THE ROLE OF THE FOUNDATION AXIOM IN THE KUNEN INCONSISTENCY

ALI SADEGH DAGHIGHI, MOHAMMAD GOLSHANI,
JOEL DAVID HAMKINS, AND EMIL JEŘÁBEK

Abstract. We consider the role of the foundation axiom in the Kunen inconsistency, the assertion that there is no nontrivial elementary embedding of the set-theoretic universe to itself. The truth or falsity of the Kunen assertion, we show, depends on the specific anti-foundational theory one adopts. On the one hand, it is relatively consistent with ZFC without foundation that the Kunen assertion fails, for there are models of $\mathrm{ZFC}^{-\mathrm{f}}$ in which there are definable nontrivial elementary embeddings $j\colon V \to V$. Indeed, in Boffa's anti-foundational theory BAFA, the Kunen assertion is outright refutable, for in this theory there are numerous nontrivial elementary embeddings $j\colon V \to V$ of the universe to itself. Meanwhile, on the other hand, Aczel's anti-foundational theory $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{AFA}$, as well as Scott's theory $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{SAFA}$ and other anti-foundational theories, continue to prove the Kunen assertion, ruling out the existence of a nontrivial elementary embedding $j\colon V \to V$.

In this article, we examine the role played by the axiom of foundation in the well-known Kunen inconsistency [Kun78], the theorem asserting that there is no nontrivial elementary embedding of the set-theoretic universe to itself. All the standard proofs of the Kunen inconsistency make use of the axiom of foundation (see [Kan04, HKP12]), and this use is essential, assuming that ZFC is consistent, because as we shall show there are models of $\mathrm{ZFC}^{-\mathrm{f}}$ that admit nontrivial elementary self-embeddings and even nontrivial definable automorphisms. Meanwhile, a fragment of the Kunen inconsistency survives without foundation

The second author's research has been supported by a grant from IPM (No. 91030417). The third author's research has been supported in part by Simons Foundation grant 209252, by PSC-CUNY grant 66563-00 44 and by grant 80209-06 20 from the CUNY Collaborative Incentive Award program. The fourth author has been supported by grant IAA100190902 of GA AV ČR, Center of Excellence CE-ITI under the grant P202/12/G061 of GA ČR, and RVO: 67985840. This inquiry grew out of a question posed by the first author on MathOverflow [Dag13] and the subsequent exchange posted by the third and fourth authors there.

as the claim in $\mathrm{ZFC}^{-\mathrm{f}}$ that there is no nontrivial elementary self-embedding of the class of well-founded sets. Nevertheless, some of the commonly considered anti-foundational theories, such as the Boffa theory BAFA, prove outright the existence of nontrivial automorphisms of the set-theoretic universe, thereby refuting the Kunen assertion in these theories. On the other hand, several other common anti-foundational theories, such as Aczel's anti-foundational theory $\mathrm{ZFC}^{-\mathrm{f}} + \mathrm{AFA}$ and Scott's theory $\mathrm{ZFC}^{-\mathrm{f}} + \mathrm{SAFA}$, reach the opposite conclusion by proving that there are no nontrivial elementary embeddings from the set-theoretic universe to itself. Our summary conclusion, therefore, is that the resolution of the Kunen inconsistency in set theory without foundation depends on the specific nature of one's anti-foundational stance.

1. Foundation and anti-foundation

The *axiom of foundation* is one of the standard axioms of the Zermelo–Fraenkel (ZFC) axiomatization of set theory, asserting that the set-membership relation $\in$ is well-founded, so that every nonempty set has an $\in$-minimal member. Thus, the foundation axiom allows for proofs by $\in$-induction, and indeed the axiom is equivalent to the $\in$-induction scheme. To give one immediate consequence, the axiom of foundation refutes $x \in x$ for any set x, for in this case $\{x\}$ would have no $\in$-minimal element; in particular, the axiom of foundation rules out the existence of *Quine atoms*, sets x for which $x = \{x\}$. Beyond this, the axiom of foundation girds much of the large-scale conceptual framework by which many set-theorists understand the cumulative universe of all sets. For example, one uses it to prove in ZF that every set appears at some level V_α of the von Neumann hierarchy, for if every element of a set x appears in some V_α, then x itself appears in $V_{\beta+1}$, where β is larger than the ranks of the elements of x, and so by $\in$-induction every set appears. Similarly, the axiom of foundation implies that every transitive set A is *rigid*, meaning that there is no nontrivial isomorphism of $\langle A, \in\rangle$ with itself, for if $\pi\colon A \to A$ is an isomorphism and π fixes every element $b \in a$ for some $a \in A$, then it follows easily that π must also fix a, and so π is the identity function by $\in$-induction.

Meanwhile, there are several commonly considered anti-foundational set theories, which we shall now briefly review. Most of these theories include the base theory we denote by $\mathrm{ZFC}^{-\mathrm{f}}$, which consists of all axioms of ZFC (including the collection and separation axiom schemes) except for the axiom of foundation.[1] Similarly, $\mathrm{GBC}^{-\mathrm{f}}$ denotes GBC

[1]Just as in the case of set theory $\mathrm{ZFC}^{-\mathrm{p}}$ without the power set axiom (see [GHJ]), one should take care to use the collection axiom plus separation in $\mathrm{ZFC}^{-\mathrm{f}}$,

without the axiom of foundation. In particular, please note that the theory $\mathrm{GBC}^{-\mathrm{f}}$ includes the global axiom of choice, which asserts that there is a class relation that is a well-ordering of all sets; this will be important in a few of our arguments. We shall sometimes also desire the global choice axiom in a context closer to $\mathrm{ZFC}^{-\mathrm{f}}$; in order to achieve this, we expand the language of set theory with a new function symbol C, and work in the theory we denote $\mathrm{ZFGC}^{-\mathrm{f}}$, which has all the $\mathrm{ZFC}^{-\mathrm{f}}$ axioms including instances of the collection and separation schemes for the expanded language, plus the assertion that $C\colon \mathrm{Ord} \to V$ is a bijection of the class of ordinals with the class of all sets.

Perhaps the most commonly used axiom in non-well-founded set theory is the *anti-foundation axiom*, denoted AFA, first investigated by Forti and Honsell [FH83] and then popularized by Aczel [Acz88] and Barwise and Moss [BM96]. The axiom has found numerous applications in computer science and formal semantics. In one formulation, AFA asserts that every directed graph $\langle A, e\rangle$ has a unique *decoration*, which is a mapping $a \mapsto f(a)$ of the elements of A to sets, such that

$$f(a) = \{f(b) \mid b\ e\ a\}$$

for every $a \in A$. This axiom therefore extends Mostowski's observation on well-founded relations to apply universally to all directed graphs. For example, by considering a graph with exactly one point and an edge from that point to itself, it follows that under AFA, there is a unique Quine atom $a = \{\, a \,\}$.

Since there are many equivalent formulations of the anti-foundation axiom, let us mention a few more of them. Define that a *partial (inverse) bisimulation* between directed graphs $\langle A, e\rangle$ and $\langle A', e'\rangle$ is a relation $\sim\ \subseteq A \times A'$ such that:[2]

(1) For every $x \in A$ and $x', y' \in A'$ such that $x \sim x'$ and $y'\ e'\ x'$, there exists a $y \in A$ such that $y\ e\ x$ and $y \sim y'$.
(2) For every $x' \in A'$ and $x, y \in A$ such that $x \sim x'$ and $y\ e\ x$, there exists a $y' \in A'$ such that $y'\ e'\ x'$ and $y \sim y'$.

rather than merely the replacement axiom, because these are no longer equivalent without foundation, although they are again equivalent without foundation in the presence of any of IE, BAFA, $V = \mathrm{WF}(t)$ for a set t, or global choice. To see the inequivalence, consider a model with ω many Quine atoms, but take only those sets in some $\mathrm{WF}(t)$ for any finite set t of such atoms, so that the full set of atoms never appears; the replacement axiom holds in this model, but not collection.

[2]These are technically *inverse* bisimulations, which have the appropriate directionality for expressing AFA using the $\in$ relation, as we have done here, but we shall drop this 'inverse' qualifier. Other accounts of AFA use the usual bisimulation directionality, but instead invert the set-membership relation to $\ni$, in effect inverting the direction of all the graphs here.

Such a bisimulation is *total* if $\mathrm{dom}(\sim) = A$ and $\mathrm{ran}(\sim) = A'$, in which case the two structures are called *bisimilar*. The anti-foundation axiom AFA is equivalent to the statement that every binary relation $\langle A, e\rangle$ is bisimilar with $\langle t, \in\rangle$ for a unique transitive set t. Thus, the axiom generalizes the situation in ZF, where Mostowski's argument shows that every well-founded directed graph is bisimilar to a unique transitive set; under AFA we get this for all directed graphs.

For another variant of AFA, define that an *accessible pointed graph* is a triple $\langle A, e, a\rangle$ where $e \subseteq A \times A$ and a is an element of A to which every element of A is related by the reflexive transitive closure of e. For example, the *canonical picture* of a set a is $\langle \mathrm{TC}(\{a\}), \in, a\rangle$, and this is an accessible pointed graph. Again generalizing the situation in ZF for well-founded relations, the anti-foundation axiom AFA is equivalent to the assertion that an accessible pointed graph $\langle A, e, a\rangle$ is isomorphic to a canonical picture of a set if and only if it is strongly extensional, meaning that every partial bisimulation from $\langle A, e\rangle$ to itself agrees with the identity relation on A.

Scott [Sco60] considered an anti-foundation axiom—following Aczel we shall denote it by SAFA—based on trees instead of arbitrary accessible pointed graphs. An accessible pointed graph $\langle A, e, a\rangle$ is a *tree* if every vertex has a unique e-path to a. For example, the *canonical tree picture* of a set a is the tree whose vertices are finite sequences $\langle x_0, \ldots, x_n\rangle$ where $x_0 = a$, and $x_{i+1} \in x_i$ for every $i < n$; we may think of the tree as growing downwards, so that child nodes correspond to elements. In ZF, one may easily prove that the canonical tree picture of a set has no nontrivial automorphisms, since every such automorphism would give rise to an automorphism of the transitive closure of the set. The Scott anti-foundation axiom SAFA generalizes this to the non-wellfounded realm by asserting that a tree is isomorphic to a canonical tree picture of a set if and only if it has no nontrivial automorphism. To illustrate, observe that under SAFA, there is precisely one Quine atom: there cannot be two, because if a and b were distinct Quine atoms, then the canonical tree picture of the doubleton set $\{\, a, b \,\}$ would have a nontrivial automorphism, swapping a and b; and there must be at least one, because the tree consisting of a single infinite descending chain is rigid, and so there must be sets a_n with $a_n = \{\, a_{n+1} \,\}$; but these must all be equal, because if $a_n \neq a_m$, then the tree picture of the doubleton $\{\, a_n, a_m \,\}$ consists of two descending chains joined at the top, and this tree is not rigid.

More generally, we say that two transitive sets s and t are isomorphic if there is an isomorphism of the structure $\langle s, \in\rangle$ with $\langle t, \in\rangle$. In the case of non-transitive sets, we say that two sets x and y are isomorphic, if

they can be placed into transitive sets having an isomorphism mapping x to y, or equivalently, if the transitive closures of $\{ x \}$ and $\{ y \}$ are isomorphic by an isomorphism mapping x to y. Thus, two sets are isomorphic if and only if their canonical pictures are isomorphic as accessible pointed graphs. Finsler [Fin26] developed non-well-founded set theory based on the informal principle that the universe of sets is maximal, subject to maintaining the axioms of extensionality and of *isomorphism extensionality*:

(IE) Isomorphic sets are equal.

This axiom implies immediately that there is at most one Quine atom, since any two would be isomorphic. Aczel [Acz88] formalized Finsler's idea in an axiom denoted FAFA, which asserts that an accessible pointed graph $G = \langle A, e, a \rangle$ is isomorphic to a canonical picture of a set if and only if it is extensional (that is, satisfies the axiom of extensionality) and $Gu \not\approx Gv$ for distinct $u, v \in A$, where Gu is the accessible pointed graph of points accessing u.

Actually, Aczel introduced an entire family of axioms AFA$^\sim$, one for each so-called regular bisimulation concept $\sim$. Each axiom characterizes canonical pictures as accessible pointed graphs satisfying a version of extensionality appropriate for $\sim$, and particular choices of $\sim$ yield the axioms AFA, SAFA, and FAFA. Each AFA$^\sim$ can be thought of as consisting of two parts: existence (stating that certain accessible pointed graph correspond to canonical pictures) and uniqueness (asserting a strengthening of the extensionality axiom). Larger $\sim$ lead to weaker existence assertions and stronger uniqueness assertions. The extremes are FAFA and AFA, which correspond to the smallest and largest regular bisimulation, respectively. In particular, the uniqueness part of every AFA$^\sim$ implies IE, which is the uniqueness part of FAFA.

Boffa [Bof72] introduced a theory which maximizes the universe of sets with respect to the plain axiom of extensionality, thus badly violating the isomorphism extensionality axiom IE. A weak version of Boffa's axiom postulates that every extensional binary relation $\langle a, e \rangle$ is isomorphic to $\langle t, \in \rangle$ for a transitive set t. The drawback of this statement is that it provides no easy way of extending existing sets: even if $\langle a, e \rangle$ already includes a set $\langle t_0, \in \rangle$ as its transitive part, the isomorphism from a to t does not have to preserve it, as there may be many sets isomorphic to t_0. For this reason, the full Boffa's axiom (the axiom of *superuniversality*), denoted here BAFA, asserts that for every transitive set t_0 and every extensional binary relation $\langle a, e \rangle$ that end-extends $\langle t_0, \in \rangle$, meaning $t_0 \subseteq a$ and $\in \restriction t_0 = e \cap (a \times t_0)$, there exists a transitive set t, and an isomorphism from $\langle a, e \rangle$ to $\langle t, \in \rangle$ that is the

identity on t_0. Thus, the axiom asserts a kind of saturation property for the transitive sets, namely, that they realize the types expressed by extensional binary relations end-extending a given transitive set. For example, under BAFA, there must be a proper class of Quine atoms, since we can extend the canonical picture of a given set of Quine atoms by a relation that describes what it would be like to have one more, or κ many more for any cardinality κ, and this new relation must be realized in a transitive set, which will have corresponding additional actual Quine atoms.

This level of saturation causes a high degree of homogeneity in any set-theoretic universe satisfying BAFA, where we have many distinct but isomorphic copies of whatever structure is produced. Such homogeneity, in turn, can cause a difficulty in class-length constructions by transfinite recursion, since the constructed objects are rarely unique, and so one cannot usually pick out the precise continuation of a given transfinite construction without a class choice principle. For this reason, it is convenient to include in Boffa's theory the global axiom of choice, which allows one to make choices in such a context, relative to a fixed class well-ordering of the sets. Thus, we work with BAFA over the theory $\mathrm{ZFGC}^{-\mathrm{f}}$ or $\mathrm{GBC}^{-\mathrm{f}}$, which includes the global axiom of choice. Boffa's theory has been used as a basis of formalization of nonstandard analysis by Ballard and Hrbáček [BH92].

In this article, we shall also make a few arguments in the theory asserting that the universe of sets is generated by a set or class of Quine atoms. Such theories are closely connected with the permutation models of set theory. Permutation models were originally constructed for the set theory ZFA with atoms (objects with no elements, but distinct from the empty set), which requires weakening of the axiom of extensionality. Alternatively, one can replace atoms with Quine atoms, simply by redefining the $\in$-relation on the atoms to make them all into self-singletons, that is, into Quine atoms; in this way extensionality is preserved at the expense of dropping foundation, which may be considered a less drastic deviation from the axioms of ZFC. Let us explain the precise meaning of being "generated". For any transitive class A, we define the *cumulative hierarchy over* A, the class denoted $\mathrm{WF}(A)$, as the elements appearing in the following recursive hierarchy:

$$\begin{aligned} \mathrm{WF}_0(A) &= A, \\ \mathrm{WF}_{\alpha+1}(A) &= \mathcal{P}(\mathrm{WF}_\alpha(A)), \\ \mathrm{WF}_\lambda(A) &= \bigcup_{\alpha<\lambda} \mathrm{WF}_\alpha(A) \quad \text{for limit } \lambda \end{aligned}$$

For example, the class of well-founded sets is simply $\mathrm{WF}(\emptyset)$. When A is a proper class, note that we take only the sub*sets* of the previous stage. In weaker set theories that may not be able to formalize this recursion on classes directly, we may equivalently define that $\mathrm{WF}(A)$ is the union of $\mathrm{WF}(t)$ for all sets $t \subseteq A$, and similarly $\mathrm{WF}_\alpha(A) = \bigcup_{t\subseteq A} \mathrm{WF}_\alpha(t)$. Every set $x \in \mathrm{WF}(A)$ has a corresponding rank, the least ordinal stage α for which $x \in \mathrm{WF}_{\alpha+1}(A)$. Such a rank function leads to a weak form of the axiom of foundation in $\mathrm{WF}(A)$, namely, every set x having an element y in $\mathrm{WF}(A)$ has such an element y of least rank, and such an element y will either be in A or have no elements in common with x, since any such element would have a lower rank than y in x. In particular, every infinite $\in$-descending sequence containing a set in $\mathrm{WF}(A)$ must eventually reach an element of A. The class $\mathrm{WF}(A)$ is transitive, as well as *full*, meaning that every subset of $\mathrm{WF}(A)$ is an element of $\mathrm{WF}(A)$, and indeed, $\mathrm{WF}(A)$ is the smallest full class containing A. When $V = \mathrm{WF}(A)$, then we say that the universe is *generated* by A, and in some of our arguments below, we shall consider models generated by a class of Quine atoms. Let us introduce the notation At to refer always to the class of Quine atoms, so that $V = \mathrm{WF}(\mathrm{At})$ just in case the universe is generated by its Quine atoms.

All of the anti-foundational theories we consider in this article are equiconsistent with ZFC. A convenient tool for showing the consistency of non-well-founded set theories is *Rieger's theorem* [Rie57] (cf. [Acz88]), which shows in $\mathrm{ZFC}^{-\mathrm{f}}$ that if M is a class endowed with a relation $E \subseteq M \times M$ that is extensional, set-like (meaning that the E-predecessors of any $a \in M$ form a set) and full (in the sense that every subset of M is the E-extension of some $a \in M$), then $\langle M, E\rangle$ satisfies all the axioms of $\mathrm{ZFC}^{-\mathrm{f}}$. If global choice is available, then we may also expand $\langle M, E\rangle$ to a model of global choice. In particular, any full transitive class, such as $\mathrm{WF}(A)$ for any transitive class A, is a model of $\mathrm{ZFC}^{-\mathrm{f}}$.

With suitable choices for such relations E (as explained in [Acz88]), we may arrange that $\langle M, E\rangle$ is extensional, set-like and full, while also satisfying AFA, SAFA, FAFA, or BAFA, whichever we prefer. In this way, inside any model of GBC we may construct models of $\mathrm{GBC}^{-\mathrm{f}}$ plus any of these anti-foundational theories. Furthermore, the WF of any full model $\langle M, E\rangle$ is isomorphic to the WF of the universe in which it is constructed. It follows that all these anti-foundational theories T are conservative over ZFC, in the sense that $T \vdash \phi^{\mathrm{WF}}$ if and only if $\mathrm{ZFC} \vdash \phi$, since every model of ZFC will arise as the WF of a model of the anti-foundational theory T.

We have mentioned that in any model of $\mathrm{ZFGC}^{-\mathrm{f}} + \mathrm{BAFA}$, there are a proper class of Quine atoms, and for any class A of Quine atoms, we may form the transitive class $\mathrm{WF}(A)$ inside this model. In particular, the theories asserting $\mathrm{GBC}^{-\mathrm{f}}$ plus "the universe is generated by a proper class of Quine atoms," or "the universe is generated by a set of five Quine atoms," or "... by a set of precisely $\aleph_6$ many Quine atoms," and so on, respectively, for any definable cardinality whose definition is absolute to $\mathrm{WF}(A)$, are each equiconsistent with ZFC and conservative over ZFC for assertions about the well-founded sets. One may also establish this directly from Rieger's theorem rather than via BAFA.

2. Some meta-mathematical issues

A number of meta-mathematical issues arise in any formalization of the Kunen inconsistency (we refer the readers to the discussion in the preliminary section of [HKP12]). The most obvious issue, of course, is that the quantifier involved in the assertion "there is no nontrivial elementary embedding j" is a second-order quantifier, not directly formalizable in the usual first-order theories such as ZFC. Many set theorists prefer to interpret all talk of classes in ZFC as referring to the first-order definable classes, and with such a formalization, the Kunen inconsistency becomes a scheme, asserting of each possible definition of j that it isn't an elementary embedding of the universe. (For example, Kanamori [Kan04] adopts this approach.) Nevertheless, this interpretation seems to miss much of the substance of the theorem, because there is an elementary proof of this formulation of the Kunen inconsistency, a simple diagonal argument not relying on the axiom of choice or any of the other combinatorial methods usually associated with the Kunen inconsistency (see [Suz99] and further discussion in [HKP12]).

So it seems natural to use a second-order set theory, such as Gödel–Bernays or Kelly–Morse set theory. In GBC, we have class quantifiers for expressing "$\exists j$," but then a second problem arises, namely, that it is not directly possible to express the assertion "j is elementary" in GBC. Kunen himself originally formulated his theorem in Kelley–Morse set theory KM precisely for this reason, since KM proves the existence of a full satisfaction class for first-order truth, making the assertion "j is elementary" expressible in KM. The Kelley-Morse theory, however, is strictly stronger theory than ZFC, even for arithmetical statements, and has a strictly higher consistency strength. Meanwhile, one can actually carry out Kunen's argument in the weaker theory GBC and even in $\mathrm{ZFC}(j)$, which is equiconsistent with and conservative over ZFC, by weakening the full elementarity of j to the assertion merely that "j

is Σ_1-elementary," which is formalizable in the first order language of set theory with a predicate for the class j. The point is that Kunen's argument actually proves this stronger statement, that there is no nontrivial Σ_1-elementary embedding j from V to V. Part of the reason for this is a lemma of Gaifman, which asserts that if $j\colon V \to V$ is Σ_1-elementary, then it is Σ_n-elementary for every meta-theoretic natural number n, by an induction carried out in the meta-theory. One issue here, however, is that Gaifman's lemma makes use of the fact that in ZF, every Σ_1-elementary embedding $j\colon V \to V$ is cofinal, in the sense that $\bigcup j[V] = V$, and this is no longer necessarily true in the absence of foundation, even with full elementarity, as shown in theorem 4 statement 3, although we may still assert that $\bigcup j[V]$ is full. It is in fact consistent with $\mathrm{ZFC}^{-\mathrm{f}}$ that there exists a Σ_1-elementary embedding $j\colon V \to V$ which is not elementary, and so this approach to formalizing the Kunen inconsistency statement does not fully succeed.

Some set theorists note that the proofs of the Kunen inconsistency show in fact that there can be no nontrivial elementary embedding of the form $j\colon V_{\lambda+2} \to V_{\lambda+2}$, a stronger statement that is expressible in the first-order language of set theory and provable in ZFC with no talk of classes. This formulation of the Kunen assertion, however, is not suitable in a context without foundation, since there could be an embedding $j\colon V \to V$ which is nontrivial, but only on ill-founded sets (as in theorem 2), and so ruling out $j\colon V_{\lambda+2} \to V_{\lambda+2}$ does not settle the question for $j\colon V \to V$ in the anti-foundational context.

In this article, we shall take a pragmatic approach. Rather than attempting to give a universally applicable definition of the Kunen assertion, we shall instead simply present the strongest results we can prove for each of the particular anti-foundational theories on a case-by-case basis. Our results on the nonexistence of embeddings usually apply already to Σ_1-elementary embeddings, and we formulate them generally over $\mathrm{GBC}^{-\mathrm{f}}$. Results on the existence of nontrivial embeddings provide definable embeddings in the $\mathrm{ZFC}^{-\mathrm{f}}$ or $\mathrm{ZFGC}^{-\mathrm{f}}$ versions of the theories, in which case the elementarity of the embedding can be formalized as an infinite schema. Sometimes the theory proves the existence of definable classes with a stronger property (given by a single formula) which implies the infinite elementarity schema; in particular, this is the case for *automorphisms* of the universe, meaning bijections $j\colon V \to V$ satisfying $x \in y \leftrightarrow j(x) \in j(y)$. A similar situation will arise with the elementary embeddings that we construct by threading a back-and-forth system of embeddings from one model to another.

Let us stress that throughout the paper, elementary embeddings and isomorphisms are only supposed to be elementary with respect to the first-order language of set theory; they need not be elementary with respect to the second-order language of classes, when working in $\mathrm{GBC}^{-\mathrm{f}}$, or with respect to the language including the global choice bijection of Ord with V, when working in $\mathrm{ZFGC}^{-\mathrm{f}}$.

3. The Kunen inconsistency on the well-founded sets

Let us now finally begin in earnest with the observation that a fragment of the Kunen inconsistency survives in set theory without the foundation axiom as the claim that there are no nontrivial elementary self-embeddings on the well-founded part of the universe.

Theorem 1. *Work either in* $\mathrm{GBC}^{-\mathrm{f}}$ *or in* $\mathrm{ZFC}^{-\mathrm{f}}$*. Then there is no nontrivial* Σ_1*-elementary embedding* $j\colon \mathrm{WF}\to\mathrm{WF}$*. In particular, every* Σ_1*-elementary embedding* $j\colon V\to V$ *fixes every well-founded set:* $j(x)=x$ *for all* $x\in\mathrm{WF}$*. Furthermore, the range* $j[V]$ *of any such embedding is a transitive full class.*

Proof. Consider the structure $\mathcal{W}$ obtained by restricting the universe to have only the objects in WF and to have only the classes A with $A\subseteq\mathrm{WF}$. Since WF is a definable class in the full universe, it follows that $A\cap\mathrm{WF}$ is a class whenever A is, and it is an elementary exercise to see that $\mathcal{W}$ is a model of GBC. In particular, if $j\colon\mathrm{WF}\to\mathrm{WF}$ were a nontrivial Σ_1-elementary embedding of WF, then j would be a class in $\mathcal{W}$ and furthermore would be a nontrivial Σ_1-elementary embedding of the entire set-theoretic universe from the perspective of $\mathcal{W}$, contrary to the original Kunen inconsistency, which shows that there can be no such embedding in any model of GBC.

Alternatively, if $j\colon\mathrm{WF}\to\mathrm{WF}$ were nontrivial and Σ_1-elementary, then since $\mathrm{WF}\models\mathrm{ZFC}$ it follows that j must have a critical point κ, and then one can define the critical sequence $\kappa_{n+1}=j(\kappa_n)$ and $\lambda=\sup_n\kappa_n$. It follows easily that $j(\lambda)=\lambda$ and consequently $j\restriction V_{\lambda+2}\colon V_{\lambda+2}\to V_{\lambda+2}$ is an elementary embedding from $V_{\lambda+2}$ to itself, which violates one of the ZFC formulations of the Kunen inconsistency inside WF.

For the further claims of the theorem, observe first that if $j\colon V\to V$ is Σ_1-elementary, then in fact $j\restriction\mathrm{WF}\colon\mathrm{WF}\to\mathrm{WF}$ is also Σ_1-elementary and hence trivial by the arguments of the previous paragraphs. To see this, note first that WF is Σ_1 definable, because a set x is in WF_α if and only if there is a ranking function $f\colon t\to\alpha$, with $a\in b\to f(a)<f(b)$, for some transitive set t with $x\in t$. This implies $j[\mathrm{WF}]\subseteq\mathrm{WF}$ and also $j(\mathrm{WF}_\alpha)=\mathrm{WF}_{j(\alpha)}$, and so $j\restriction\mathrm{WF}\colon\mathrm{WF}\to\mathrm{WF}$. One may see similarly that $j[\mathrm{Ord}]$ is cofinal in Ord. To see that

$j \restriction \mathrm{WF}\colon \mathrm{WF} \to \mathrm{WF}$ is Σ_1-elementary, notice that if there is an existential witness on the right side, $\mathrm{WF} \models \exists x\, \phi(x, j(u))$, where ϕ is Δ_0, then the witness x is found in some WF_β, which is contained in some $\mathrm{WF}_{j(\alpha)}$, since $j[\mathrm{Ord}]$ is cofinal in Ord, and so $V \models \exists x \in \mathrm{WF}_{j(\alpha)}\, \phi(x, j(u))$, which implies $\exists x \in \mathrm{WF}_\alpha\, \phi(x, u)$ by the Σ_1-elementarity of j, and so $\mathrm{WF} \models \exists x\, \phi(x, u)$ as desired.

In order to show that the range of j is transitive, assume $x \in j(u)$, and by the axiom of choice fix a bijection $f\colon \kappa \to u$ with some ordinal κ. It follows that $j(f)\colon j(\kappa) \to j(u)$ is also a bijection, and so $x = j(f)(\alpha)$ for some ordinal α. But since j is the identity on ordinals, it follows that $\alpha = j(\alpha)$ and so $x = j(f)(\alpha) = j(f)(j(\alpha)) = j(f(\alpha))$, which places x into the range of j, as desired.

Finally, to verify fullness, suppose $x \subseteq j[V]$. Let $u = j^{-1}[x]$, so that $x = j[u]$. Clearly $j[u] \subseteq j(u)$, but since $j(u) \subseteq j[V]$ by transitivity, we achieve also the converse, and so $j(u) = j[u] = x$. □

The fact that the range of such an embedding j must be transitive makes them totally different in character than the kinds of embeddings that are usually considered in the ZFC context, which never have transitive range.

4. Theories with nontrivial self-embeddings of the universe

We shall now prove that ZFC without the foundation axiom is relatively consistent with the existence of nontrivial automorphisms of the set-theoretic universe. In particular, if ZFC is consistent, then the Kunen inconsistency assertion is not provable in ZFC or GBC without making use of the axiom of foundation. We begin with the theories where the universe is generated by its Quine atoms, a situation where the automorphisms and elementary embeddings have a very transparent structure, admitting a complete description.

Theorem 2. *Work in* $\mathrm{GBC}^{-\mathrm{f}}$ *or* $\mathrm{ZFC}^{-\mathrm{f}}$ *and assume that the universe is generated by its Quine atoms, so that* $V = \mathrm{WF}(\mathrm{At})$.

1. *If* $j\colon V \to V$ *is an automorphism of the universe* V*, then* $j \restriction \mathrm{At}$ *is a permutation of the class of atoms* At.
2. *Every permutation* $\sigma\colon \mathrm{At} \to \mathrm{At}$ *of the atoms has a unique extension to an automorphism* $\overline{\sigma}\colon V \to V$ *of the entire universe.*

Proof. The first assertion is clear, as the class of Quine atoms is definable. For the second assertion, suppose that σ is a permutation of the class of Quine atoms At. We may extend this to a permutation $\overline{\sigma}$ of

all of $\mathrm{WF}(\mathrm{At})$ by the following recursion:

$$(*) \qquad \overline{\sigma}(x) = \begin{cases} \sigma(x) & \text{if } x \in \mathrm{At}, \\ \overline{\sigma}[x] & \text{otherwise} \end{cases}$$

We may view this directly as a recursion on the set-like well-founded relation $\in \setminus \mathrm{id}_{\mathrm{At}}$, or alternatively, as defining $\overline{\sigma} \restriction \mathrm{WF}(\mathrm{At})_\alpha$ by recursion on the rank α. It follows easily by induction that $\overline{\sigma} \restriction \mathrm{WF}(\mathrm{At})_\alpha$ is an $\in$-preserving automorphism of $\mathrm{WF}(\mathrm{At})_\alpha$, for every ordinal α, and so $\overline{\sigma}\colon V \to V$ is an automorphism of the entire universe V. Furthermore, any automorphism extending σ must obey $(*)$, and so $\overline{\sigma}$ is the unique such extension of σ to all of V. □

In particular, if At consists of at least two Quine atoms, then we have a permutation swapping two of them (leaving the rest in place), and so we may extend this to an automorphism of the entire universe $\mathrm{WF}(\mathrm{At})$, which will be definable from the Quine atoms to be swapped. In theorem 4, we shall use the recursion $(*)$ even when $\sigma\colon \mathrm{At} \to \mathrm{At}$ is merely injective, yielding an embedding $\overline{\sigma}\colon \mathrm{WF}(\mathrm{At}) \to \mathrm{WF}(\mathrm{At})$.

Corollary 3. *Work in* $\mathrm{ZFC}^{-\mathrm{f}}$*. If the universe is generated by its Quine atoms,* $V = \mathrm{WF}(\mathrm{At})$*, and there are at least two Quine atoms, then there is a nontrivial automorphism of* V*, definable from parameters.*

If the universe is generated from exactly two atoms, then we don't need any parameters to define the automorphism, since in this case σ and hence $\overline{\sigma}$ are both definable without parameters, since there is only one nontrivial permutation of a two-element set. The theorem generalizes from automorphisms to elementary embeddings as follows.

Theorem 4. *Work in* $\mathrm{GBC}^{-\mathrm{f}}$ *or* $\mathrm{ZFC}^{-\mathrm{f}}$*, and assume the universe is generated by its Quine atoms, so that* $V = \mathrm{WF}(\mathrm{At})$*.*

1. *If* $j\colon V \to V$ *is a* Σ_1*-elementary embedding, then* $j \restriction \mathrm{At}$ *is an injection of* At *to* At*, and* $j = \overline{j \restriction \mathrm{At}}$ *using the notation from* $(*)$*.*
2. *If* At *is a set, every* Σ_1*-elementary embedding* $j\colon V \to V$ *is an automorphism.*
3. *(Assuming global choice.) If the class of Quine atoms* At *is a proper class and* $\sigma\colon \mathrm{At} \to \mathrm{At}$ *is injective, then* $j = \overline{\sigma}\colon V \to V$ *is an elementary embedding, that is,* Σ_n*-elementary for every particular natural number* n*.*

Proof. (1) The first assertion is clear as $j(a) \in j(a)$ for any $a \in \mathrm{At}$. Also, $j(x) \supseteq j[x]$. On the other hand, since $j[V]$ is transitive by theorem 1, every element of $j(x)$ is of the form $j(y)$, where necessarily $y \in x$, and so $j(x) = j[x]$, which implies that j obeys $(*)$.

(2) The property of being the set of all atoms is Π_1, and so $j(\mathrm{At}) = \mathrm{At}$. But since the range of j is transitive, it follows that $\overline{j[\mathrm{At}]} = \mathrm{At}$ and hence $j \restriction \mathrm{At}$ is a permutation, which implies that $j = \overline{j \restriction \mathrm{At}}$ is an automorphism.

(3) This statement is a theorem scheme, a separate statement for each meta-theoretic natural number n. Assume that At is a proper class, with an injection $\sigma\colon \mathrm{At} \to \mathrm{At}$. Let $j = \overline{\sigma}\colon V \to V$ be the corresponding embedding arising from σ via $(*)$, and let $A = \sigma[\mathrm{At}]$, which is a proper class subclass of At. Since inductively j is an isomorphism of $\mathrm{WF}_\alpha(\mathrm{At})$ with $\mathrm{WF}_\alpha(A)$, for every ordinal α, it follows that j is an isomorphism of $\mathrm{WF}(\mathrm{At})$ with $\mathrm{WF}(A)$. In order to see that j is Σ_n-elementary from V to V, therefore, it therefore suffices to show that $\mathrm{WF}(A) \prec_{\Sigma_n} \mathrm{WF}(\mathrm{At})$. For this, we verify the Tarski–Vaught criterion: assume that a Σ_n statement $\phi(u, x)$ holds in V with $u \in \mathrm{WF}(A)$ and $x \in V$, we have to find $v \in \mathrm{WF}(A)$ such that $\phi(u, v)$. Fix sets $a \subseteq A$ and $a \subseteq b \subseteq \mathrm{At}$ such that $u \in \mathrm{WF}(a)$ and $x \in \mathrm{WF}(b)$. Using global choice, we can find an injection of b into A identical on a, which we can extend to a permutation $\tau\colon \mathrm{At} \to \mathrm{At}$. Since $\overline{\tau}$ is an automorphism identical on $\mathrm{WF}(a)$, we have $\phi(u, \overline{\tau}(x))$, where $\overline{\tau}(x) \in \mathrm{WF}(\tau[b]) \subseteq \mathrm{WF}(A)$. □

We remark that statement (2) applies more generally: if $V = \mathrm{WF}(t)$ for some set t, then every Σ_1-elementary embedding $j\colon V \to V$ is an automorphism: since $V = \mathrm{WF}(t)$ is a Π_1 property of t, asserting that every infinite $\in$-descending sequence contains an element of t, it follows that $V = \mathrm{WF}(j(t))$, which implies $V = j[V]$ as $j[V]$ is full.

Corollary 5. *Work in* $\mathrm{ZFC}^{-\mathrm{f}}$ *and assume that the universe is generated from a proper class of Quine atoms. Then there is an embedding* $j\colon V \to V$*, definable from parameters and* Σ_n*-elementary for every particular* n*, which is not an automorphism.*

Proof. The result follows from theorem 4 statement 3, if we have global choice. But let us prove it without global choice, using only the usual first-order axiom of choice. Since At is a proper class, we can prove the existence of at least n atoms for every $n \in \omega$ by induction on n. Using collection and choice, we can thus find an infinite countable set $\{a_n : n \in \omega\} \subseteq \mathrm{At}$. Define $\sigma\colon \mathrm{At} \to \mathrm{At}$ by $\sigma(a_n) = a_{n+1}$ and $\sigma(a) = a$ otherwise. This injection extends as in theorem 4 to an isomorphism

$\overline{\sigma}$ of WF(At) with WF(A), where $A = \text{At} \setminus \{a_0\}$. In the proof of the elementarity of WF(A) in WF(At) in theorem 4, we can assume $b = a \cup \{a_0\}$, hence we can define τ as a transposition of a_0 with any atom in At $\setminus\, b$ without using any choice.

Thus, the embedding j we produce arises via $(*)$ from an injection on At, and our argument shows as a theorem scheme that any embedding arising this way is Σ_n-elementary for any meta-theoretic natural number n. □

As we mentioned after corollary 3, in $\text{ZFC}^{-\text{f}} + V = \text{WF(At)}$, where At has precisely two Quine atoms, the unique nontrivial automorphism from theorem 2 is definable without parameters, as it is induced by swapping the two atoms. On the other hand, if $|\text{At}| \neq 2$, there is no nontrivial parameter-free definable Σ_1-elementary embedding, because any parameter-free definable class is preserved by all automorphisms, but the only function At $\to$ At that commutes with every permutation is the identity.

Next, we turn to Boffa's theory, which turns out to be rich in elementary embeddings and automorphisms. We shall start with some general remarks on basic consequences of BAFA. First, the axiom can be equivalently formulated in the following more convenient form: if $\langle a, e\rangle$ is an extensional binary relation which is an end-extension of $\langle a_0, e_0\rangle$, and f_0 is an isomorphism from $\langle a_0, e_0\rangle$ to $\langle t_0, \in\rangle$ with transitive t_0, there exists an isomorphism $f\colon \langle a, e\rangle \to \langle t, \in\rangle$ extending f_0 with transitive t.

Using global choice, BAFA implies a generalized Mostowski collapse lemma, asserting that every set-like extensional relation E on a class A is isomorphic to the $\in$ relation on some transitive class T. (The usual Mostowski collapse applies only to well-founded relations.) This can be shown by writing A as the union of an increasing chain $\{a_\alpha : \alpha \in \text{Ord}\}$ of E-transitive subsets $a_\alpha \subseteq A$, and constructing an increasing chain of isomorphisms $f_\alpha\colon \langle a_\alpha, E \restriction a_\alpha\rangle \to \langle t_\alpha, \in\rangle$ with transitive t_α. Its union is then an isomorphism of $\langle A, E\rangle$ to $T = \bigcup_\alpha t_\alpha$. This has immediate consequences for the construction of elementary embeddings of V into transitive classes M. For example, every ultrapower $\langle V^I/U, \in_U\rangle$ of the universe by an ultrafilter U on a set I gives rise to the corresponding set-like extensional relation $\in_U$, which therefore is realized as a transitive class $\langle M, \in\rangle \cong \langle V^I/U, \in_U\rangle$ via the generalized Mostowski collapse, and so the ultrapower map provides an elementary embedding $j\colon V \to M$ into a transitive class M, even when U is not countably complete, which of course does not happen in ZFC. This feature is essential for the development of nonstandard analysis in the framework

of [BH92]. However, one cannot construct a nontrivial elementary embedding into V itself in this way.

An important special case of BAFA is that every isomorphism of transitive sets can be extended to an isomorphism whose domain and range contains any given set. This means that the class of all isomorphisms of transitive sets forms a *back-and-forth system* from V to V (see [Mar02]), and consequently BAFA proves (as a scheme) that every isomorphism of transitive sets preserves the truth of any particular formula. With global choice, one can carry out the full back-and-forth construction:

Theorem 6. $\mathrm{ZFGC}^{-\mathrm{f}} + \mathrm{BAFA}$ *proves that every isomorphism of transitive sets can be extended to an automorphism of the universe. In particular, there exist nontrivial automorphisms.*

Proof. Let f_0 be any isomorphism of transitive sets. Using global choice and the back-and-forth property we just mentioned above, we may construct an increasing chain $\{f_\alpha : \alpha \in \mathrm{Ord}\}$ of isomorphisms of transitive sets such that every set eventually belongs to $\mathrm{dom}(f_\alpha) \cap \mathrm{ran}(f_\alpha)$ as α becomes large. The union $j = \bigcup_{\alpha\in\mathrm{Ord}} f_\alpha$ is therefore an automorphism of V extending f_0. If f_0 itself is nontrivial, such as an isomorphism of one Quine atom with another, we thereby ensure that j is nontrivial. □

Remark 7. It is not difficult to prove in BAFA that if a is any set and $b \notin \mathrm{WF}(\mathrm{TC}(a))$, then there is a proper class of sets isomorphic to b by an isomorphism fixing a. In view of theorem 6, this means that the orbit of b under automorphisms of V fixing a is a proper class, suggesting a rich Galois theory here.

Turning from automorphisms to general elementary embeddings, we have the following criterion.

Theorem 8. *Work in* $\mathrm{ZFGC}^{-\mathrm{f}} + \mathrm{BAFA}$ *or* $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{BAFA}$. *The following are equivalent for any class M.*

(1) *$M = j[V]$ for some elementary embedding $j : V \to V$.*
(2) *$M = j[V]$ for some Σ_1-elementary embedding $j : V \to V$.*
(3) *M is transitive and isomorphic to V.*
(4) *M is a full transitive model of* BAFA.

Proof. The formalization of statement 1 is problematic, since we cannot express it directly even as a scheme. What we mean by including it here is that, first, it contains statement 2 as an immediate special case; and second, if statement 4 holds, then using M and the global choice function we may define a particular class embedding $j : V \to V$ for

which $M = j[V]$ and prove that any class satisfying this definition is Σ_n-elementary for any particular natural number n. The implication (1) $\to$ (2) is thereby clear, and (3) $\to$ (4) also, since isomorphisms are truth-preserving for any given assertion. The implication (2) $\to$ (3) follows from theorem 1.

For the remaining implication (4) $\to$ (1), let I be the class of all isomorphisms $f\colon t \to s$ of transitive sets, with $s \subseteq M$. We claim that I is a back-and-forth system from V to M. The back direction follows directly from BAFA; for the forth direction, let $f \in I$ be as above, and $t' \supseteq t$ be a transitive set. Using the axiom of choice, we can construct an isomorphism $g\colon \langle t', \in\rangle \to \langle a, e\rangle$ extending f, where $\langle a, e\rangle$ is an extensional structure end-extending $\langle s, \in\rangle$ such that $a \setminus s \subseteq \mathrm{Ord}$. Since M is full, $\langle a, e\rangle \in M$, hence there is an isomorphism $h\colon \langle a, e\rangle \to \langle s', \in\rangle$ extending id_s for some transitive set $s' \in M$ using $M \models \mathrm{BAFA}$. Then $f' = h \circ g \in I$ extends f to t'.

Using transfinite recursion as in the proof of theorem 6, we can construct an isomorphism $j\colon V \simeq M$, which will be definable from M and the global choice function. Moreover, if $a \in M$, then $\mathrm{id}_{\mathrm{TC}(\{a\})} \in I$ can be extended in the same way to an isomorphism of V to M fixing a. This implies (as a scheme) that M and V must agree on the truth of any particular formula having parameters in M, and so $j\colon V \to V$ is an elementary embedding, in the sense that it is Σ_n-elementary for any particular n. □

Theorem 9. *In the theory* $\mathrm{ZFGC}^{-\mathrm{f}} + \mathrm{BAFA}$*, there is a definable class elementary embedding* $j\colon V \to V$*, which is not an automorphism.*

Proof. This is technically a theorem scheme, since we assert that the particular class embedding $j\colon V \to V$ that we define is Σ_n-elementary for every particular natural number n. The class j will be defined relative to the global choice function that is available in $\mathrm{ZFGC}^{-\mathrm{f}}$. Let $A = (V \times \{0\}) \cup \{a\}$, $a = \langle 1, 1\rangle$ and $E = \{\langle\langle x, 0\rangle, \langle y, 0\rangle\rangle : x \in y\} \cup \{\langle a, a\rangle\}$; thus, $\langle A, E\rangle$ is a disjoint union of an isomorphic copy of V and an additional Quine atom. By the generalized Mostowski collapse, there is an isomorphism $F\colon \langle A, E\rangle \to \langle T, \in\rangle$ to a transitive class T. Let j be the composition of F with the natural inclusion of V in A. Then j is an isomorphism of V to a transitive class $M = T \setminus \{F(a)\} \subsetneq V$, and M is an elementary substructure of V by theorem 8, hence $j\colon V \to V$ is an elementary embedding which is not an automorphism. □

Although the nontrivial automorphisms and elementary embeddings arising in theorems 6 and 9 were definable, these definitions made essential use of the global choice function, in order to carry out the transfinite

back-and-forth construction. We now show that one cannot define such embeddings in the first-order language of set theory alone:

Theorem 10. $\mathrm{ZFC}^{-\mathrm{f}} + \mathrm{BAFA}$ *proves that there is no nontrivial* Δ_0*-elementary embedding* $j\colon V \to V$ *of the universe with itself that is definable (with set parameters) in the first-order language of set theory.*

Proof. This is a theorem scheme, a separate statement for each possible definition, asserting that it does not define a nontrivial Δ_0-elementary embedding of the universe V to itself. Assume that such an embedding j is definable using parameters from a transitive set u. Since set isomorphisms preserve any particular formula, remark 7 implies that $j(x) \in \mathrm{WF}(u \cup \mathrm{TC}(x))$ for every x. Given any set x, let y be a set such that $y = \{y, x\}$ using Boffa's axiom. Then $j(y) = \{j(y), j(x)\}$, and $j(y) \in \mathrm{WF}(u \cup \mathrm{TC}(y)) = \mathrm{WF}(u \cup \mathrm{TC}(\{x\}) \cup \{y\})$. Since $j(y) \in j(y)$, the least α such that $j(y) \in \mathrm{WF}_\alpha(u \cup \mathrm{TC}(\{x\}) \cup \{y\})$ must be 0, so $j(y) \in u \cup \mathrm{TC}(\{x\}) \cup \{y\}$. Since $j^{-1}[u \cup \mathrm{TC}(\{x\})]$ is a set, but there is a proper class of solutions to $y = \{y, x\}$, we may assume that we choose y so that $j(y) \notin u \cup \mathrm{TC}(\{x\})$. But in this case $j(y) = y$ and consequently $j(x) = x$. □

Much of the material we have presented in this section on BAFA has been already known; for more information on automorphisms and elementary embeddings in $\mathrm{ZFGC}^{-\mathrm{f}} + \mathrm{BAFA}$, we refer the reader to [Paj99, Jer01].

5. Automorphisms and elementary embeddings of the universe

In the previous section, we produced models exhibiting several patterns of possibility for the existence of nontrivial automorphisms and nontrivial elementary embeddings of the universe. Specifically, we have a model with nontrivial automorphisms, but no other nontrivial elementary embeddings (corollary 3 and theorem 4 statement 2); we have a model with nontrivial automorphisms and other nontrivial elementary embeddings (theorem 4 statement 3); and we have the models of ZFC, which have no nontrivial automorphisms and no nontrivial elementary embeddings (by the Kunen inconsistency itself). We should like now to round out these possibilities with the missing case, namely, a model of set theory having nontrivial elementary embeddings, but no nontrivial automorphisms.

Theorem 11. *If* ZFC *is consistent, there is a model of* $\mathrm{GB}^{-\mathrm{f}} + \mathrm{AC}$ *having a class* $j\colon V \to V$ *that is an elementary embedding of the universe to itself, but in which no class is a nontrivial automorphism of the universe.*

Proof. Let us explain more precisely what we mean. Assuming ZFC is consistent, we shall construct a model of $\mathrm{GB}^{-\mathrm{f}} + \mathrm{AC}$ that has a class j that is a Σ_n-elementary embedding $j\colon V \to V$ of the universe to itself, for every meta-theoretic natural number n, but no class in the model is a nontrivial automorphism of the universe. Actually, we may make a more uniform claim, producing a model of $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{BAFA}$, for which there is a formula $\Phi(X)$ such that the model satisfies the single statement "the collection of classes satisfying Φ is closed under the class-formation axioms of $\mathrm{GB}^{-\mathrm{f}}$, and there is a non-identical function $j\colon V \to V$ preserving all classes satisfying Φ".

If ZFC is consistent, then there is a model of $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{BAFA}$, and we shall work inside that model. Therefore, assume $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{BAFA}$ and let $\overline{\mathrm{Ord}} = \mathrm{Ord} \cup \{\infty\}$, where $\alpha < \infty$ for all $\alpha \in \mathrm{Ord}$. Fix a bijection between a proper class A of Quine atoms, and the class of all sequences $a\colon \omega \to \overline{\mathrm{Ord}}$ such that $a(m) = \infty$ for all but finitely many m. We will identify $a \in A$ with the corresponding sequence in notation. Let $<$ denote the lexicographic order on A, namely,

$$a < b \leftrightarrow \exists n \in \omega\, (a \restriction n = b \restriction n \wedge a(n) < b(n)),$$

and let $A_n = \{a \in A : \forall m \geq n\ a(m) = \infty\}$. Notice that $\langle A, <\rangle$ is a dense linear order with largest element ∞^* (the constant ∞ sequence) and no least element, whereas $\langle A_n, <\rangle$ is a well-order isomorphic to the lexicographic order on $\overline{\mathrm{Ord}}^n$.

Let us also fix a set $r_{a,b}$ such that $r_{a,b} = \langle r_{a,b}, a, b\rangle$ for every $a, b \in A$ such that $a < b$. Note that the sets $r_{a,b}$ are pairwise distinct, and they are not in $\mathrm{WF}(A)$. For any subclass $B \subseteq A$, denote

$$\mathrm{WF}^{<}(B) = \mathrm{WF}(B \cup \{r_{a,b} : a, b \in B, a < b\}).$$

Let $M_n = \mathrm{WF}^{<}(A_n)$ and $M = \bigcup_{n\in\omega} M_n$. The first-order part of our desired model will be $\langle M, \in\rangle$. The purpose of adding the sets $r_{a,b}$ is to make $<$ definable in M, since $a < b \leftrightarrow M \models a \neq b \wedge \exists x\, x = \langle x, a, b\rangle$.

The classes of our model will be the subclasses of M that are invariant under certain partial isomorphisms, which we now describe. Let $\mathcal{P}^M(A)$ be the class of all subsets of A that belong to M, that is, subsets of A_n for some $n \in \omega$. Let I denote the class of all order-preserving isomorphisms $f\colon u \to v$ where $u, v \in \mathcal{P}^M(A)$ and $\infty^* \in u, v$.

Lemma 11.1. *Let $f \in I$, $f\colon u \to v$, $v \subseteq A_n$. For every $u \subseteq u' \in \mathcal{P}^M(A)$, there exists $f' \in I$ such that $f \subseteq f'$ and $f'\colon u' \to v'$ where $v' \subseteq A_{n+1}$.*

Proof. Since u' is well-ordered, we can find an increasing enumeration $u' \setminus u = \{a'_\alpha : \alpha < \gamma\}$. For each $\alpha < \gamma$, let a_α be the smallest element of u larger than a'_α (note that $a'_\alpha < \infty^* \in u$, so a_α exists). By assumption, $(f(a_\alpha))(m) = \infty$ for all $m \geq n$; let $f'(a'_\alpha)$ be the sequence in A_{n+1} which differs from $f(a_\alpha)$ only in the nth coordinate, where we put $(f'(a'_\alpha))(n) = \alpha$. Then f' has the required properties. □

Every order-isomorphism $f\colon u \to v$ from I uniquely extends to an $\in$-isomorphism $\overline{f}\colon \mathrm{WF}^<(u) \to \mathrm{WF}^<(v)$, since $\overline{f}(r_{a,b}) = r_{f(a),f(b)}$ is the unique set in M satisfying $x = \langle x, f(a), f(b)\rangle$, and then we use well-founded recursion on $\in$ to make $\overline{f}(x) = \overline{f}[x]$. More generally, for any $k \in \omega$, let $s_k\colon A \to A$ denote the shift operator $(s_k(a))(n) = a(n+k)$, and let $\overline{f}^k$ be the unique isomorphism $\mathrm{WF}^<(s_k^{-1}[u]) \to \mathrm{WF}^<(s_k^{-1}[v])$ such that for any $a \in s_k^{-1}[u]$, $\overline{f}^k(a) \restriction k = a \restriction k$ and $s_k(\overline{f}^k(a)) = f(s_k(a))$. That is, we leave the first k elements of a unchanged and apply f to the tail of the sequence. If $k \in \omega$ and $u \in \mathcal{P}^M(A)$, we say that a relation $R \subseteq M^r$ is *k-invariant*, if $\overline{f}^k$ preserves R for every $f \in I$, meaning that

$$R(x_1, \ldots, x_r) \leftrightarrow R(\overline{f}^k(x_1), \ldots, \overline{f}^k(x_r))$$

for all $x_1, \ldots, x_r \in \mathrm{dom}(\overline{f}^k)$.

Observe that R is k-invariant as an r-ary relation if and only if it is as a (unary) class of r-tuples, because $\overline{f}^k$ is an $\in$-isomorphism. If R is k-invariant, it is also k'-invariant for any $k' > k$.

Our model will be $\mathcal{M} = \langle M, \mathcal{X}, \in\rangle$, where $\mathcal{X}$ is the collection of all subclasses of M that are k-invariant for some $k \in \omega$. Formally, $\mathcal{X}$ is not an object of any kind in our model; rather, we are defining a (parametric) interpretation of $\mathrm{GB}^{-\mathrm{f}}$ in $\mathrm{GBC}^{-\mathrm{f}} + \mathrm{BAFA}$, delimiting classes of the interpreted theory by a formula of the background theory.

Lemma 11.2.

(1) *For every $k \in \omega$, the collection of k-invariant relations is closed under first-order definability and contains $\in$.*
(2) $\mathcal{M} \models \mathrm{GB}^{-\mathrm{f}} + \mathrm{AC}$.
(3) $\mathcal{M} \models$ *"every automorphism of $\langle V, \in\rangle$ is the identity".*

Proof. (1) Closure under Boolean operations is trivial. Let $S(x) \leftrightarrow \exists y\, R(x, y)$, so that $S = \mathrm{dom}(R)$, where R is k-invariant, and fix $f \in I$, $f\colon u \to v$, and $x \in \mathrm{WF}^<(s_k^{-1}[u])$. If $S(x)$, fix a $y \in M$ such that $R(x, y)$,

and $u' \in \mathcal{P}^M(A)$ such that $y \in \mathrm{WF}^<(u')$. By lemma 11.1, there is $g \supseteq f$ in I such that $\mathrm{dom}(g) \supseteq s_k[u']$. Then $y \in \mathrm{dom}(\overline{g}^k)$, hence $R(\overline{g}^k(x), \overline{g}^k(y))$, which implies $S(\overline{g}^k(x))$, where $\overline{g}^k(x) = \overline{f}^k(x)$. The other direction is symmetric.

(2) Being a full transitive class, each $\mathrm{WF}^<(u)$ is a model of $\mathrm{ZFC}^{-\mathrm{f}}$, hence $\mathcal{M}$ is a model of $\mathrm{ZFC}^{-\mathrm{f}}$ without collection. In $\mathcal{M}$, any set is a class as every $x \in \mathrm{WF}^<(u)$ with $u \subseteq A_n$ is n-invariant, and $\mathcal{M}$ satisfies the class formation axioms of $\mathrm{GB}^{-\mathrm{f}}$ by (1), it remains to show that it satisfies collection. Assume that $\forall x \in z\, \exists y\, R(x,y)$, where $R \subseteq M^2$ is k-invariant, and $z \in M_k$. If $x \in z$ and $y \in M$ satisfy $R(x,y)$, fix $u \in \mathcal{P}^M(A)$ such that $y \in \mathrm{WF}^<(u)$. By lemma 11.1, there exists $f \in I$ such that $\mathrm{dom}(f) \supseteq s_k[u]$ and $\mathrm{ran}(f) \subseteq A_{k+1}$. Then $\overline{f}^k(x) = x$ as $f(\infty^*) = \infty^*$, $\overline{f}^k(y) \in M_{k+1}$, and $R(x, \overline{f}^k(y))$. Thus $\forall x \in z\, \exists y \in M_{k+1}\, R(x,y)$. Using collection in the background theory, there is a subset $w \subseteq M_{k+1}$ such that $\forall x \in z\, \exists y \in w\, R(x,y)$; we have $w \in M_{k+1} \subseteq M$ as M_{k+1} is full.

(3) Let $j\colon M \to M$ be a k-invariant automorphism. Since A and $<$ are definable in M, it follows that j restricts to an automorphism of $\langle A, <\rangle$. Fix $n \geq k$; we claim that $j[A_n] = A_n$. Assume for contradiction that $j(a) = b$ where $a \in A_n \not\ni b$ (the other case is symmetric). Then $s_n(a) = \infty^* \neq s_n(b)$. Let f be an order-preserving function with domain $\{\infty^*, s_n(b)\}$ such that $f(s_n(b)) \neq s_n(b)$. Then $\overline{f}^n(a) = a$ and $\overline{f}^n(b) \neq b$, contradicting the n-invariance of j. Thus, j restricts to an automorphism of $\langle A_n, <\rangle$. Since this is a well order, it follows that $j \restriction A_n = \mathrm{id}$, and this implies $j \restriction M_n = \mathrm{id}$. Since $n \geq k$ was arbitrary, the entire automorphism is trivial $j = \mathrm{id}$, as desired. □

Let $\sigma\colon A \to A$ be the function such that $(\sigma(a))(0) = 1 + a(0)$ and $s_1(\sigma(a)) = s_1(a)$, so that $\sigma(\langle a_0, a_1, a_2, \dots\rangle) = \langle 1 + a_0, a_1, a_2, \dots\rangle$. Since σ is an order-preserving bijection from A to

$$A' = \{a \in A : a(0) \neq 0\},$$

it has a unique extension to an isomorphism $j\colon M \to \mathrm{WF}^<(A')$.

Lemma 11.3. *Both σ and j are 1-invariant and consequently belong to $\mathcal{X}$. The embedding j preserves all 0-invariant classes, and in particular, $j\colon M \to M$ is a nontrivial elementary embedding.*

Proof. The 1-invariance of σ is clear from the definition, as

$$\begin{aligned}\sigma(\overline{f}^1(\langle a_0, a_1, a_2, \dots\rangle)) &= \sigma(\langle a_0, f(\langle a_1, a_2, \dots\rangle)\rangle)\\ &= \langle 1 + a_0, f(\langle a_1, a_2, \dots\rangle)\rangle\\ &= \overline{f}^1(\langle 1 + a_0, a_1, a_2, \dots\rangle)\\ &= \overline{f}^1(\sigma(\langle a_0, a_1, a_2, \dots\rangle)),\end{aligned}$$

and j is definable in $\langle M, \in, \sigma\rangle$, hence it is also 1-invariant by lemma 11.2.

The preservation of 0-invariant relations follows from the fact that if $x \in \mathrm{WF}^{<}(u)$ with $\infty^* \in u$, then $f := \sigma \restriction u \in I$, and $\overline{f}^0(x) = j(x)$. This implies any instance of elementarity, since isomorphisms of transitive sets preserve the truth of any given formula, and so the Σ_n satisfaction relations are 0-invariant and hence preserved by j. □

Theorem 11 now follows from lemmas 11.2 and 11.3, and the proof is complete. □

Extending the idea of the previous argument, notice that we may extend $\sigma_k(\langle a_0, a_1, \dots\rangle) = \langle a_0, \dots, a_{k-1}, 1 + a_k, a_{k+1}, \dots\rangle$ to a nontrivial $(k+1)$-invariant embedding $j_k \colon M \to M$ which preserves all k-invariant classes. In particular, we may form the structure $\langle M, \in, j_0, j_1, j_2, \dots\rangle$, which is a model of $\mathrm{ZFC}^{-\mathrm{f}}(j_0, j_1, \dots)$ + "there is no nontrivial automorphism of $\langle V, \in\rangle$" + "j_k is a nontrivial elementary self-embedding of $\langle V, \in, j_0, \dots, j_{k-1}\rangle$" for every k.

Let us introduce some notation to help summarize what we've done. Let $\mathrm{Aut}(V)$ denote the collection of automorphisms of V and, informally, let $\mathrm{Eem}(V)$ the collection of elementary embeddings of V. This latter notation is informal, because in light of the issues mentioned in section 2, we are not actually able to express the property "$j \colon V \to V$ is an elementary embedding" as a single assertion about the class j, even in the full second-order language of Gödel–Bernays set theory. We are able to express that a class $j \colon V \to V$ is Σ_n-elementary for any particular natural number n in the meta-theory, and in this way we can say of any specific class j that "$j \colon V \to V$ is elementary" as an infinite scheme of statements, and this scheme-theoretic treatment of elementarity suffices for many applications. Meanwhile, in contrast, the property "$j \colon V \to V$ is an automorphism" *is* a first-order expressible property of j, and one can prove that any such automorphism is Σ_n-elementary for any particular natural number n in the meta-theory.

We have produced models of $\mathrm{ZFC}^{-\mathrm{f}}$ realizing all four separating refinements of the fact that $\{\mathrm{id}_V\} \subseteq \mathrm{Aut}(V) \subseteq \mathrm{Eem}(V)$.

(1) $\{\mathrm{id}_V\} = \mathrm{Aut}(V) = \mathrm{Eem}(V)$. Models of ZFC have no nontrivial automorphisms or elementary self-embeddings of universe.
(2) $\{\mathrm{id}_V\} \subsetneq \mathrm{Aut}(V) = \mathrm{Eem}(V)$. If $V = \mathrm{WF}(A)$ for a set A of at least two Quine atoms, then there are nontrivial automorphisms, but no other nontrivial elementary embeddings.
(3) $\{\mathrm{id}_V\} = \mathrm{Aut}(V) \subsetneq \mathrm{Eem}(V)$. The model of theorem 11 has no nontrivial automorphisms, but does have a nontrivial elementary embedding.
(4) $\{\mathrm{id}_V\} \subsetneq \mathrm{Aut}(V) \subsetneq \mathrm{Eem}(V)$. If $V = \mathrm{WF}(A)$ for a proper class of Quine atoms, then there are nontrivial automorphisms, as well as non-automorphic nontrivial elementary embeddings.

In the case of statement (2), the model $V = \mathrm{WF}(A)$ for a set A of Quine atoms, we have $\mathrm{Aut}(V) = \mathrm{Aut}(A)$, the permutation group of the set A. In fact, let us now show that every group can arise this way.

Theorem 12. *Assume* $\mathrm{GBC}^{-\mathrm{f}}$+BAFA. *Then for every group G, there is a transitive set A_G whose automorphism group is isomorphic to G, and the automorphism group of the corresponding cumulative universe* $\mathrm{WF}(A_G)$ *generated over this set is also isomorphic to G, in the sense that every automorphism of A_G extends to a unique automorphism of* $\mathrm{WF}(A_G)$ *and every automorphism of* $\mathrm{WF}(A_G)$ *arises this way.*

Proof. Work in $\mathrm{GBC}^{-\mathrm{f}}$ + BAFA, and fix any group G, which we may assume is in WF, since it has an isomorphic copy there. Let A_G consist of the transitive closure of the following objects:

(1) Quine atoms a_g for every $g \in G$, and
(2) Sets $r_{g,h}$ satisfying $r_{g,h} = \langle r_{g,h}, a_g, h, a_{gh}\rangle$ for every $g, h \in G$.

If $j\colon A_G \to A_G$ is an automorphism of A_G, then j must fix every element of G, as these are in WF, and it must permute the Quine atoms of A_G. It follows that $j(a_g) = a_{\pi(g)}$ for some permutation π of G and furthermore that $j(r_{g,h}) = \langle j(r_{g,h}), a_{\pi(g)}, h, a_{\pi(gh)}\rangle = r_{\pi(g),h}$. Thus, the embedding j is determined uniquely by π, and we may see also that $\pi(gh) = \pi(g)h$, which implies that $\pi(h) = gh$ for every h, where $g = j(1)$. Conversely, for every $g \in G$, the permutation $\pi(h) = gh$ extends to an automorphism j_g of A_G. Furthermore, $j_{gh} = j_g \circ j_h$, and so $g \mapsto j_g$ is an isomorphism of G with $\mathrm{Aut}(A_G)$, meaning the automorphisms of the structure $\langle A_G, \in\rangle$. So the automorphism group of the transitive set A_G is isomorphic to G. Every automorphism of A_G extends canonically to an automorphism of $\mathrm{WF}(A_G)$ via $(*)$, and conversely, every automorphism of $\mathrm{WF}(A_G)$ arises from an automorphism of A_G, since it must permute the Quine atoms and the sets of the form r_{gh}. So in $\mathrm{WF}(A_G)$, the full automorphism group of the universe is

definably isomorphic with G, in the sense that every automorphism of A_G extends to an automorphism of $\mathrm{WF}(A_G)$ and every automorphism $\mathrm{WF}(A_G)$ arises as such an extension. □

6. Theories without nontrivial self-embeddings

Let us show now in contrast that there are no nontrivial elementary embeddings or automorphisms under Aczel's anti-foundation axiom and similar anti-foundational theories where equality of sets is determined by the isomorphism type of the underlying $\in$-relation on the hereditary members of the set.

Theorem 13. *Under* $\mathrm{GBC}^{-\mathrm{f}}+\mathrm{IE}$*, there is no nontrivial* Σ_1*-elementary embedding* $j\colon V\to V$ *of the universe to itself.*

Proof. Suppose that $j\colon V\to V$ is a Σ_1-elementary embedding of the universe to itself. Take any $x\in V$, and let t be the transitive closure of $\{x\}$. Since the range $j[V]$ is transitive by theorem 1, it follows that $j[t]$ is also transitive, and $j\restriction t$ is an isomorphism of $\langle t,\in\rangle$ with $\langle j(t),\in\rangle$. By IE, therefore, $j\restriction t$ is the identity, and so $j(x)=x$. □

What is going on is this: under the IE principle, every transitive set t is determined by the isomorphism type of the underlying directed graph $\langle t,\in\rangle$; but by the axiom of choice, this graph has an isomorphic copy in the well-founded universe WF, which is consequently fixed by j. Thus, $\langle t,\in\rangle$ and $\langle j(t),\in\rangle$ are both isomorphic to the same graph and hence to each other, and so $j(t)=t$. Applying this to the transitive closure t of $\{\,x\,\}$, it follows that $j(x)=x$ for every set x.

Corollary 14. *Under* $\mathrm{GBC}^{-\mathrm{f}}+\mathrm{AFA}$*,* $\mathrm{GBC}^{-\mathrm{f}}+\mathrm{SAFA}$*,* $\mathrm{GBC}^{-\mathrm{f}}+\mathrm{FAFA}$*, and more generally,* $\mathrm{GBC}^{-\mathrm{f}}+\mathrm{AFA}^{\sim}$ *for any regular bisimulation concept* $\sim$*, there is no nontrivial* Σ_1*-elementary embedding of the universe.*

The proof of theorem 13 makes a fundamental use of the axiom of choice, both in order to know that $j\restriction \mathrm{WF}$ is trivial and to find a surjection from a well-founded set to a given set. The proof of theorem 13 applies as is to $\mathrm{GB}^{-\mathrm{f}}+\mathrm{IE}$ if we know a priori that $j[V]$ is transitive, in particular $\mathrm{GB}^{-\mathrm{f}}+\mathrm{IE}$ proves that there are no nontrivial *automorphisms* of the universe. However, this does not resolve the case of general elementary embeddings, so it is natural to inquire whether one may prove theorem 13 without the axiom of choice.

Question 15. *Is it consistent with* $\mathrm{GB}^{-\mathrm{f}}+\mathrm{AFA}$*,* $\mathrm{ZF}^{-\mathrm{f}}+\mathrm{AFA}$*, or a similar extension of* $\mathrm{ZF}^{-\mathrm{f}}+\mathrm{IE}$ *that there is a nontrivial elementary embedding of the universe to itself?*

In light of the fact that it remains a prominent open question whether one can prove the Kunen inconsistency in GB, that is, with the axiom of foundation but without the axiom of choice, we shouldn't expect an easy negative answer to this question. But perhaps one might hope for a positive answer by building a suitable model of AFA without the axiom of choice, where some ill-founded sets have no well-founded copies of their hereditary $\in$-graphs and can be subject to nontrivial embeddings.

References

[Acz88] Peter Aczel. *Non-well-founded sets.* CSLI Lecture Notes 14, 1988.

[BH92] David Ballard and Karel Hrbáček. Standard foundations for nonstandard analysis. *Journal of Symbolic Logic,* 57(2):741–748, 1992.

[BM96] Jon Barwise and Lawrence Moss. *Vicious Circles: On the Mathematics of Non-Wellfounded Phenomena.* CSLI Lecture Notes 60, 1996.

[Bof72] Maurice Boffa. Forcing et négation de l'axiome de Fondement. *Acad. Roy. Belg. Cl. Sci. Mém. Coll. in-8° (2)*, 40(7):52, 1972.

[Dag13] Ali Sadegh Daghighi. Is there any large cardinal beyond Kunen inconsistency? MathOverflow. URL: http://mathoverflow.net/q/136057 (version: 2013-07-08).

[Fin26] Paul Finsler. Über die Grundlagen der Mengenlehre, I. *Mathematische Zeitschrift*, 25:683–713, 1926.

[FH83] Marco Forti and Furio Honsell. Set theory with free construction principles. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze, Serie IV*, 10:493–522, 1983.

[GHJ] Victoria Gitman, Joel David Hamkins, and Thomas A. Johnstone. What is the theory ZFC without Powerset? under review, arxiv.org/abs/110.2430.

[HKP12] Joel David Hamkins, Greg Kirmayer, and Norman Lewis Perlmutter. Generalizations of the Kunen inconsistency. *Annals of Pure and Applied Logic*, 163(12):1872–1890, 2012.

[Jer01] Emil Jeřábek. *Reflexe v neregulárních univerzech.* Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague, 2001.

[Kan04] Akihiro Kanamori. *The Higher Infinite, Corrected Second Edition.* Springer-Verlag, 2004.

[Kun78] Kenneth Kunen. Saturated ideals. *Journal of Symbolic Logic*, 43(1):65–76, March 1978.

[Mar02] David Marker. *Model Theory: An Introduction.* Graduate Texts in Mathematics 217, Springer, 2002.

[Paj99] Petr Pajas. *Endomorfismy, invariantní třídy a nestandardní principy v neregulárním universu množin.* Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague, 1999.

[Rie57] Ladislav Rieger. A contribution to Gödel's axiomatic set theory, I. *Czechoslovak Mathematical Journal,* 7(3):323–357, 1957.

[Sco60] Dana Scott. *A Different Kind of Model for Set Theory.* Unpublished paper given at the 1960 Stanford Congress of Logic, Methodology and Philosophy of Science.

[Suz99] Akira Suzuki. No elementary embedding from V into V is definable from parameters. *J. Symbolic Logic*, 64(4):1591–1594, 1999.

(A. S. Daghighi) School of Mathematics, Amirkabir University of Technology, Hafez avenue 15194, Tehran, Iran
E-mail address: a.s.daghighi@gmail.com

(M. Golshani) School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P.O. Box: 19395-5746, Tehran, Iran
E-mail address: golshani.m@gmail.com
URL: http://math.ipm.ac.ir/golshani/

(J. D. Hamkins) Mathematics, Philosophy, Computer Science, The Graduate Center of The City University of New York, 365 Fifth Avenue, New York, NY 10016 & Mathematics, College of Staten Island of CUNY, Staten Island, NY 10314
E-mail address: jhamkins@gc.cuny.edu
URL: http://jdh.hamkins.org

(E. Jeřábek) Institute of Mathematics of the Academy of Sciences, Žitná 25, 115 67 Praha 1, Czech Republic
E-mail address: jerabek@math.cas.cz
URL: http://math.cas.cz/~jerabek