that $X_z \cap x = y$. Abbreviated*: $\mathsf{st}(X, x)$.

When we add these assumptions to R^*_2, we get the principle: whatever is true in the sets and classes is true in some V_α and a set-sized collection of classes standard for V_α. Formally:

$$\varphi \rightarrow \exists\alpha, X(\mathsf{sm}(X) \wedge \mathsf{st}(X, V_\alpha) \wedge \varphi^{V_\alpha, X}) \qquad (\mathsf{R}_2)$$

for $\varphi \in \mathcal{L}^2_\in$.

So, we can see Bernays' principle as imposing further constraints on R_2: namely, that the classes in X are all co-extensive with subsets of V_α, and that class parameters are re-interpreted by their intersections with V_α. For the language of second-order set theory, it turns out that these further constraints are redundant: R_2 is equivalent to BR_2 (Theorem 7.3). However, once we move to extensions of that language, they have significant consequences.

§4. A strong reflection principle.

It is notoriously difficult to generalise Bernays' principle to extensions of $\mathcal{L}^2_\in$. To see this, consider a predicate P which applies to all and only those classes that are co-extensive with some set. Formally:

$$\forall X(P(X) \leftrightarrow \exists x\forall y(y \in x \leftrightarrow y \in X)).$$

[6] See Koellner (2009).

[7] See Kanamori (2003) §6.

[8] Without these assumptions, BR_2 would be as weak as R^*_2. Indeed, the resulting version of BR_2 would be equivalent to R^*_2. The proof would run along the same lines as the proof of Theorem 7.3.

[9] In other words, that the reflecting structure reflects these features in addition to φ.

[10] This in turn requires re-interpretation of class parameters, which Bernays does by taking their intersections with the sets.

Since BR_2 re-interprets class parameters by their intersections with V_α, the most natural way to re-interpret P is as the set of intersections with V_α of the classes satisfying it: that is, $\{Y \cap V_\alpha : P(Y)\}$. So, the most natural way to re-interpret occurrences of $P(X)$ is as $X \cap V_\alpha \in \{Y \cap V_\alpha : P(Y)\}$.[11,12]

But now note that the class of all sets, V, is not co-extensive with any set: in other words, $\neg P(V)$. So, if BR_2 held for "$\neg P(V)$", it would follow that there is some α such that $(\neg P(V))^\alpha$, which is to say $V \cap V_\alpha \notin \{Y \cap V_\alpha : P(Y)\}$. But that is false: when X is the class co-extensive with V_α, $P(X)$ and thus $X \cap V_\alpha \in \{Y \cap V_\alpha : P(Y)\}$, but $X \cap V_\alpha = V_\alpha = V \cap V_\alpha$.

In contrast, R_2 generalises straightforwardly and consistently to formulas involving P. Since R does not re-interpret parameters or predicates, $(\neg P(V))^{V_\alpha, X}$ is just $\neg P(V)$. Indeed, it is routine to modify the proof of Theorem 7.12 to show that the obvious generalisation of R_2 to the language of α^{th}-order set theory is consistent relative to the existence of large cardinals.[13] And in some extensions of $\mathcal{L}^2_\in$, it is remarkably strong. Let me now consider one such extension, where we add new predicates for a satisfaction relation.

Definition 4.1. *Let Var_1 be the set of $\mathcal{L}^2_\in$'s first-order variables, Var_2 the set of its second-order variables, and $Var = Var_1 \cup Var_2$.*[14] *Say that a class X is a variable assignment if* (*i*) $X \subseteq V \times V$, (*ii*) $Var_1 \subseteq \mathsf{dom}(X) \subseteq Var$, *and* (*iii*) $X \upharpoonright Var_1$ *is a function* (*where* $X \upharpoonright x = \{\langle y, z\rangle \in X : y \in x\}$). *If X is a variable assignment, let $X(x)$ be the unique y such that $\langle x, y\rangle \in X$ when $x \in Var_1$, and X_x otherwise.*[15] *In other words, X("y") is the set X assigns to "y" and X("Y") is the class it assigns to "Y".*

Definition 4.2. *Let $\mathcal{L}_S$ be $\mathcal{L}^2_\in$ extended with predicates $Sat(x, X)$* (*intended to express that the formula $x \in \mathcal{L}^2_\in$ is true on the variable assignment X*), *$As_0(X)$* (*intended to express that X is a variable assignment*), *$As_1(X, x, y)$* (*intended to express that y is the set assigned to the first-order variable x by X*), *and $As_2(X, x, Y)$* (*intended to express that Y is the class assigned to the second-order variable x by X*).

Definition 4.3. *Let* SAT *be the conjunction of the standard Tarski clauses for Sat,*[16] *and let* AS *be the conjunction of the following defining axioms for the other new predicates*:

(i) $\forall X(As_0(X) \leftrightarrow$ *X is a variable assignment*),
(ii) $\forall X(As_1(X, x, y) \leftrightarrow As_0(X) \wedge X(x) = y)$,
(iii) $\forall X(As_2(X, x, Y) \leftrightarrow As_0(X) \wedge X(x) = Y)$.

[11] Of course, we could re-interpret P as the set of subsets of V_α that satisfy its defining condition *in* V_α: that is, we could re-interpret it as $\{x \subseteq V_\alpha : V_\alpha \vDash \exists z \forall y (y \in z \leftrightarrow y \in x)\}$, which is just to say V_α! But this strategy is limited: there are predicates that do not have definitions in $\mathcal{L}^2_\in$. For example, a satisfaction predicate for $\mathcal{L}^2_\in$ will not have a definition in $\mathcal{L}^2_\in$, by Tarski's theorem on the undefinability of truth.

[12] See Tait (1998) and Koellner (2009) for discussion of this way of generalising Bernays' principle to extensions of $\mathcal{L}^2_\in$. See also Marshall R. (1989) for a less straightforward generalisation.

[13] In particular, it is consistent relative to the existence of an α-extendible cardinal.

[14] As usual, I will assume that each set has been coded as a recursive subset of ω.

[15] So, $X(x)$ is the empty class when $x \notin \mathsf{dom}(X)$.

[16] For example, one conjunct will say that for all variable assignments X:

$$Sat(\text{``}x \in Y\text{"}, X) \leftrightarrow X(\text{``}x\text{"}) \in X(\text{``}Y\text{"})$$

and one will say that for all variable assignments X:

$$Sat(\varphi \wedge \psi, X) \leftrightarrow Sat(\varphi, X) \wedge Sat(\psi, X).$$

Finally, let $\mathsf{ZFC2}_S$ *be* $\mathsf{ZFC2} + \mathsf{SAT} + \mathsf{AS}$*, with* comp *extended to* $\mathcal{L}_S$*, and let* R_S *be* R_2 *extended to* $\mathcal{L}_S$*.*

THEOREM. $\mathsf{ZFC2}_S + \mathsf{R}_S$ *implies that there is a proper class of 1-extendible cardinals, and thus that* $V \neq L$ *and* $AD^{L(\mathbb{R})}$*.*[17,18]

THEOREM. ZFC *implies that if there is a 2-extendible cardinal, then there is a model of* $\mathsf{ZFC2}_S + \mathsf{R}_S$*.*

§5. Virtues of R. Let me now outline the main virtues of R over its rivals. I will focus the comparison on Bernays' principle, though most of what I say also applies to other principles in the literature, like those in Reinhardt (1974), Marshall R. (1989), and Welch (forthcoming).

- *Generality.* We saw that R_2 easily generalises to extensions of $\mathcal{L}^2_\in$ like $\mathcal{L}_S$, whereas BR_2 does not.[19] But it also easily generalises to different interpretations of $\mathcal{L}^2_\in$, whereas BR_2 does not.

  So far, I have not provided an interpretation of $\mathcal{L}^2_\in$'s second-order variables: in other words, I have not said what classes are. Nonetheless, I have assumed that they satisfy ext and comp. And although these assumptions are plausible on some ways of thinking about classes, they are not on others.

  For example, they are plausible if we think of classes as pluralities.[20] But suppose we think of classes as properties, and read "$x \in X$" as "X applies to x".[21] The property X of being my favourite ordinal is distinct from the property Y of being the number 7. Although the number 7 *is* my favourite ordinal, it might not have been. In that case, X and Y would have applied to different things. So, X and Y happen to be co-extensive—they happen to apply to the same thingsbut they are distinct. In other words, ext is false for properties. Alternatively, suppose we think of classes as formulas in the language of first-order set theory, and read "$x \in X$" as "x satisfies the formula X in its one free variable". Then, many instances of comp would be false. For example, there would be no satisfaction class for the language

[17] These are Theorems 7.9 and 7.12 in the appendix. See Kanamori (2003) for all the undefined technical terms in this article and for a proof that $V \neq L$ follows from the existence of a 1-extendible cardinal. See chapter 22 in Foreman & Kanamori (2009) for a proof that $AD^{L(\mathbb{R})}$ follows from the existence of a proper class of 1-extendible cardinals.

[18] Since it implies that there are 1-extendible cardinals, R_S also implies that there are models of the principle S4 in Reinhardt (1974) and GRP in Welch (forthcoming).

[19] The principles in Marshall R. (1989) and Welch (forthcoming), however, do generalise naturally to such languages.

[20] On this account, X *are some things*, and "$x \in X$" is read as "it$_x$ is one of them$_X$". Trivially, some things are nothing over and above the individual things they comprise. So, when X and Y comprise the very same things—that is, when $\forall x(x \in X \leftrightarrow x \in Y)$—they must be identical. If X and Y are nothing over and above the things they comprise and they comprise the same things, then *nothing more* is required for X and Y to be identical. Similarly, since each individual φ exists trivially—that is, $\forall x(\varphi \to \exists y(y = x))$—there must be some things which are all and only the φs—that is, $\exists X \forall x(x \in X \leftrightarrow \varphi)$. If some things are nothing over and above the individual things they comprise and each individual φ exists, then *nothing more* is required for there to be some things that comprise the φs. See Boolos (1984), Uzquiano (2003), and Burgess (2004) for discussion.

[21] Properties have found use in metaphysics (see, for example, Williamson (2013)) and in the philosophy of set theory (see, for example, Reinhardt (1980)).

of first-order set theory by Tarski's theorem on the undefinability of truth, but it is a standard result that comp implies the existence of such a class (in the presence of the other axioms of $\mathsf{ZFC2}$).[22]

But Bernays' principle implies both ext and comp, and is thus incompatible with each of these interpretations of $\mathcal{L}^2_\in$. The reason is that failures of ext or comp would have to be reflected down to the subsets of some V_α by BR_2, but ext and comp always hold in those subsets.[23] In contrast, R is perfectly compatible with them. Because the second-order domain of the reflecting structure in R is a collection of classes and not a collection of sets, they need not satisfy ext. Moreover, when comp fails, the standardness assumption is no longer plausible and can simply be dropped.[24] To get the strength of R_S, we just need *some* interpretation of $\mathcal{L}^2_\in$ for which its assumptions are plausible, and we have that with the plural interpretation. In general, as long there is a serviceable notion of set-sized collection for some kind of entity, R can apply to them.

- *Uniformity.* BR_2 treats first- and second-order quantifiers in radically different ways: first-order quantifiers that range over sets continue to range over sets in the reflecting structure, whereas second-order quantifiers are re-interpreted to range over sets. In contrast, R applies in the same way to all quantifiers. A quantifier which ranges over entities of some kind continues to range over entities of the same kind in the reflecting structure, albeit a set-sized collection of them.
- *Simplicity.* BR_2 involves a complicated re-interpretation of second-order quantifiers by first-order quantifiers over subsets of V_α, and of class parameters by their intersection with V_α.[25] In contrast, R only restricts quantifiers: it does not re-interpret them as ranging over entities of some other kind, and it does not re-interpret parameters or predicates at all.[26]

§6. Is R_S intrinsically justified? In his influential article on reflection principles, Koellner concludes with the following challenge:

[22] Since there are distinct formulas that are satisfied by the same objects, ext will also fail on this account.

[23] More precisely, BR_2 implies:

$$\forall\alpha\varphi^\alpha \to \varphi \qquad (*)$$

for $\varphi \in \mathcal{L}^2_\in$ by contraposition. So, since it is trivial in ZFC to show that both ext and comp are true in all V_α, it follows from (*) that they are true simpliciter.

[24] Similarly, if the first-order quantifiers of the language range over nonsets, then the assumption that the first-order domain of the reflecting structure is a V_α can also be dropped.

[25] The re-interpretation of parameters and predicates is more complicated in the principles proposed by Reinhardt (1974), Welch (forthcoming), and Marshall R. (1989). Reinhardt's $\mathsf{S4}$ and Welch's GRP both postulate the existence of a function J that simultaneously re-interprets *all* subsets of the first-order domain of the reflecting structure as classes, where the only constraint on J is that it satisfy the principle. Marshall's $\mathsf{A3}$ is formulated in the language of third-order set theory, and re-interprets third-order classes relative to some other third-order class $\mathcal{X}$, where again the only constraint on $\mathcal{X}$ is that it satisfy the principle. For example, the predicate P from §4 gets re-interpreted relative to $\mathcal{X}$ as $\{X \cap V_\alpha : P(X) \wedge X \in \mathcal{X}\}$.

[26] It thus also avoids the explanatory burden of saying why parameters and predicates are re-interpreted the way they are. Koellner (2009) raises this problem for Reinhardt's $\mathsf{S5}$, and so by extension Welch's GRP, and calls it the 'problem of tracking'; and Linnebo (2007) raises it for Bernays' principle on a plural interpretation of $\mathcal{L}^2_\in$, and calls it the 'problem of plural parameters'. It is easy to see that the problem arises in general for Bernays' principle, and also for Marshall's $\mathsf{A3}$.

> 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. (p. 217, 2009)

Does R_S meet Koellner's challenge? Since it implies the existence of large cardinals far above $\kappa(\omega)$, this turns on whether it is intrinsically justified.

Usually, a statement is taken to be intrinsically justified if it follows (in some appropriate sense) from the *iterative conception of set*. According to that conception, the sets occur in an absolutely infinite series of stages: essentially, the V_αs. The standard arguments that BR_2 follows from this conception typically rely on the claim that the stages are absolutely infinite. For example, the most direct argument is that since the stages are absolutely infinite, whenever a claim φ is true, they must extend far enough to reach a stage at which it is true: that is, a V_α for which φ^α.[27] But, as (Koellner, 2009, p. 209) effectively points out, these arguments are prone to overgeneration. For example, they do not distinguish the consistent cases, where φ is a formula in the language of second-order set theory with class parameters, from the inconsistent cases, where it includes predicates in definitional expansions of that language, like the predicate P discussed in §4. In particular, since $\neg P(V)$ is true, it would seem that the stages should extend far enough to reach a stage at which $\neg P(V)$, which is impossible.[28] It is thus unsurprising that the arguments can easily be extended to R_S. And, as they stand, there is no principled reason to block those extensions. In general, it is unclear whether there is an interesting notion of intrinsic justification according to which Bernays' principle is justified but R_S is not.

Let me conclude with an argument that R_S is *not* intrinsically justified. The crucial thought is that the existence of classes does not follow from the iterative conception alone. It is, after all, a conception of *sets*, not of *classes*. But R_S implies that there are classes, since its consequent asserts that there is a class coding a set-sized collections of classes. So, R_S is not intrinsically justified.

It may at first seem like BR_2 is not subject to this problem, since its consequent merely asserts the existence of sets: $\exists\alpha\varphi^\alpha$ is a formula in the language of first-order set theory. However, BR_2 does imply that there are many and varied classes. As I mentioned in footnote 23, it implies that there are classes of some kind whenever every V_α thinks there are such classes. So, for example, it implies comp and that there is a class coding a

[27] See Burgess (2004) and Tait (2005) for more sophisticated arguments.

[28] I actually think there is a more fundamental problem with intrinsic justification. Even if we grant that a statement φ follows from the iterative conception, that would at most give us conditional evidence for φ: if the sets are as the iterative conception says they are, then φ is true of them. But, as Boolos points out:

> It does not follow that the iterative conception shows that the theorems of [...] Z^- [which is ZFC minus the axioms of extensionality, choice, and replacement] are *true*, for there is no reason to think that stages (whatever *they* might be) and sets are as the conception maintains, i.e., that the conception is correct about sets and stages. Certainly, if matters are as the conception has them, then Z^- is true, for, unexceptionably, it can be *deduced* from the iterative conception. However, no independent reason has been given to believe that sets and stages are as they are according to the iterative conception. (p. 6, 1989)

well-order of the sets.[29] If the argument shows that R_S is not intrinsically justified, then, it also shows that BR_2 is not intrinsically justified.

§7. Appendix.

THEOREM 7.1 (ZFC). *If there is a strongly inaccessible cardinal, then there is a model of* $\mathsf{ZFC2} + \mathsf{R}_2^*$.

Proof. Let κ be strongly inaccessible. I claim that $\langle V_\kappa, V_{\kappa+1}\rangle$ models $\mathsf{ZFC2} + \mathsf{R}_2^*$. Clearly, $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \mathsf{ZFC2}$. So it suffices to show that $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \mathsf{R}_2^*$. By the Lowenheim-Skolem theorem, there is a countable $M \subseteq V_\kappa$ and a countable $M' \subseteq V_{\kappa+1}$ such that $\langle M, M'\rangle$ is an elementary substructure of $\langle V_\kappa, V_{\kappa+1}\rangle$. Since M is countable, it will be in V_κ. Moreover, M' can be coded as a set-sized collection of classes X in $\langle V_\kappa, V_{\kappa+1}\rangle$. For example, let $\langle X_n : n < \omega\rangle$ enumerate the elements of M' and let $X = \{\langle n, x\rangle : x \in X_n\} \subseteq V_\kappa$. It is easy to see that $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \mathsf{sm}(X)$. A simple induction then shows that for $\vec{y} \in M$ and $\vec{Y} \in X$:

$$\langle V_\kappa, V_{\kappa+1}\rangle \vDash \varphi^{M,X} \leftrightarrow \langle M, M'\rangle \vDash \varphi,$$

where $\varphi \in \mathcal{L}^2_\in$ with free variables among $\vec{y}, \vec{Y}$. Now, suppose $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \varphi(\vec{y}, \vec{Y})$. Then we can pick M, M' as above but with $x \in M$ and $Y \in M'$. It follows that $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \varphi^{M,X}(\vec{y}, \vec{Y})$ with X as above, and so $\langle V_\kappa, V_{\kappa+1}\rangle \vDash \exists x, X(\mathsf{sm}(X) \wedge \varphi^{x,X}(\vec{y}, \vec{Y}))$.[30] □

LEMMA 7.2 ($\mathsf{ZFC2}$). *Suppose that X is standard for V_α, and $\mathsf{ext}^{V_\alpha,X}$. Then, for each $y \subseteq V_\alpha$, there is a unique $Y \in X$ such that $Y \cap V_\alpha = y$. By* comp*, let J be the class with $\mathsf{dom}(J) = V_{\alpha+1}$ such that J_y is that unique class for each $y \subseteq V_\alpha$. Then, the identity function on V_α together with J on $V_{\alpha+1}$ give an isomorphism between $\langle V_\alpha, V_{\alpha+1}\rangle$ and V_α, X.*

THEOREM 7.3 ($\mathsf{ZFC2}$). R_2 *and* BR_2 *are equivalent.*

Proof. $\mathsf{BR}_2 \Rightarrow \mathsf{R}_2$. Suppose $\varphi(x, Y)$. Applying BR_2 to $\varphi(x, Y)$ plus the claim that x exists, we get an V_α for which $\varphi^\alpha(x, Y \cap V_\alpha)$ and $x \in V_\alpha$. Now, let X be a set-sized collection of sets such that $\mathsf{dom}(X) = V_{\alpha+1}$, $X_{Y\cap V_\alpha} = Y$, and X_y is the class co-extensive with y for all $y \subseteq V_\alpha$ distinct from $Y \cap V_\alpha$. It is easy to see that V_α, X satisfies ext. It then follows immediately from Lemma 1 that $\varphi^{V_\alpha,X}(x, Y)$, since $j(Y \cap V_\alpha) = Y$, where j is the relevant isomorphism.

$\mathsf{R}_2 \Rightarrow \mathsf{BR}_2$. Suppose $\varphi(x, Y)$. Applying R_2 to $\varphi(x, Y)$, ext, and the claim that x and Y exist, we get an X standard for some V_α such that $\varphi^{V_\alpha,X}(x, Y)$, $x \in V_\alpha$, and $Y \in X$. So, it follows from Lemma 1 that $\varphi^\alpha(x, Y \cap V_\alpha)$, since $j(Y \cap V_\alpha) = Y$, where again j is the relevant isomorphism. □

[29] To see this, note that it follows from Choice that there is a subset of each limit V_λ coding a well-order of V_λ.

[30] To get a sharper bound on the strength of $\mathsf{ZFC2} + \mathsf{R}_2^*$, the Lowenheim-Skolem argument can be carried out in $\mathsf{ZFC2}$ supplemented with suitable choice principles. In particular, if we add the schema of collection:

$$\forall x \exists X \varphi(x, X) \to \exists X \forall x \varphi(x, X_x)$$

and the schema of ω-dependent choice:

$$\forall X \exists Y \varphi(X, Y) \to \exists X \forall n \varphi(X_n, X_{n+1})$$

for $\varphi \in \mathcal{L}^2_\in$, then R_2^* becomes provable. See Hamkins *et al.* (Accessed 24th June 2016) for further discussion.

DEFINITION 7.4. *A set a is a variable assignment over $\langle V_\alpha, V_{\alpha+1}\rangle$ if $a : Var \to V_{\alpha+1}$ and* $\mathsf{rng}(a \restriction Var_1) \subseteq V_\alpha$.

DEFINITION 7.5. *Say that an ordinal α is 1-extendible to β if $\alpha < \beta$ and there is an elementary embedding $j : \langle V_\alpha, V_{\alpha+1}\rangle \prec \langle V_\beta, V_{\beta+1}\rangle$ such that j is the identity on V_α. Say that α is 1-extendible if it is 1-extendible to some β.*[31]

The next definition is stated in $\mathcal{L}_S$.

DEFINITION 7.6. *Say that an ordinal α is 1-extendible to Ω if there is a class J such that for all variable assignments a over $\langle V_\alpha, V_{\alpha+1}\rangle$ and $\varphi \in \mathcal{L}^2_\in$:*

$$\langle V_\alpha, V_{\alpha+1}\rangle \vDash \varphi[a] \leftrightarrow Sat(\varphi, J^a),$$

where J^a is the variable assignment such that $J^a(x) = a(x)$ for $x \in Var_1$, and $J^a(x) = J_{a(x)}$ for $x \in Var_2$.

The next two easy lemmas, which I state without proof, highlight the connection between satisfaction classes and satisfaction in a structure, on the one hand, and the two notions of 1-extendibility, on the other. Let λ be a limit ordinal.

LEMMA 7.7 (ZFC). *Let $S \subseteq V_\lambda \times V_{\lambda+1}$ be such that $\langle V_\lambda, V_{\lambda+1}, S\rangle \vDash \mathsf{SAT}$. Let $A \subseteq V_\lambda$ be a variable assignment according to $\langle V_\lambda, V_{\lambda+1}\rangle$, and let a be the corresponding variable assignment over $\langle V_\lambda, V_{\lambda+1}\rangle$: that is, $\langle V_\lambda, V_{\lambda+1}\rangle \vDash A(x) = a(x)$, for all $x \in Var$. Then*:

$$\langle V_\lambda, V_{\lambda+1}\rangle \vDash \varphi[a] \quad \leftrightarrow \quad \langle \varphi, A\rangle \in S$$

for $\varphi \in \mathcal{L}^2_\in$.

LEMMA 7.8 (ZFC). *Let $S \subseteq V_\lambda \times V_{\lambda+1}$ be such that $\langle V_\lambda, V_{\lambda+1}, S\rangle \vDash \mathsf{SAT}$, and let $\alpha < \lambda$. Then*:

$$(\langle V_\lambda, V_{\lambda+1}, S\rangle \vDash \alpha \text{ is 1-extendible to } \Omega) \quad \leftrightarrow \quad \alpha \text{ is 1-extendible to } \lambda.$$

THEOREM 7.9 ($\mathsf{ZFC2}_S$). R_S *implies that there is a proper class of 1-extendible cardinals.*

Proof. I will show something stronger, namely that there is a proper class Y of ordinals which form a 1-extendible *chain*. That is, for any $\alpha, \beta \in Y$ with $\alpha < \beta$, α is 1-extendible to β.

Suppose that x is a set of ordinals that (1) form a 1-extendible chain and (2) are each 1-extendible to Ω. I will show that there is an ordinal outside x such that each ordinal in x is 1-extendible to it and which is itself 1-extendible to Ω. It will follow by a simple transfinite induction that there is a proper class satisfying (1) and (2).

Using R_S, we can get a V_α and a set-sized collection of classes X standard for V_α such that it is true in V_α, X that (i) x exists, (ii) each $\beta \in x$ is 1-extendible to Ω, (iii) ext, (iv) there is no greatest ordinal, and (v) $\mathsf{SAT} + \mathsf{AS}$. (iv) guarantees that α is a limit ordinal, and thus that Lemmas 2 and 3 are applicable.

By (iii) and Lemma 1, some J together with the identity on V_α gives an isomorphism from $\langle V_\alpha, V_{\alpha+1}\rangle$ to V_α, X. Trivially, they also give an isomorphism from $\langle V_\alpha, V_{\alpha+1}, J^{-1}[Sat]\rangle$ to V_α, X (where $J^{-1}[Sat] = \{\langle x, y\rangle \in V_\alpha \times V_{\alpha+1} : Sat(x, J_y)\}$). So, $\langle V_\alpha, V_{\alpha+1}, J^{-1}[Sat]\rangle \vDash \mathsf{SAT}$ by (v). By (i), x is in V_α. So, by (ii) and Lemma 3, each $\beta \in x$ is 1-extendible to α. It thus suffices to show that α is 1-extendible to Ω.

[31] See Kanamori (2003) §23.

By Lemma 2:

$$\langle V_\alpha, V_{\alpha+1}\rangle \vDash \varphi[a] \quad \leftrightarrow \quad \langle \varphi, A\rangle \in J^{-1}[Sat] \quad \leftrightarrow \quad Sat(\varphi, J_A),$$

where a is any variable assignment over $\langle V_\alpha, V_{\alpha+1}\rangle$, and $A \subseteq V_\alpha$ is the corresponding variable assignment in $\langle V_\alpha, V_{\alpha+1}\rangle$. To finish the proof, we just need to show that $J_A = J^a$.

First, since A is a variable assignment in $\langle V_\alpha, V_{\alpha+1}\rangle$, J_A is a variable assignment in V_α, X. So, because AS holds in V_α, X, we have $As_0(J_A)$, and thus that J_A is a variable assignment simpliciter. Similarly, since $A(x) = a(x)$ is true in $\langle V_\alpha, V_{\alpha+1}\rangle$, it follows that when $x \in Var_1$, $J_A(x) = a(x)$ is true in V_α, X (since $x, a(x) \in V_\alpha$). So, because AS is true in V_α, X, we have $As_1(J_A, x, a(x))$, and thus that $J_A(x) = a(x)$ simpliciter. An analogous argument shows that $J_A(x) = J_{a(x)}$ when $x \in Var_2$. □

This proof also suggests a way to obtain much of the strength of R_S without a satisfaction predicate. To see this, let $\mathcal{L}_Q$ be $\mathcal{L}^2_\in$ extended with a new predicate Q with the defining axiom:

$$\forall \vec{x}, \vec{X}(Q(\vec{x}, \vec{X}) \leftrightarrow \varphi), \qquad (\mathrm{Def}_Q)$$

where φ's free variables are among $\vec{x}, \vec{X}$. Let $\mathsf{ZFC2}_Q$ be $\mathsf{ZFC2}$ + Def_Q with comp extended to $\mathcal{L}_Q$, and let R_Q be R_2 extended to $\mathcal{L}_Q$.

Now, working in $\mathsf{ZFC2}_Q$ + R_Q, we can apply R_Q to ext + Def_Q we get a V_α and a set-sized collection of classes X standard for V_α for which $(\mathsf{ext} + \mathrm{Def}_Q)^{V_\alpha,X}$. Then, since $\mathsf{ext}^{V_\alpha,X}$, it follows from Lemma 1 that there is a J which gives an isomorphism between $\langle V_\alpha, V_{\alpha+1}\rangle$ and V_α, X. So, $\varphi^\alpha(\vec{x}, \vec{y})$ is equivalent to $\varphi^{V_\alpha,X}(\vec{x}, \vec{J_y})$. But, because $(\mathrm{Def}_Q)^{V_\alpha,X}$, it follows that $\varphi^{V_\alpha,X}(\vec{x}, \vec{J_y})$ is equivalent $Q(\vec{x}, \vec{J_y})$ and thus to $\varphi(\vec{x}, \vec{J_y})$ simpliciter. In other words, we have:

$$\forall \vec{x} \in V_\alpha \forall \vec{y} \subseteq V_\alpha(\varphi^\alpha(\vec{x}, \vec{y}) \quad \leftrightarrow \quad \varphi(\vec{x}, \vec{J_y})),$$

which is essentially just the instance for φ of the schema $\mathsf{S4}$ proposed in Reinhardt (1974) and GRP proposed in Welch (forthcoming). As Welch has shown, GRP already implies the existence of a proper class of measurable Woodin cardinals, and thus $V \neq L$ and $AD^{L(\mathbb{R})}$.

Definition 7.10. *Say that an ordinal α is 2-extendible to β if $\alpha < \beta$ and there is an elementary embedding $j : \langle V_\alpha, V_{\alpha+1}, V_{\alpha+2}\rangle \prec \langle V_\beta, V_{\beta+1}, V_{\beta+2}\rangle$ which is the identity on V_α. Say that α is a 2-extendible cardinal if it is 2-extendible to some β.*[32]

The following lemma, which I state without proof, is a simple consequence of this definition.

Lemma 7.11 (ZFC). *Suppose α is 2-extendible to β via j, and let $R_0, \ldots, R_n$ be relations over $\langle V_\alpha, V_{\alpha+1}\rangle$. Then, there are relations $R'_0, \ldots, R'_n$ over $\langle V_\beta, V_{\beta+1}\rangle$ such that*:

$$j : \langle V_\alpha, V_{\alpha+1}, R_0, \ldots, R_n\rangle \prec \langle V_\beta, V_{\beta+1}, R'_0, \ldots, R'_n\rangle.$$

Theorem 7.12 (ZFC). *If there is a 2-extendible cardinal, then there is a model of* $\mathsf{ZFC2_S}$ + $\mathsf{R_S}$.

Proof. Let α be 2-extendible to β via j, and let $M = \langle V_\alpha, V_{\alpha+1}, S, A_0, A_1, A_2\rangle \vDash \mathsf{SAT} + \mathsf{AS}$, where S, A_0, A_1, A_2 are relations over $\langle V_\alpha, V_{\alpha+1}\rangle$ interpreting Sat, As_0, As_1,

[32] Again, see Kanamori (2003) §23.

and As_2, respectively. By Lemma 4, there are S', A'_0, A'_1, A'_2 such that:

$$j : M \prec M' = \langle V_\beta, V_{\beta+1}, S', A'_0, A'_1, A'_2 \rangle.$$

I claim that $M \vDash \mathsf{ZFC2_S} + \mathsf{R_S}$. Since every 1-extendible cardinal is strongly inaccessible, it follows that $M \vDash \mathsf{ZFC2_S}$. It thus suffices to show that $M \vDash \mathsf{R_S}$.

So, suppose $M \vDash \varphi(\vec{x}, \vec{X})$. Trivially, j is an isomorphism between M and $M'' = \langle V_\alpha, rng(j \restriction V_{\alpha+1}), S', A'_0, A'_1, A'_2 \rangle$. Thus, $M'' \vDash \varphi(\vec{x}, j(\vec{X}))$. Moreover, $rng(j \restriction V_{\alpha+1})$ can be coded as a $Y \subseteq V_\beta$ which is set-sized and standard for V_α from the perspective of $\langle V_\beta, V_{\beta+1} \rangle$ because $\alpha < \beta$. It is thus straightforward to verify that:

$$M' \vDash \mathrm{st}(Y, V_\alpha) \wedge \mathrm{sm}(Y) \wedge \varphi^{V_\alpha, Y}(\vec{x}, j(\vec{X}))$$

and thus:

$$M \vDash \exists \alpha, Y(\mathrm{st}(Y, V_\alpha) \wedge \mathrm{sm}(Y) \wedge \varphi^{V_\alpha, Y}(x, X))$$

by elementarity.[33] □

It is routine to generalise this proof to show that higher-order versions of R_2 are consistent relative to the corresponding α-extendible cardinals.[34]

§8. Acknowledgments. Thanks to Leon Horsten, Øystein Linnebo, Pen Maddy, Rupert McCallum, Agustin Rayo, Gabriel Uzquiano, Philip Welch, two anonymous referees, and audiences at Bristol, Leeds, and Oslo for helpful comments and discussion.

BIBLIOGRAPHY

Bernays, P. (1976). On the problem of schemata of infinity in axiomatic set theory. In Müller, G. H., editor. *Sets and Classes: On the Work by Paul Bernays*. Studies in Logic and the Foundations of Mathematics, Vol. 84. North-Holland: Amsterdam, pp. 121–172.

Boolos, G. (1984). To be is to be a value of a variable (or to be some values of some variables). *Journal of Philosophy*, **81**(8), 430–449.

Boolos, G. (1989). Iteration again. *Philosophical Topics*, **17**, 5–21.

Burgess, J. P. (2004). E pluribus unum: Plural logic and set theory. *Philosophia Mathematica*, **12**(3), 193–221.

Foreman, M. & Kanamori, A. (2009). *Handbook of Set Theory*. Netherlands: Springer.

Hamkins, J., Gitman, V., & Johnstone, T. (2015). *Kelley-morse set theory and choice principles for classes*. Available at: http://boolesrings.org/victoriagitman/files/2015/01/kelleymorse2.pdf (accessed June 24, 2016).

Kanamori, A. (2003). *The Higher Infinite* (second edition). Berlin: Springer.

Koellner, P. (2009). On reflection principles. *Annals of Pure and Applied Logic*, **157**(2–3), 206–219.

Linnebo, O. (2007). Burgess on plural logic and set theory. *Philosophia Mathematica*, **15**, 79–93.

Linnebo, O. & Rayo, A. (2012). Hierarchies ontological and ideological. *Mind*, **121**(482), 269–308.

Marshall R., M. V. (1989). Higher order reflection principles. *Journal of Symbolic Logic*, **54**(2), 474–489.

[33] A sharper bound could be obtained by running the argument, with minor changes, using a subcompact cardinal.

[34] See Linnebo & Rayo (2012) for an interesting discussion of these languages.

Reinhardt, W. N. (1974). Remarks on reflection principles, large cardinals, and elementary embeddings. In Jech, T., editor. *Axiomatic Set Theory*. Providence: American Mathematical Society, pp. 189–205.

Reinhardt, W. N. (1980). Satisfaction definitions and axioms of infinity in a theory of properties with necessity operator. In *Mathematical logic in Latin America*. Studies in Logic and the Foundations of Mathematics, Vol. 99. Amsterdam: North-Holland, pp. 267–303.

Tait, W. W. (2003). Zermelo's conception of set theory and reflection principles. In Schirn, M., editor, *Philosophy of Mathematics Today*. New York: Oxford University Press, pp. 469–483.

Tait, W. (2005). Constructing cardinals from below. In *The Provenance of Pure Reason*. Logic and Computation in Philosophy. New York: Oxford University Press, pp. 133–154.

Uzquiano, G. (2003). Plural quantification and classes. *Philosophia Mathematica*, **11**(3), 67–81.

Welch, P. (forthcoming). Global reflection principles. In Sober, E., Niiniluoto, I., and Leitgeb, H., editors. *Proceedings of the CLMPS, Helsinki 2015*. London: College Publications.

Williamson, T. (2013). *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

DEPARTMENT OF PHILOSOPHY
IFIKK
UNIVERSITY OF OSLO
POSTBOKS 1020 BLINDERN
0315 OSLO, NORWAY
E-mail: sam.roberts@ifikk.uio.no
URL: http://samrroberts.net