NONLINEARITY IN THE HIERARCHY OF LARGE CARDINAL CONSISTENCY STRENGTH

JOEL DAVID HAMKINS

ABSTRACT. I present instances of nonlinearity and illfoundedness in the hierarchy of large cardinal consistency strength—as natural or as nearly natural as I can make them—and consider philosophical aspects of the question of naturality with regard to this phenomenon. I provide various cautious enumerations of the ZFC axioms which succeed in enumerating all the ZFC axioms, but with a strictly weaker consistency strength than the usual (incautious) enumeration. And similarly there are various cautious versions of many large cardinal hypothesis, which are natural but also incomparable in consistency strength.

It is a mystery often mentioned in the foundations of mathematics, a fundamental phenomenon to be explained, that our best and strongest mathematical theories seem to be linearly ordered and indeed well-ordered by consistency strength. Given any two of the familiar large cardinal hypotheses, for example, generally one of them will prove the consistency of the other.

Why should it be linear? Why should the large cardinal notions line up like this, when they often arise from completely different mathematical matters? Measurable cardinals arise from set-theoretic issues in measure theory; Ramsey cardinals generalize ideas in graph coloring combinatorics; compact cardinals arise with compactness properties of infinitary logic. Why should these disparate considerations lead to principles that are linearly related by direct implication and consistency strength?

The phenomenon is viewed by many in the philosophy of mathematics as significant in our quest for mathematical truth. In light of Gödel incompleteness, after all, we must eternally seek to strengthen even our best and strongest theories. Is the linear hierarchy of consistency strength directing us along the elusive path, the "one road upward" as John Steel [Ste13] describes it, toward the final, ultimate mathematical truth? That is the tantalizing possibility.

Meanwhile, we do know as a purely formal matter that the hierarchy of consistency strength is not actually well-ordered—it is ill-founded, densely ordered, and nonlinear. The statements usually used to illustrate these features, however, are weird self-referential assertions constructed in the Gödelian manner via the fixed-point lemma—logic-game trickery, often dismissed as unnatural.

Many set theorists claim that amongst the *natural* assertions, consistency strengths remain linearly ordered and indeed well ordered. H. Friedman [Fri98] refers to "the apparent comparability of naturally occurring logical strengths as one of the great mysteries of [the foundations of mathematics]."

Andrés Caicedo says,

> It is a remarkable empirical phenomenon that we indeed have comparability for natural theories. We expect this to always be the case, and a significant amount of work in inner model theory is guided by this belief. [Cai11]

Stephen G. Simpson writes:

> It is striking that a great many foundational theories are linearly ordered by $<$. Of course it is possible to construct pairs of artificial theories which are incomparable under $<$. However, this is not the case for the "natural" or non-artificial theories which are usually regarded as significant in the foundations of mathematics. The problem of explaining this observed regularity is a challenge for future foundational research. [Sim09]

John Steel writes "The large cardinal hypotheses [the ones we know] are themselves wellordered by consistency strength," and he formulates what he calls the "vague conjecture" asserting that

> If T is a natural extension of ZFC, then there is an extension H axiomatized by large cardinal hypotheses such that $T \equiv_{\mathrm{Con}} H$. Moreover, $\leq_{\mathrm{Con}}$ is a prewellorder of the natural extensions of ZFC. In particular, if T and U are natural extensions of ZFC, then either $T \leq_{\mathrm{Con}} U$ or $U \leq_{\mathrm{Con}} T$. [Ste14]

Peter Koellner writes

> Remarkably, it turns out that when one restricts to those theories that "arise in nature" the interpretability ordering is quite simple: There are no descending chains and there are no incomparable elements—the interpretability ordering on theories that "arise in nature" is a wellordering. [Koe11]

Let me refer to this position as the *natural linearity position*, the assertion that all natural assertions of mathematics are linearly ordered by consistency strength. The strong form of the position, asserted by some of those whom I have cited above, asserts that the natural assertions of mathematics are indeed well-ordered by consistency strength. By all accounts, this view appears to be widely held in large cardinal set theory and the philosophy of set theory.

Despite the popularity of this position, I should like in this article to explore the contrary view and directly to challenge the natural linearity position.

Main Question. *Can we find natural instances of nonlinearity and illfoundedness in the hierarchy of consistency strength?*

I shall try my best.

1. Formal instances of consistency-strength nonlinearity

Let me begin by setting aside the naturality requirement (for this section only) and reviewing as a purely formal matter that both nonlinearity and ill-foundedness occur in the hierarchy of consistency strength. This will be established with certain self-referential sentences constructed via the Gödel fixed-point lemma—precisely the sentences often dismissed as unnatural. The results of this section are well

known and have essentially become part of the mathematical logic folklore; I shall mention several references presently.

The consistency-strength order relation at the center of the discussion is defined as follows, where $\mathrm{Con}(T)$ is the assertion "T is consistent," expressed using a given arithmetically expressible definition of the theory T.

Definition 1. Theory S is below theory T in consistency strength, written $S \leq T$, if the implication $\mathrm{Con}(T) \to \mathrm{Con}(S)$ is provable in a base theory, fixed for the purpose of this relation. The theory S is strictly below T, written $S < T$, if $S \leq T$ but not conversely; and the theories are equiconsistent, written $S \equiv T$, when $S \leq T$ and $T \leq S$, which provides a hierarchy of degrees of consistency strength.

One commonly sees Peano arithmetic PA as the base theory in arithmetic contexts and ZFC in set theory. Much weaker base theories, however, such as primitive recursive arithmetic PRA, actually suffice for nearly all of the usual features one seeks in the hierarchy of consistency strength, whether in arithmetic or set theory, including large cardinal set theory. A strong base theory, such as ZFC, can erase distinctions in the hierarchy that a weaker theory reveals; the iterated consistency tower over PA, for example, with $\mathrm{Con}(\mathrm{PA})$, $\mathrm{Con}(\mathrm{PA} + \mathrm{Con}(\mathrm{PA}))$ and so on, is trivialized in ZFC, which proves outright all those consistency statements. Meanwhile, instances of nonlinearity are more compelling over a stronger base theory than a weak one, since incomparability persists to weaker base theories; and so in this article I shall adopt PA as the default base theory for arithmetic and ZFC as the default base theory for set theory, although very little of my analysis depends on this.

A convenient sufficient condition for the strict relation $S < T$, when T extends the base theory, occurs when $T \vdash \mathrm{Con}(S)$. In this case $\mathrm{Con}(T)$ implies $\mathrm{Con}(T + \mathrm{Con}(S))$, which implies $\mathrm{Con}(S)$, but $\mathrm{Con}(S)$ cannot prove $\mathrm{Con}(T)$ over the base theory, for then the base theory plus $\mathrm{Con}(S)$ would prove its own consistency, contrary to the incompleteness theorem. This method enables one to prove instances of the strict relation $S < T$ without ever having explicitly to prove an instance of nonprovability, since this part is in effect offloaded to the incompleteness theorem. In the large cardinal hierarchy, many instances of the strict relation in consistency strength are often proved in just this way—if there is a Mahlo cardinal κ, for example, then V_κ is a set-sized model of ZFC with a proper class of inaccessible cardinals, showing the consistency of this theory, which is therefore strictly weaker than a Mahlo cardinal.

In the research literature, not everyone is using exactly the same hierarchy of strength; there are several closely related notions. First, of course, there is the hierarchy of direct implication, but nearly everyone recognizes that this isn't really what we want, for we already have numerous natural instances of nonlinearity for direct implication, even for several completely standard large cardinal notions. Beyond this, Per Lindstrom [Lin03] provides a well-developed account for the hierarchy of interpretability strength, which is closely related to but not identical with the hierarchy of consistency strength; nevertheless, much of Lindstrom's analysis carries over easily to consistency strength. Koellner [Koe11] similarly treats the interpretability hierarchy in his discussion of large cardinal strength. It may be that for most purposes including philosophical analysis we should ultimately be using interpretability strength rather than consistency strength. These two hierarchies are different, however, even in the case of ZFC and Gödel-Bernays set theory

GBC, since these two theories are equiconsistent but not mutually interpretable. Meanwhile, Harvey Friedman [Fri98] uses the consistency-strength relation of definition 1, but only for sentences over the base theory, rather than theories. John Steel [Ste14] uses exactly the relation of definition 1. Stephen Simpson [Sim09], however, defines the consistency-strength order as $S < T \iff T \vdash \mathrm{Con}(S)$. This is the sufficiency property I mentioned earlier, which suffices for the strict order of definition 1, but Simpson's order is not actually identical with the strict order of definition 1—for example, Simpson's order is not dense amongst extensions of the base theory, since it can have nothing strictly between PA and PA + Con(PA). To my way of thinking, we should want to analyze consistency strength through a reflexive preorder relation $S \leq T$, which gives rise to the equiconsistency degrees via $S \equiv T \iff S \leq T \leq S$, and Simpson's approach does not seem to do this. For the rest of this article, therefore, I shall proceed with the notion of consistency strength provided by definition 1.

Theorem 2. *There are statements σ and τ in the language of arithmetic with incomparable consistency strengths over* PA*. That is, neither* $\mathrm{Con}(\mathrm{PA} + \sigma)$ *nor* $\mathrm{Con}(\mathrm{PA} + \tau)$ *provably implies the other over* PA*.*

Proof. Using the double version of the fixed-point lemma, we can find distinct sentences σ and τ, each asserting that for any refutation of the other sentence in the theory PA + Con(PA), there is a smaller refutation of itself, one with a smaller Gödel code. That is, σ asserts that for any proof of $\neg\tau$ in PA + Con(PA), there is a smaller proof of $\neg\sigma$; and similarly vice versa with τ.

Neither of these sentences, I claim, is actually refutable in PA + Con(PA), since if one of them were refutable, then one of them would have the smallest refutation, and this would make that sentence also provably true in PA, which would contradict the consistency of the theory PA + Con(PA). So neither sentence is actually refutable and hence both are (vacuously) true.

Since σ is not refutable, it follows that PA+Con(PA)+σ is consistent, and so it is also consistent with the assertion of its own inconsistency $\neg\,\mathrm{Con}(\mathrm{PA}+\mathrm{Con}(\mathrm{PA})+\sigma)$. In any model of this combined theory, σ is refutable in PA+Con(PA), but since also σ is true there, there must not be any smaller refutation of τ. Since this syntactic situation will be provable in PA, it follows in light of what the sentences assert that the model thinks that PA proves that σ is true and τ is false. So from Con(PA) it follows both that Con(PA + σ) and $\neg\,\mathrm{Con}(\mathrm{PA} + \tau)$ in this model.

Similarly, since τ is not refutable, we may consider the theory PA + Con(PA) + τ analogously, and thereby find a model in which Con(PA+τ) but $\neg\,\mathrm{Con}(\mathrm{PA}+\sigma)$. So the two sentences have incomparable consistency strength over PA, as claimed.   □

In fact, we don't need the double fixed-point method to prove theorem 2, since it is an immediate consequence of the following stronger result, which avoids the double fixed-points, while also achieving the incomparability for a sentence with its negation.

Theorem 3. *There is a statement η in the language of arithmetic, such that η and $\neg\eta$ have incomparable consistency strengths over* PA*.*

Proof. Let η be the Rosser sentence of the theory PA + Con(PA), that is, the sentence asserting of itself that for any proof of η from this theory, there is a smaller proof of $\neg\eta$, smaller in the sense of Gödel codes. The usual Rosser argument shows

that η is neither provable nor refutable in this theory. Namely, if η is actually provable, then because of what η asserts, we will have proved that $\neg\eta$ has a proof with a smaller Gödel code, but since the theory is consistent, none of those can be proofs of $\neg\eta$, and we can prove this; similarly, if η were actually refutable, then we will have proved that η is provable with a smaller code, which again would contradict the consistency of the theory.

Since η is not provable, there is a model of $\mathrm{PA} + \mathrm{Con}(\mathrm{PA}) + \neg\eta$. In light of what η asserts, this model thinks that there is a proof of η from $\mathrm{PA} + \mathrm{Con}(\mathrm{PA})$ with no smaller proof of $\neg\eta$. Since the model thinks that PA can prove these concrete syntactic facts, which suffice for the failure of η, the model thinks that PA proves $\neg\eta$. Since $\mathrm{Con}(\mathrm{PA})$ holds in this model, it follows that this is a model of $\mathrm{Con}(\mathrm{PA} + \neg\eta) + \neg\,\mathrm{Con}(\mathrm{PA} + \eta)$.

Conversely, since η is not refutable, it follows that $\mathrm{PA}+\mathrm{Con}(\mathrm{PA})+\eta$ is consistent. By the incompleteness theorem, there is a model M of this theory that also thinks this theory is inconsistent. So M has a proof of $\neg\eta$ from $\mathrm{PA}+\mathrm{Con}(\mathrm{PA})$, and because η is true in the model, there must be such a proof of $\neg\eta$ with no smaller proof of η. Since PA can prove this concrete syntactic fact, which furthermore suffices to validate η, it follows that the model thinks that PA proves η. So from $\mathrm{Con}(\mathrm{PA})$ in this model, we must also have $\mathrm{Con}(\mathrm{PA} + \eta) + \neg\,\mathrm{Con}(\mathrm{PA} + \neg\eta)$.

So we have exhibited models showing that neither the consistency statement $\mathrm{Con}(\mathrm{PA}+\eta)$ nor $\mathrm{Con}(\mathrm{PA}+\neg\eta)$ implies the other over PA, and so these consistency strengths are incomparable.   □

The sentences η and $\neg\eta$ of theorem 3 exhibit what is known as the *double-jump* phenomenon for consistency strength, which occurs when both a sentence and its negation jump up in consistency strength. Precisely because η and $\neg\eta$ have incomparable consistency strengths, it follows that neither $\mathrm{PA} + \eta$ nor $\mathrm{PA} + \neg\eta$ can be equiconsistent with PA alone (for then they would be comparable), and so both of them jump. Contrast this situation with the usual Rosser sentence ρ defined with respect to PA, since this sentence ρ has no jumps—both $\mathrm{PA} + \rho$ and $\mathrm{PA} + \neg\rho$ are equiconsistent with PA itself. Meanwhile, the Gödel sentence γ has one jump, because $\mathrm{PA}+\gamma$ is equivalent to $\mathrm{PA}+\mathrm{Con}(\mathrm{PA})$, which has strictly higher consistency strength than PA, but $\mathrm{PA} + \neg\gamma$ is equivalent to $\mathrm{PA} + \neg\,\mathrm{Con}(\mathrm{PA})$, which remains equiconsistent with PA.

The single and double jump phenomenon is also commonly considered for the hierarchy of interpretative strength, as in [Koe11], but I'd like to mention a few differences between these hierarchies. Consistency strength is closely related to interpretative strength, to be sure, because if a theory T proves $\mathrm{Con}(S)$, then S is interpretable in T with strictly lower interpretability strength simply by constructing the Henkin model; so these instances of strong increase in consistency strength are also instances of strict increase in interpretative strength. But the hierarchies are not the same. To see one difference, notice that [Koe11] points out that no Π_1^0 sentence can realize the no-jumping situation for interpretative strength, whereas we have said that the Rosser sentence ρ is no-jumping in consistency strength, and this sentence has complexity Π_1^0. The difficulty for interpretation is that any model of PA satisfying a Σ_1^0 statement will think that this statement must be true in all the models it interprets, since the very same existential instance is in effect inserted into the interpreted models. But this problem does not arise with consistency strength, since one can think an existential statement is consistent without

yet having a specific concrete instance. With the Rosser sentence ρ, for example, the mere consistency of the existential case $\neg\rho$ does not cause a difficulty for the consistency of ρ itself.

One can easily modify the argument of theorem 3 to look instead for proofs from $\mathrm{PA} + \mathrm{Con}(S)$, where S is a consistent extension of PA. The result is a sentence η for which $S + \eta$ and $S + \neg\eta$ have incomparable consistency strengths. A slight generalization of this method shows in fact that the hierarchy of consistency strengths is a dense order:

Theorem 4. *If theory S has strictly weaker consistency strength than theory T, both extending* PA*, then there is a theory U of strictly intermediate consistency strength, $S < U < T$. Indeed, there are two such theories U and U', both strictly between S and T, but with incomparable consistency strength to each other.*

Proof. Suppose that S is weaker than T in consistency strength, with both theories extending PA. It follows that $\mathrm{PA} + \mathrm{Con}(S) + \neg\,\mathrm{Con}(T)$ is consistent. Let δ be the Rosser sentence of this theory, asserting that for any proof of δ from this theory, there is a smaller proof of $\neg\delta$. The usual Rosser argument shows that δ is neither provable nor refutable in this theory.

Because δ is not provable, we have a model of $\mathrm{Con}(S) + \neg\,\mathrm{Con}(T) + \neg\delta$. In light of what $\neg\delta$ asserts, this model thinks there is proof of δ with no smaller proof of $\neg\delta$. Since this concrete situation can be verified by PA, the model thinks that PA proves $\neg\delta$, and so this is a model of $\mathrm{Con}(S + \neg\delta) + \neg\,\mathrm{Con}(S + \delta) + \neg\,\mathrm{Con}(T)$.

Conversely, because δ is not refutable, the theory $\mathrm{PA} + \mathrm{Con}(S) + \neg\,\mathrm{Con}(T) + \delta$ is consistent, and so there is a model of this theory, and we may furthermore assume that the model thinks this theory is inconsistent. So this model will think that $\neg\delta$ is provable from $\mathrm{PA} + \mathrm{Con}(S) + \neg\,\mathrm{Con}(T)$, and since δ is true here, it must be that the smallest proof of $\neg\delta$ has no smaller proof of δ. The model thinks that PA can prove these concrete syntactic facts, which suffice to validate δ, and so the model thinks that PA proves δ. Therefore $\mathrm{Con}(S)$ implies $\mathrm{Con}(S + \delta)$ and $\neg\,\mathrm{Con}(S + \neg\delta)$ here. So this is a model of $\mathrm{Con}(S + \delta) + \neg\,\mathrm{Con}(S + \neg\delta) + \neg\,\mathrm{Con}(T)$.

To prove the theorem, let U be the theory $T \vee (S + \delta)$, meaning the theory with all sentences of the form $\tau \vee (\sigma \wedge \delta)$, for any $\tau \in T$ and $\sigma \in S$. The models of U are precisely the models of T and the models of $S + \delta$, and so $\mathrm{Con}(U)$ is simply the disjunction $\mathrm{Con}(T) \vee \mathrm{Con}(S + \delta)$. Similarly, let U' be the theory $T \vee (S + \neg\delta)$, which has $\mathrm{Con}(U') = \mathrm{Con}(T) \vee \mathrm{Con}(S + \neg\delta)$. It follows easily that U and U' are both at least weakly intermediate in consistency strength, $S \leq U, U' \leq T$.

We complete the proof by showing that U and U' are incomparable in consistency strength, which also implies they are both strictly intermediate between S and T. For this, observe that the first model above had $\mathrm{Con}(S + \delta)$ and hence $\mathrm{Con}(U)$, but neither $\mathrm{Con}(S + \neg\delta)$ nor $\mathrm{Con}(T)$ and hence not $\mathrm{Con}(U')$. The second model, in contrast, had $\mathrm{Con}(S + \neg\delta)$ and hence $\mathrm{Con}(U')$, but neither $\mathrm{Con}(S + \delta)$ nor $\mathrm{Con}(T)$ and hence not $\mathrm{Con}(U)$. So neither $\mathrm{Con}(U)$ nor $\mathrm{Con}(U')$ provably implies the other, and so these theories are incomparable in consistency strength.   □

Let me conclude this section by presenting H. Friedman's [Fri98] method of establishing density and incomparability in the hierarchy of consistency strength, proving it as a consequence of density in the simpler context of the derivability hierarchy, that is, in the Lindenbaum algebra, which for Π_1^0 assertions, he proves, is isomorphic to the hierarchy of consistency strength. Recall the *Lindenbaum algebra*

over a base theory T, the algebra of derivability (or direct implication), for which $\sigma \leq \tau$ when $T \vdash \tau \rightarrow \sigma$. This is a Boolean algebra modulo the induced equivalence of provable equivalence. The strict order $\sigma < \tau$ holds when $\sigma \leq \tau$ but $\tau \not\leq \sigma$.

Lemma 5. *The Lindenbaum algebra over any computably enumerable base theory T extending* PA *is dense. That is, if $\sigma < \tau$ with respect to derivability over T, then there is a sentence θ with $\sigma < \theta < \tau$. If σ and τ are Π_1^0, then there is a Π_1^0 such sentence θ.*

Proof. Since $\sigma < \tau$, the theory $T + \sigma + \neg\tau$ is consistent. Since this provides a consistent computably axiomatizable theory of arithmetic, it is incomplete. Let ρ be any statement independent of this theory, such as the Rosser sentence, which has complexity Π_1^0. Let $\theta = \sigma \wedge (\tau \vee \rho)$. So $\sigma \leq \theta \leq \tau$. But since we have models of $T + \sigma + \neg\tau + \rho$ and $T + \sigma + \neg\tau + \neg\rho$, which are therefore models of $T + \sigma + \neg\tau + \theta$ and $T + \sigma + \neg\tau + \neg\theta$, respectively, it follows that $\sigma < \theta < \tau$, as desired. Note that if σ and τ have complexity Π_1^0, then so does θ. $\square$

Next, Friedman transfers the consistency strength hierarchy to the Lindenbaum algebra simply by observing that every sufficient Π_1^0 statement is a consistency statement.

Lemma 6. *If θ is any Π_1^0 sentence with $\mathrm{Con}(\mathrm{PA}) \leq \theta$, then θ is* PA-*provably equivalent to $\mathrm{Con}(\mathrm{PA} + \eta)$ for some Π_1^0 sentence η.*

Proof. Using the fixed-point lemma, we can form a sentence η that asserts of itself, "for any counterexample to θ, there is a smaller number coding a proof of $\neg\eta$." Assume $\mathrm{PA} + \theta$, and hence also $\mathrm{Con}(\mathrm{PA})$. If $\mathrm{Con}(\mathrm{PA} + \eta)$ fails, then there is a proof of $\neg\eta$ from PA. Since θ holds, this proof is coded below any counterexample to θ, and so we can also prove η, contrary to $\mathrm{Con}(\mathrm{PA})$. So $\mathrm{Con}(\mathrm{PA} + \eta) \leq \theta$. Conversely, if θ fails yet $\mathrm{Con}(\mathrm{PA} + \eta)$ holds, then there is a counterexample to θ, but no smaller proof of $\neg\eta$, and from this we can prove $\neg\eta$, contrary to assumption. $\square$

The argument works not just with PA, but with any sufficient base theory. This lemma has a consequence I find surprising, namely, that every computably enumerable theory extending the base theory is equiconsistent with an individual sentence, since $\mathrm{Con}(T)$ is a Π_1^0 sentence and implies $\mathrm{Con}(\mathrm{PA})$, and so by the lemma $\mathrm{Con}(T)$ is equivalent to $\mathrm{Con}(\mathrm{PA} + \eta)$ for some sentence η. For example, even though the theory ZFC is not finitely axiomatizable, nevertheless there is an arithmetic sentence ζ that is equiconsistent with ZFC over PA.

Using lemma 6, Friedman deduces the following consequence, which shows how the fundamental nature of consistency strength must parallel that of direct implication.

Theorem 7. *The hierarchy of consistency degrees over* PA *is isomorphic to the derivability algebra over* PA *of Π_1^0 sentences at or above $\mathrm{Con}(\mathrm{PA})$.*

In particular, since the Lindenbaum algebra is dense by lemma 5, it follows that the consistency strength hierarchy also is dense; and since it is known that there are incomparable Π_1^0 statements (incomparable with respect to provable implication), it similarly follows that there are incomparable consistency strengths.

2. Natural instances of consistency-strength nonlinearity

Nobody likes the examples of nonlinearity and illfoundedness provided in section 1. Those sentences are viewed as unnatural—weird self-referential logic-game trickery. Therefore, let me now embark in earnest on the hard task I have set out for myself. Everyone seems to think it is impossible, but I shall try my best.

> **The hard task.** To provide natural, or at least nearly natural, instances of nonlinear incomparability and ill-foundedness in the hierarchy of consistency strength, particularly in the hierarchy of large-cardinal consistency strength.

To begin, consider the assertions that there are some finite number of inaccessible cardinals.

> "There are n inaccessible cardinals."

As n increases, these statements increase strictly in consistency strength. But there is a subtlety here, concerning how we describe the number n. Must we write out n as $1 + 1 + \cdots + 1$? That would be odd to insist upon if n were very large.

Suppose that we had said instead that

> "The number of inaccessible cardinals is at least the number of prime pairs."

This sentence surely makes a large cardinal existence assertion, but since it is an open question exactly how many prime pairs there are or whether there are infinitely many, perhaps the consistency strength of this assertion is not clear. Even in more concrete cases such as "there are 2^{100} inaccessible cardinals," we would ordinarily describe this number in effect by providing a method of computing it—multiply 2 by itself 100 times. More generally, we might want to say that "there are n inaccessible cardinals," where n is defined as the output of a certain explicitly provided computational procedure. Allowing such statements, however, opens the door wide to nonlinearity in the hierarchy of large cardinal consistency strength.

Theorem 8. *Amongst the large cardinal existence assertions of the form,*

> *"There are n many inaccessible cardinals,"*

where n is the output of a specific computational processs, there are instances of incomparable consistency strength. Indeed, there is a computable function f for which the statements

> *"there are f(n) inaccessible cardinals"*

are strongly independent with respect to consistency strength.

A list of statements is strongly independent in consistency strength when no nontrivial implication is provable between Boolean combinations of the consistency statements. This is equivalent to asserting that the consistency statements freely generate the free countably infinite Boolean algebra.

The proof of theorem 8 will rely on an elementary variant of the universal algorithm, as follows.

Theorem 9. *For any consistent theory T extending* ZFC, *there is a Turing machine program e, which we can write down, such that for any partial function $f : \mathbb{N} \to \mathbb{N}$, there is a model of the theory, such that if we run the program e on input n inside the model, then for $n \in \mathrm{dom}(f)$ the result is $f(n)$, but if $n \notin \mathrm{dom}(f)$, then the program does not halt.*

Proof. The program e searches for a proof from T of a statement of the form "the function computed by e is not precisely the function determined entirely by these specific input/output pairs: $(k_0, n_0), \ldots, (k_r, n_r)$." If such a proof is found, then the program proceeds to halt on exactly those inputs k_i giving output n_i; for inputs not of the form k_i on that list, the program loops endlessly. We use the Kleene recursion theorem in order to know that indeed there such a program e defined by this self-referential recursion.

In the standard model, I claim, the program will never halt. If it ever did halt, then it will have done so because it found such a proof, but then proceeded to halt anyway on exactly those inputs with exactly those outputs. By inspecting the computation, we would be able to prove this is the behavior, and this would show that T is inconsistent. So in the standard model, there are no such proofs to be found.

But precisely because of this, it follows that for any particular desired finite list of input/output behavior $(k_0, n_0), \ldots, (k_r, n_r)$, it is consistent with T that the program e halts on these k_i with output n_i, and diverges on all other input. Therefore, for any partial function $f : \mathbb{N} \to \mathbb{N}$, it is finitely consistent with T that the operation of e is in accordance with f. And so the whole theory is consistent. So there is a model $M \models T$ in which the function computed by e is altogether in accordance with f on standard input n. $\qed$

Kindly notice that if in some model of arithmetic the program e should happen to halt on one input n and not on another input m, then this will be provable by PA in the model. The reason is that if the program halts on n, then it was because a certain proof was found involving an input/output lookup table, which did not include m. And so in that model PA will be able to prove that the program does not halt on m.

Theorem 9 is part of a long history, perhaps tracing back to Mostowski [Mos60] and Kripke [Kri62]. The particular proof above follows the presentation on my blog post [Ham16b], following a suggestion of Vadim Kosoy in the comment section there. The theorem is mainly to be viewed, however, as a baby version of the universal algorithm theorem, due to W. Hugh Woodin [Woo11]; but see my simplified proof in [Ham18], and also [Ham17a; Ham17c], [BE17; Bla17]. Shavrukov made similar arguments in 2012 private communications under the slogan "On risks of accruing assets against increasingly better advice," and he pointed out connections with [Ber90] and [Jap94]. Albert Visser has pointed out a similar affinity with the classical proof-theoretic 'exile' argument (for example, see 'refugee' in [AB04]). Further set-theoretic analogues of the universal algorithm have been explored in [HW17; HW21]. The main difference between theorem 9 and the full universal algorithm (as in [Ham18] etc.) is the remarkable extension property of the full version. Namely, the algorithm always produces a finite sequence, but in any model of arithmetic M in which the sequence produced is s, then for any finite extension t of that sequence in M, there is an end-extension of M to a model N in which the computed sequence is t. Since I haven't seemed to need this extension property for the applications of this article, I have proceeded with just the baby version of theorem 9.

Proof of theorem 8. Let us now use the universal computable function to establish nonlinearity in the large cardinal hierarchy. Let f be the universal computable

function, computed by the algorithm as described in theorem 9 using any consistent theory T extending ZFC plus the claim that there are infinitely many inaccessible cardinals. Let σ_n be the statement, "there are $f(n)$ many inaccessible cardinals." We interpret the statement as asserting also that $f(n)$ exists. We want to show that $\mathrm{Con}(\mathrm{ZFC}+\sigma_n)$ does not ZFC-provably imply $\mathrm{Con}(\mathrm{ZFC}+\sigma_m)$ for any $n \neq m$. For this, we need a model of ZFC satisfying the first consistency statement, but not the second. By the universality theorem 9, there is a model M of T in which $f(n) < f(m)$ and both are defined. Since the model M has what it thinks is infinitely many inaccessible cardinals, it also has at least $f(n)+1$ many. By cutting the universe to V_κ at the $(f(n)+1)$th inaccessible cardinal, the model thinks that the theory "ZFC$+f(n) < f(m)$ and there are exactly $f(n)$ many inaccessible cardinals" is consistent. Since this theory does not prove its own consistency, there is a model U inside M that thinks $f^M(n)$ many inaccessible cardinals are consistent with ZFC, but not $f^M(n)+1$ many. That model must have $\mathbb{N}^U$ end-extending $\mathbb{N}^M$, and so the operation of e on n and m will agree with M, and so $f^U(n) = f^M(n) < f^U(m) = f^M(m)$. So U is a model of ZFC in which $\mathrm{Con}(\mathrm{ZFC}+\sigma_n) + \neg\,\mathrm{Con}(\mathrm{ZFC}+\sigma_m)$, as desired. □

The same argument works with Mahlo cardinals, measurable cardinals, and similarly with almost any of the usual large cardinal notions.

Perhaps one is tempted to object to theorem 8 on the grounds that the program computing f is not provably total, and this could cause ambiguity in the meaning of the statement "there are $f(n)$ inaccessible cardinals" if the computation of $f(n)$ does not halt. In theorem 8, I had interpreted the statement as false in that case. Indeed, $f(n)$ is defined only in models of $\neg\,\mathrm{Con}(T)$, which might be considered rather strange to consider if we are interested in T as an aspirational large cardinal theory.

This objection is easily addressed, however, simply by using slightly different statements. Namely, let us consider the sentences asserting "there are as many measurable cardinals as the running time of program e on input n." This sentence is naturally interpreted whether or not the program halts on input n; the number of asserted measurable cardinals will be finite if e halts on n and otherwise infinite. In the natural case for the die-hard large cardinal set theorist, therefore, who consider both T and $\mathrm{Con}(T)$ to be true, the sentences will all assert the existence of infinitely many inaccessible cardinals. And yet, still they will have incomparable consistency strengths.

Theorem 10. *Amongst the large cardinal existence assertions of the form,*

> *"There are as many measurable cardinals as the running time of this specific computational process,"*

there are instances with incomparable consistency strength. Indeed, there is a computable function f for which the statements about the running time of the computation of $f(n)$ are strongly independent with respect to consistency strength.

Proof. We proceed essentially the same as in theorem 8. Let e be the program for the universal function f of theorem 9 defined relative to the theory $T = \mathrm{ZFC}+$"there are infinitely many measurable cardinals," and let τ_n be the statement "there are as many measurable cardinals as the running time of e on input n." If $n \neq m$, then there is a model $M \models T$ in which e halts on n, but not on m. If t is the running time of e on n, then we may cut off M at the $(t+1)$th measurable cardinal to see

that the theory "ZFC + e halts on n in t steps but fails to halt on input m and there are exactly $t+1$ many measurable cardinals" is consistent. Since this theory does not prove its own consistency, there is a model U inside M that thinks t many measurable cardinals are consistent with ZFC, but not $t+1$ many. That model must have a $\mathbb{N}$ end-extending $\mathbb{N}^M$, and so the operation of e on n will still halt by time t and m will not. So U is a model of ZFC in which $\mathrm{Con}(\mathrm{ZFC}+\tau_n)+\neg\,\mathrm{Con}(\mathrm{ZFC}+\tau_m)$, as desired. □

An exactly similar argument will apply with statements of the form

"There are as many supercompact cardinals as the sizes of squares in the plane that can be covered by tilings using this specific set of polygonal tiles,"

and similarly with many other such kinds of assertions, using any provably m-complete decision problem.

3. Natural arithmetic instances of nonlinearity in consistency strength

A careful reader will recognize that many of the arguments of the previous section have little actually to do with large cardinals, and the essential ideas can be used to exhibit natural instances of ill-foundedness and incomparability in the consistency strengths of simple arithmetic assertions.

Theorem 11. *Amongst the assertions of the form,*

"This specific computational process halts,"

there are instances of incomparable consistency strength. Indeed, there are instances whose consistency strengths are double-jumping. Furthermore, there is a program e for which the statements "e halts on input n" are strongly independent with respect to consistency strength.

Proof. Since every Σ^0_1 statement is provably equivalent to the halting of a certain computational process, the existence of double-jumping statements of this form is an immediate consequence of theorem 3, which provided a Π^0_1 sentence η, so $\neg\eta$ is Σ^0_1.

But let me also argue directly, using the universal computable function f described in theorem 9. The universal property of f implies that it is consistent with LC that e halts exactly on any desired finite set of n and only on those numbers. Consider any $n \neq m$. There is a model of LC in which e halts on n, but not on m. In this case, the model thinks that it is consistent with PA that e halts on n, but inconsistent that it does so on m, since the reason that it halted on n was because of a specific list of numbers, which did not include m, and in the model, PA can prove that that list is what arises in the computation of f. So we have a model of $\mathrm{PA} + \mathrm{Con}(\mathrm{PA} + e\text{ halts on }n) + \neg\,\mathrm{Con}(\mathrm{PA} + e\text{ halts on }m)$. And similarly in the model where e halts only on m. So these two statements have incomparable consistency strengths.

We can argue similarly that these statements are strongly independent. The consistency assertions for any nontrivial Boolean combination of the statements is determined by the halting pattern of e on the finite set of numbers m mentioned in the expression. And for any n that is not mentioned, we can make models where the original pattern of halting is the same, except that e also halts on n, or does not

halt on n, respectively. So both the consistency of this statement or its negation can be added to the expression consistently, as desired.   □

I should like specifically to call attention to the fact that the assertions used in theorem 11 express instances of halting, rather than non-halting. It is an easy matter to show that every consistency assertion $\mathrm{Con}(T)$ for a computably enumerable theory T is equivalent to a statement asserting that a certain program does *not* halt on a certain input—we need only consider the algorithm that searches for a proof of a contradiction in T and halts when such a proof is found. So the assertion that this program does not halt is provably equivalent to $\mathrm{Con}(T)$, and therefore has consistency strength strictly exceeding T.

A similar observation reveals the tight intermingling of the two meanings of 'undecidable,' namely, the undecidability of a decision problem versus the undecidability of a sentence in a theory.

Theorem 12. *Suppose that A is a computably enumerable nondecidable decision problem.*

(1) *For any consistent theory extending* PA*, there are true instances of $n \notin A$ that are independent of that theory.*

(2) *Furthermore, the true assertions of $n \notin A$ are not bounded in consistency strength by any consistent theory, including any consistent large cardinal hypothesis.*

Proof. Suppose A is a computably enumerable nondecidable decision problem, and consider any consistent extension S of PA. If all true instances of $n \notin A$ were provable in S, then we would have a decision procedure for A. Namely, by day we run the enumeration algorithm of A, learning more of the positive instances; by night, we search for proofs in S that $n \notin A$, for the negative instances. Since S is consistent and PA proves any true instance of halting, it follows that S can be trusted when it proves $n \notin A$. Since A is undecidable, this process must fail as a decision procedure, and so there are $n \notin A$ for which this is not provable in the theory S.

Consider now any consistent computably enumerable theory T, such as any consistent large cardinal theory. By applying the previous observation with the theory $\mathrm{PA} + \mathrm{Con}(T)$, we see that there are instances $n \notin A$ for which this is not provable in $\mathrm{PA} + \mathrm{Con}(T)$. So there is a model of PA in which $\mathrm{Con}(T) + (n \in A)$. Since $n \in A$ implies that PA proves it, the model thinks that $\mathrm{PA} + (n \notin A)$ is inconsistent. So this is a model of PA in which T is consistent, but $\mathrm{PA} + (n \notin A)$ is not, showing that the consistency strength of the assertion $n \notin A$ is not bounded by T, as claimed.   □

Thus, every computably enumerable computably undecidable set is saturated with logical undecidability and nontrivial consistency strength. By (2), the true assertions of $n \notin A$ can have no largest instance of consistency strength. From this, it follows that either there is incomparability in consistent strength amongst these statements, or else the statements form a linear hierarchy of consistency strength, not bounded above by any given consistency strength.

Recall that a computably enumerable set $A \subseteq \mathbb{N}$ is said to be *m-complete*, if for every computably enumerable set B there is a computable total function $f : \mathbb{N} \to \mathbb{N}$ such that $b \in B \leftrightarrow f(b) \in A$. For example, the halting problem, the word problem

for groups, the nontiling problem, and many other problems are well known to be m-complete. Meanwhile, observation 15 below will show that it is a strictly stronger hypothesis about a program that it enumerates a PA-*provably* m-complete set, and strictly stronger still to assume that the enumerated set admits specific provable reductions f from the familiar m-complete sets.

Theorem 13. *Suppose that A is any computably enumerable m-complete decision problem.*

(1) *If A is PA-provably m-complete, with a provable reduction of the halting problem, then amongst the true statements $n \notin A$, there are instances strictly exceeding any given consistency strength.*

(2) *But even without that extra provability assumption, amongst the true assertions $n \notin A$ there are instances with incomparable consistency strength, and instances of double-jumping consistency strength.*

(3) *Amongst the consistent statements $n \in A$, there are instances of incomparable consistency strength and of double-jumping consistency strength.*

(4) *Within both kinds of statements $n \notin A$, $n \in A$, there are effective enumerations that are strongly independent in consistency strength.*

Proof. For statement (1) only, we assume that A is provably m-complete, with a provable reduction from the halting problem. For any computably enumerable theory T, let e_T be the program that searches for a proof of a contradiction in T, halting only when found. By the reduction of the halting problem to A, we get a number n such that e_T halts if and only if $n \in A$, provably in PA. So $\text{Con}(T)$ is provably equivalent to the assertion $n \notin A$, which therefore has consistency strength strictly exceeding T.

For the rest of the statements, let us now drop the extra assumption about provable m-completeness. We now assume only that A is actually computably enumerable and m-complete. Let π be a specific computable function that is a reduction to A of the halting problem. Consider a version of the universal computable function f, as in theorem 9, but defined relative to the theory $\text{PA} + \text{Con(ZFC)} + \text{"}\pi$ is a reduction of the halting problem to A." The proof of theorem 9 shows that there is a computable procedure e such that for any $n \neq m$, there is a model of this theory in which e halts on n but not m, and another in which e halts on m and not n. In the first case, we'll have $\pi(e, n) \in A$, $\pi(e, m) \notin A$ and in the second case vice versa. Because of the way e is defined, if it halts on one number but not another, then in such a model it will be thought inconsistent that it could halt on the other number. Thus, the statements of the form $\pi(e, n) \notin A$ are incomparable in consistency strength, and indeed these statements form a strongly independent family with respect to consistency strength, just as in theorem 11. Furthermore, any one of these statements is double-jumping, because, once $\pi(e, n)$ is in A, then it is inconsistent for it to be out of A, and if it is out of A, but another number is in A, then it is inconsistent to be in A. This establishes statement (2), and the part of 4 for assertions of the form $n \notin A$.

The rest of the claims, statement (3) and the part of (4) for $n \in A$, are proved by essentially the same argument, which exhibits a certain symmetry as to whether we were asserting membership or nonmembership of $\pi(e, n)$ in A.   □

Corollary 14.

(1) *Amongst assertions of the form*

> *"This specific set of polygonal tiles admits a tiling of the plane"*
>
> *there are consistent instances strictly exceeding any given consistency strength.*

(2) There are such assertions with incomparable consistency strength.

(3) There are such assertions having incomparable consistency strength with their own negations—both the statement and its negation jump in consistency strength.

(4) There is an effective enumeration of finite tile sets $t_0, t_1, \ldots$, such that the assertions "tile set t_n admits a tiling of the plane" have strongly independent consistency strengths.

Proof. The tiling problem is provably m-complete, with a provable reduction of the halting problem. □

We can just as easily provide corresponding results for statements of the form

> "this specific diophantine equation $p(\vec{x}) = 0$ has no solution in the integers,"

where p is explicitly provided as a polynomial over the integers. Such statement are not bounded in consistency strength by any consistent theory, and there are specific instances with incomparable consistency strength and double-jumping consistency strength. There is a particular integer polynomial $p(\vec{x}, y)$ for which the statements "$p(\vec{x}, n) = 0$ has a solution in the integers," as n varies, have strongly independent consistency strengths.

We can provide similar results for assertions such as

> "this specific finite group presentation is the trivial group"

or

> "this cell in this specific Game of Life position will eventually become alive."

And so on. In each case, there are specific such statements with incomparable consistency strength, with double-jumping consistency strength, and schemes of such statements with strongly independent consistency strength.

What I take these arguments to show is that there is pervasive nonlinearity at every level of the consistency-strength hierarchy, and this is a direct consequence of the difficulty of interpreting even the names of natural numbers. We can describe numbers, even specifying them by concrete computable procedures, but our base theory may just not settle the question of which number is larger, and this leads directly and almost immediately to incomparable consistency strengths.

Meanwhile, regarding the difference between being m-complete and being provably m-complete, or having provable reductions of specific computably enumerable sets, let me prove that indeed these are not the same.

Observation 15.

(1) There are m-complete computably enumerable sets that are not PA-provably m-complete.

(2) There are PA-provably m-complete sets A, for which none of the usual m-complete sets B have a specific reduction function that is PA-provably a reduction of B to A.

Proof. The halting problem is, of course, a PA-provably m-complete computably enumerable set. Consider the computably enumerable set A, which enumerates

the halting problem, unless it happens to find a proof of a contradiction in some
fixed strong theory T, in which case it enumerates every number into A, spoiling
the set. If T is actually consistent, then this algorithm will enumerate the halting
problem, which is m-complete. But this set is not PA-provably complete, since it
is consistent with PA that $\neg \operatorname{Con}(T)$, in which case the set will be all of $\mathbb{N}$ and
therefore not m-complete. This proves statement (1).

For statement (2), let us simply modify the previous algorithm, so that when
the proof of a contradiction in T is found, the program enumerates all numbers
up to the code of that proof into the set, but then starts over above this point by
enumerating a shifted copy of the halting problem. In any model of PA, this set
will either be the halting problem or a shifted copy of the halting problem (with
a filled-in block below), and hence m-complete. So we can prove in PA that this
algorithm enumerates an m-complete computably enumerable set. But there will
be no particular reduction function from the usual halting problem (or any of the
other commonly considered m-complete sets) that we can prove in PA is a reduction
to A, since it is consistent with PA that every standard number is in A, and so we
will get wrong answers for the instances of provable nonhalting programs in such a
model.   □

It it very natural to inquire whether one can prove theorem 13 under the weaker
assumption only that A is computably enumerable and undecidable, rather than m-
complete. Theorem 12 gets part of the result, but does not achieve incomparability
in the consistency assertions.

Question 16. *Does every computably enumerable noncomputable set A admit
statements $n \notin A$ of incomparable consistency strength?*

The answer is no, proved by Uri Andrews after a talk I gave on the topics of this
article for the Madison Logic Seminar.

Theorem 17 (Uri Andrews). *There is a c.e. undecidable set A, which is Turing
equivalent to the halting problem, but for which the assertions $n \notin A$ are linearly
ordered by consistency strength. Also the assertions $n \in A$ are linearly ordered by
consistency strength.*

Proof. For any c.e. undecidable set $I \subseteq \mathbb{N}$, let $r = \sum_{n \in I} \frac{1}{2^n}$ be the real number
having binary bit 1 in the places of I. Let A be the rational cut determined by r,
the set $\{\, q \in \mathbb{Q} \mid q < r \,\}$. Let us assume we have naturally encoded rational numbers
with natural numbers. The set A is c.e., because as natural numbers are enumerated
into I, we learn improved lower bounds to r and can enumerate the corresponding
rational numbers into A. The set A is Turing equivalent to I, since with I as an
oracle we can compute r well enough so as to answer definitively about any given
rational number; and conversely with A as an oracle we can compute I. So by
taking I to be Turing complete we shall have A Turing equivalent to the halting
problem.

The key thing to notice about A, however, is that for rational numbers, if $p < q$
and $q \in A$, then $p \in A$. Equivalently, $p < q$ and $p \notin A$ implies $q \notin A$, and this will
be provable in PA. So there is a provable linear relation about any two nonmembers
of A. The consequence of this is that the assertions $p \notin A$ and $q \notin A$ cannot be
incomparable in consistency strength, since any model of the former is necessarily
a model of the latter, and so $\operatorname{Con}(\mathrm{PA} + p \notin A)$ implies $\operatorname{Con}(\mathrm{PA} + q \notin A)$. So the

nonmembership assertions $p \notin A$ are linearly ordered by consistency strength. An essentially similar argument works with the positive assertions $p \in A$.   □

In light of theorem 12, it follows that the consistency strengths of the assertions $p \notin A$ for the set A mentioned in the proof of theorem 17 form a linear order of height ω. And the consistency strengths of the statements $p \in A$ form a linear hierarchy of order type ω^*.

4. Natural instances of ill-foundedness and incomparability via cautious theory enumerations

Let me now present another huge class of examples. Imagine that we believed in a certain computably enumerable theory, such as PA, ZFC, or ZFC plus large cardinals, whose axioms we intended to enumerate, but *cautiously*. By doing so, we can often enumerate the theory in a sensible, natural manner, but with a strictly lower consistency strength or with incomparable consistency strengths. These cautious enumeration theories therefore provide natural instances both of ill-foundedness and nonlinearity in the hierarchy of consistency strength.

As a general umbrella term, I shall say that one has a *cautious enumeration* of a theory, if one is enumerating the axioms of the theory according to a procedure that will continue as long as one hasn't encountered a certain kind of contrary indicator, a reason to doubt the truth of the theory. Different cautious enumerations of a theory will arise depending on the specific contrary indicator that is required to halt the enumeration.

I shall also use the term to refer to specific natural theories. For example, let the *cautious enumeration of* ZFC be the enumeration of ZFC that continues as long as we have not yet found a proof in what we have enumerated so far that ZFC is inconsistent. I denote the resulting theory by ZFC°. In order to halt the enumeration, we do not require an explicit contradiction in ZFC, but rather only a proof that there is such a contradiction. If Con(ZFC) is actually true, as we have assumed, and in particular, if Con(ZFC) is consistent with ZFC, then no confounding proof of inconsistency will ever be found. As a theory, therefore, the cautious enumeration ZFC° actually has all the same axioms as ZFC; they are the same theory, but with a different manner of enumeration. Thus, this situation reveals an intentional aspect of the consistency strength hierarchy, for the consistency strength of a theory T is not determined extensionally by the theory itself, by the set of sentences that are actually in T, but rather by how the theory has been described, in order to form the assertion Con(T).

I find the cautious enumeration to be both sensible and realistic—in this sense it is a natural theory—for if we were actually enumerating ZFC and a proof was pointed out to us along the way that the theory we have already committed to proves the full ZFC theory inconsistent, then we would have ample reason to pause and reflect on whether we should continue with the enumeration. I think we would pause the enumeration right then and reconsider. The cautious enumeration is what we would actually do—so it is a natural theory.

Theorem 18. *The cautious enumeration* ZFC° *is an alternative computable enumeration of* ZFC, *with a strictly lower consistency strength than* ZFC.

Proof. We have already explained that if $\mathrm{Con}(\mathrm{ZFC})$ is consistent with ZFC, then the cautious enumeration ZFC° will never encounter the confounding proof of inconsistency, and so it will fully enumerate all the ZFC axioms. Meanwhile, I claim that the cautious theory ZFC° has a strictly weaker consistency strength than ZFC. We can easily prove, of course, that $\mathrm{ZFC}^\circ \subseteq \mathrm{ZFC}$ and consequently that $\mathrm{Con}(\mathrm{ZFC}) \rightarrow \mathrm{Con}(\mathrm{ZFC}^\circ)$. But if $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})$ is consistent, then there is a model of the theory $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}) + \neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}))$. In such a model M, while ZFC is consistent, nevertheless there will be a proof from ZFC that it is not, and so M will think that ZFC° consists of only finitely many axioms of ZFC. Since we can prove even in PA that ZFC proves each of its finite subsets is consistent, it follows that M thinks that ZFC proves $\mathrm{Con}(\mathrm{ZFC}^\circ)$. By the incompleteness theorem, it thinks $\mathrm{ZFC} + \neg\,\mathrm{Con}(\mathrm{ZFC})$ is consistent, and so it can build a model of ZFC in which $\mathrm{Con}(\mathrm{ZFC}^\circ)$ holds, but not $\mathrm{Con}(\mathrm{ZFC})$. So the cautious theory ZFC° is strictly weaker than ZFC in consistency strength, even though they enumerate the same theory. $\square$

It might be a little surprising that a theory can have a strictly weaker consistency strength over a base theory than the base theory itself. But it shouldn't be too surprising, since there is room between ZFC and $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})$ for other consistency statements, and that is what is going on here. The theorem amounts to the claim that $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}^\circ)$ is a strictly weaker theory than $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})$.

Consider next the *doubly cautious* enumeration $\mathrm{ZFC}^{\circ\circ}$, where we enumerate the ZFC axioms as usual, but continue only as long as we have not yet found a proof in ZFC that ZFC is inconsistent or even a proof in ZFC that there is such a proof of inconsistency. In other words, we stop the enumeration when we find a proof from ZFC either of $\neg\,\mathrm{Con}(\mathrm{ZFC})$ or of $\neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}))$.

Theorem 19. *The doubly cautious enumeration of* ZFC *is an alternative computable enumeration of* ZFC*, with strictly weaker consistency strength than even the cautious enumeration.*

Proof. Since $\mathrm{ZFC}^{\circ\circ}$ is a subtheory of ZFC°, we easily prove in ZFC that $\mathrm{Con}(\mathrm{ZFC}^\circ)$ implies $\mathrm{Con}(\mathrm{ZFC}^{\circ\circ})$. To show the implication is not reversible, consider a model of $\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}) + \mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})) + \neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})))$. In this model, because $\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}))$ holds, the cautious enumeration will never find a proof of $\neg\,\mathrm{Con}(\mathrm{ZFC})$ from ZFC, and so it will enumerate the full theory, $\mathrm{ZFC}^\circ = \mathrm{ZFC}$. But because the models believes $\neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC})))$, it will find a proof from ZFC of $\neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(\mathrm{ZFC}))$. In other words, it will find a proof from ZFC that there is a proof from ZFC that ZFC is inconsistent. Thus, the doubly cautious enumeration will find its stopping point, and so M thinks $\mathrm{ZFC}^{\circ\circ}$ is a finite fragment of ZFC. So it thinks that ZFC° proves $\mathrm{Con}(\mathrm{ZFC}^{\circ\circ})$. Since by $\mathrm{Con}(\mathrm{ZFC})$ it thinks $\mathrm{ZFC}^\circ + \neg\,\mathrm{Con}(\mathrm{ZFC}^\circ)$ is consistent, it can therefore build a model of ZFC in which $\mathrm{Con}(\mathrm{ZFC}^{\circ\circ})$ holds, but not $\mathrm{Con}(\mathrm{ZFC}^\circ)$. So the doubly cautious theory $\mathrm{ZFC}^{\circ\circ}$ is strictly weaker in consistency strength than the cautious theory ZFC°. $\square$

We can obviously iterate this to the triply cautious enumeration, and so on, resulting in an effective list of alternative enumerations of ZFC, all enumerating exactly the same full ZFC theory, but doing so with progressively weaker consistency strengths.

$$\cdots < \mathrm{ZFC}^{\circ\circ\circ} < \mathrm{ZFC}^{\circ\circ} < \mathrm{ZFC}^\circ < \mathrm{ZFC}$$

Thus, we have ill-foundedness in consistency strength even amongst the enumerations of the same ZFC theory. And all these enumerations are quite sensible, since if we believe in ZFC, then the discovery that it is provably inconsistent or that it is provably provably inconsistent would surely make us pause and reflect. So I find this to be a natural instance of ill-foundedness in the hierarchy of consistency strength.

One might also be curious about another slightly less cautious enumeration of ZFC, what I call the *stop-when-hopeless* enumeration $\text{ZFC}^{\otimes}$, where we enumerate ZFC as usual, but continue adding axioms only so long as we have not found any explicit contradiction yet in what has been enumerated. If ZFC is consistent, then no such contradiction will ever be found, and so the resulting theory $\text{ZFC}^{\otimes}$ has all the same axioms as ZFC. Is this theory also strictly weaker in consistency strength than ZFC? The answer is no, I claim, the stop-when-hopeless enumeration $\text{ZFC}^{\otimes}$ is equiconsistent with ZFC. Since $\text{ZFC}^{\otimes}$ is provably a subtheory of ZFC, we can easily prove that $\text{Con}(\text{ZFC}) \rightarrow \text{Con}(\text{ZFC}^{\otimes})$. Conversely, if we are in a model of ZFC in which $\text{ZFC}^{\otimes}$ is consistent, then the enumeration process must not ever have stopped, and so it is the same theory as ZFC. So in ZFC, we can prove $\text{Con}(\text{ZFC}) \leftrightarrow \text{Con}(\text{ZFC}^{\otimes})$. So the stop-when-hopeless enumeration offers no advantage in consistency strength.

In general, for any theories S and T, we may consider the *S-cautious enumeration of T*, where we enumerate the axioms of T as long as we have not yet seen an explicit contradiction in S. In this terminology, ZFC° is the $\text{ZFC} + \text{Con}(\text{ZFC})$-cautious enumeration of ZFC, and $\text{ZFC}^{\circ\circ}$ is the $\text{ZFC} + \text{Con}(\text{ZFC}) + \text{Con}(\text{ZFC} + \text{Con}(\text{ZFC}))$-cautious enumeration of ZFC. The stop-when-hopeless enumeration $\text{ZFC}^{\otimes}$ is simply the ZFC-cautious enumeration of ZFC.

Meanwhile, let me now offer several additional alternative cautious enumerations of ZFC, each looking for a slightly different particular reason to halt the enumeration, each of them sensible and compelling. Yet the resulting theories have strictly incomparable consistency strengths, even though they all actually enumerate the same ZFC theory.

Theorem 20. *There are diverse alternative cautious enumerations of the* ZFC *theory, of incomparable consistency strengths to one another and hence all strictly weaker than* ZFC. *Indeed, there is an effective enumeration of theories* $\text{ZFC}^{(n)}$, *each actually having exactly the same axioms as* ZFC, *but whose consistency assertions are strongly independent, generating the countable free Boolean algebra.*

Proof. Let f be the universal computable function defined relative to the theory $\text{ZFC} + \text{Con}(\text{ZFC})$. This algorithm searches for a proof in that theory that f does not agree exactly with a certain finite list of input/output pairs, and when found, it computes exactly in accordance with that list, diverging on all other input. Let $\text{ZFC}^{(n)}$ be the theory that continues enumerating the ZFC axioms as long as $f(n)$ has not yet halted.

Each $\text{ZFC}^{(n)}$ is sensible, each cautious in a slightly different and independent way, for each of them enumerates the ZFC axioms provided a certain kind of contrary information is not encountered. Namely, the nth theory $\text{ZFC}^{(n)}$ is watching to see whether $f(n)$ halts, and if it does, then this is a perfectly sound reason to doubt the veracity of ZFC, because the halting of $f(n)$ occurs only when there is a proof that f shouldn't behave as it did, showing that $\text{ZFC} + \text{Con}(\text{ZFC})$ has false consequences;

on these grounds, $\text{ZFC}^{(n)}$ stops its enumeration. Meanwhile, since the program f does not actually ever halt (in the standard model $\mathbb{N}$), each theory $\text{ZFC}^{(n)}$ actually has all the same axioms as ZFC. The disagreements between these theories are merely theoretical possibilities, occurring only possibly in a nonstandard model and even then only on nonstandard-length axioms.

The theories are bounded above in consistency strength by ZFC, simply because each $\text{ZFC}^{(n)}$ is provably a subtheory of ZFC. But indeed, I claim that they are each strictly strictly weaker than ZFC in consistency strength, for if $n \neq m$, then I shall now prove that $\text{ZFC}^{(n)}$ and $\text{ZFC}^{(m)}$ have incomparable consistency strength over ZFC. Assume $n \neq m$. By the properties of the universal computable function, there is a model M of $\text{ZFC} + \text{Con}(\text{ZFC})$ in which $f(n)$ halts but $f(m)$ does not. In this model, $\text{ZFC}^{(n)}$ is a finite fragment of ZFC, but $\text{ZFC}^{(m)}$ is fully the same as ZFC. Since M thinks $\text{Con}(\text{ZFC})$, it can build a model N that it thinks satisfies $\text{ZFC} + \neg\,\text{Con}(\text{ZFC})$. Because the function f halted at a finite stage of M, the model N will agree with this, since $\mathbb{N}^M$ must be an initial segment of $\mathbb{N}^N$. So the computation of f on inputs n and m is the same in N as it is in M. Since we can prove even in PA that the theory ZFC proves the consistency of any of its finite fragments, it follows that M thinks $\text{Con}(\text{ZFC}^{(n)})$ holds in N. But since $\neg\,\text{Con}(\text{ZFC})$ also holds and $f(m)$ does not halt, it follows that N satisfies $\neg\,\text{Con}(\text{ZFC}^{(m)})$. So we have a model of ZFC which thinks $\text{ZFC}^{(n)}$ is consistent, but not $\text{ZFC}^{(m)}$. This shows that the consistency of $\text{ZFC}^{(n)}$ does not provably imply the consistency of $\text{ZFC}^{(m)}$ over ZFC, and so all the theories have incomparable consistency strength.

It is not much more difficult to show that these consistency statements are strongly independent and freely generate the free Boolean algebra. The point is that for any conjunction of the statements $\text{Con}(\text{ZFC}^{(n)})$ or their negations, then we can find a model where f halts on the ns for which $\text{Con}(\text{ZFC}^{(n)})$ occurs positively in the conjunction and not on the ns that occur negatively, and this will make the conjunction altogether true; or conversely we can find a model making it false. By controlling which n the universal algorithm halts on, we can control exactly which consistency statements are strong and which are weak, showing that every possible combination of those consistency statements is nontrivial. So these statements are strongly independent and therefore freely generate the free Boolean algebra. □

Let I be the assertion that there is an inaccessible cardinal.

Theorem 21. *The cautious enumeration of $\text{ZFC} + I$ has consistency strength strictly between ZFC and $\text{ZFC} + I$. Indeed, there is an effective enumeration of infinitely many cautious enumerations of this theory, with strongly independent incomparable consistency strengths, whose consistency statements freely generate the countable free Boolean algebra.*

Proof. Consider first the cautious enumeration $(\text{ZFC} + I)^\circ$ of the theory $\text{ZFC} + I$, where we enumerate the axioms into this theory, including I itself, up until a stage at which we see a proof in $\text{ZFC} + I$ that $\text{ZFC} + I$ is inconsistent. I should like to emphasize again that we don't insist on seeing an actual proof of a contradiction from this theory, but rather merely a proof that there is one. Since the theory is actually consistent, we have assumed, this enumeration will actually provide all of the axioms of the full theory $\text{ZFC} + I$. The difficulty is that we are not able to prove this in ZFC—we had made an extra consistency assumption to deduce it.

The theory ZFC+I proves Con(ZFC), since we can prove that if κ is inaccessible, then V_κ is a transitive model of ZFC. Since this proof can be made explicit, the proof can be made using axioms that appear before any supposed (nonstandard) proof that there is a contradiction. Therefore, Con(ZFC) will be provable in $(\text{ZFC} + I)^\circ$, provably so, and so this cautious theory is strictly stronger than ZFC in consistency strength. This argument extends to the iterated consistency statements Con(ZFC+ Con(ZFC)) and so on.

Meanwhile, since we have assumed that ZFC+I+Con(ZFC+I) is consistent, by the incompleteness theorem it is consistent with $\neg$ Con(ZFC+I+Con(ZFC+I)). In a model M of this theory, there will come a stage where ZFC+I proves $\neg$ Con(ZFC+ I), which will stop the cautious enumeration. So in M, the theory $(\text{ZFC} + I)^\circ$ is a finite fragment of $\text{ZFC} + I$. But we can prove in a weak meta theory that $\text{ZFC} + I$ proves the consistency of each of its finite fragments, and so this model thinks that ZFC+I proves Con($(\text{ZFC}+I)^\circ$). Since M has a model of ZFC+I+$\neg$ Con(ZFC+I), it will therefore have a model of $\text{ZFC} + I + \text{Con}((\text{ZFC} + I)^\circ) + \neg\,\text{Con}(\text{ZFC} + I)$. Therefore, we cannot prove in ZFC, nor even in $\text{ZFC} + I$, that the consistency of the cautious theory $(\text{ZFC}+I)^\circ$ implies the consistency of ZFC+I. So it has strictly intermediate consistency strength.

For the incomparability result, let f be the universal computable function, defined relative to the theory $\text{ZFC} + I + \text{Con}(\text{ZFC} + I)$. That is, f looks for a proof in this theory that the input/output pattern of f does not agree with some explicit finite list, and when found, the function f halts or diverges on its input exactly in accordance with that list. Let $(\text{ZFC} + I)^{(n)}$ be the cautious enumeration in which one enumerates the axioms of $\text{ZFC} + I$ until a stage at which $f(n)$ halts, at which point the enumeration also halts. This is sensible, since the halting of $f(n)$ is a sensible reason not to trust $\text{ZFC} + I$, since the assertion that it was consistent led to false statements.

Meanwhile, I claim these theories are incomparable in consistency strength, and strongly independent. For any pair $n \neq m$, there is a model $M \models \text{ZFC} + I + \text{Con}(\text{ZFC} + I)$ in which $f(n)$ halts but $f(m)$ does not, and provably so. In such a model, the theory $(\text{ZFC}+I)^{(n)}$ is a finite fragment of ZFC+I, whereas $(\text{ZFC}+I)^{(m)}$ is provably the full theory ZFC+I. By the same method as earlier, therefore, we can find a model of $\text{ZFC} + I$ in which Con($(\text{ZFC}+I)^{(n)}$) holds, but Con($(\text{ZFC}+I)^{(m)}$) does not. So the two theories have incomparable consistency strength, even over ZFC+I. Similar reasoning shows that these theories are strongly independent. $\quad\square$

A similar analysis applies to the axioms asserting more than one inaccessible cardinal. Let I_n be the assertion that there are n inaccessible cardinals, where n is any natural number, and more generally, let I_α be the assertion that there are at least α many inaccessible cardinals, in their natural order, for any ordinal α. Consider the cautious assertion of infinitely many inaccessible cardinals, namely the axiom $I_{\omega\circ}$ asserting that there are n inaccessible cardinals for every natural number n up to the size of the smallest proof of a contradiction, if any, in the theory $\text{ZFC} + I_\omega$, which asserts that there are infinitely many inaccessible cardinals. Note that the assertion $I_{\omega\circ}$ is a single axiom, in contrast with the cautious enumerations of theories we considered earlier.

Theorem 22. *The cautious assertion of infinitely many inaccessible cardinals is strictly stronger than any particular finite number of inaccessible cardinals and*

strictly weaker than infinitely many.

$$I < I_2 < I_3 < \quad \cdots \quad < I_{\omega^\circ} < I_\omega$$

Proof. We have assumed that it is consistent with ZFC for there to be infinitely many inaccessible cardinals, and so the theory $\mathrm{ZFC} + I_{\omega^\circ}$ will actually include every I_n as an axiom, and every particular instance of this will be provable. So the consistency strength of I_{ω° will strictly exceed that of any particular I_n. On the other hand, there is a model M satisfying $\mathrm{ZFC} + I_\omega + \neg\,\mathrm{Con}(\mathrm{ZFC} + I_\omega)$. This model will think that I_{ω° stops the assertions at some (nonstandard) n, the stage at which ZFC proves a contradiction from the theory $\mathrm{ZFC} + I_\omega$. The model thinks I_{n+1} is true, and hence that $\mathrm{Con}(\mathrm{ZFC} + I_n)$ and consequently that $\mathrm{Con}(\mathrm{ZFC} + I_{\omega^\circ})$ is true. But since it thinks $\neg\,\mathrm{Con}(\mathrm{ZFC} + I_\omega)$, this shows the cautious axiom is strictly weaker in consistency strength.   □

A similar idea will now enable us to exhibit a similar natural family of strongly independent axioms, incomparable in consistency strength. Let f be the universal computable function, defined relative to the theory $\mathrm{ZFC} + I_\omega$. Let $I_{\omega^\circ}^{(k)}$ be the axiom asserting that the number of inaccessible cardinals is at least the running time of $f(k)$. If $f(k)$ halts, after all, this is a reason to distrust the theory $\mathrm{ZFC} + I_\omega$, which asserts infinitely many inaccessible cardinals, since it is an instance where this theory proved something false, and so in this case $I_{\omega^\circ}^{(k)}$ retreats to assert only some finitely many inaccessible cardinals. But just as proved in theorem 10, it will turn out that these particular reasons for limiting the full axiom I_ω are strongly independent over $\mathrm{ZFC} + I_\omega$. Each axiom $I_{\omega^\circ}^{(k)}$ is being cautious about I_ω in a different, independent manner; but every one of them is quite reasonable.

Theorem 23. *The cautious inaccessibility assertions $I_{\omega^\circ}^{(k)}$ are pairwise incomparable in consistency strength over* ZFC *and indeed strongly independent—their consistency statements freely generate the countable free Boolean algebra. Each of these axioms is strictly stronger than every I_n, but strictly weaker than I_ω.*

Proof. If $k \neq r$, then there is a model of $\mathrm{ZFC} + I_\omega$ in which $f(k)$ is defined and $f(r)$ is not, and the model thinks that ZFC can prove this. The model must also think $\neg\,\mathrm{Con}(\mathrm{ZFC} + I_\omega)$, and so this is a model in which $\mathrm{ZFC} + I_{\omega^\circ}^{(k)}$ is thought to be consistent and $\mathrm{ZFC} + I_{\omega^\circ}^{(r)}$ is not. A similar argument works with any finite pattern for the consistency of the theories $I_{\omega^\circ}^{(k)}$, and so these axioms are strongly independent in consistency strength.   □

One might have been tempted to try to formulate a cautious version of I_ω by asserting that there are as many inaccessible cardinals as are consistent, up to ω. That is, the assertion that for every n, if $\mathrm{ZFC} + I_n$ is consistent, then I_n. This statement, however has no consistency strength at all—it is equiconsistent with ZFC—because it is equiconsistent with ZFC that ZFC itself is inconsistent, which trivializes the stated axiom. The cautious formulation I_{ω°, in contrast, strictly exceeds every particular I_n.

Generalizing the previous temptation, let us define the *provisional* assertion of any sentence A to be the assertion $\mathrm{Con}(\mathrm{ZFC} + A) \to A$. This sentence asserts that A is true, provided that it is consistent. But again, it has no consistency strength at all over ZFC, because it is relatively consistent with ZFC that $\neg\,\mathrm{Con}(\mathrm{ZFC})$, which trivializes the provisional assertion of A.

One might also consider a dual to the cautious enumerations. Namely, in the *petulant enumeration* ZFC$^\bullet$, we enumerate the axioms of ZFC, but if we should ever find an explicit proof of contradiction in our favored very strong consistent theory T, then in a petulant rage we immediately add a contradiction also to the current enumeration, thereby spoiling ZFC$^\bullet$. (The general idea is proposed in [Fef60, theorem 7.6], attributed to Steven Orey.) This enumeration is arguably unnatural, but it is not beyond imagination—we might consider a large cardinal set theorist who truly and sincerely believes in the hierarchy of theories from ZFC up through inaccessible cardinals, measurable cardinals, supercompact cardinals and so on, but is so vested in their success that for an unexpected contradiction to arise higher up would undermine the whole picture in a way that was too upsetting to contemplate. Such is the personality behind the petulant enumeration.

In any case, since we have assumed that the strong theory T is actually consistent, the petulant reaction will never come, and so the petulant enumeration results in exactly the usual theory ZFC. The difference between ZFC and ZFC$^\bullet$ is in this sense merely theoretical and can only be realized in nonstandard models and even then only with nonstandard instances of the axioms. Nevertheless, the petulant theory is much stronger than ZFC in consistency strength, for it is equiconsistent with T. The reason is that if Con(ZFC$^\bullet$), then the enumeration must never have reached the petulant stage, and so Con(T) as well. And conversely, from the consistency of T we know both that ZFC is consistent and that ZFC$^\bullet$ agrees with ZFC. So the petulant theory ZFC$^\bullet$ is equiconsistent with the much stronger theory T.

Let me next consider whether there is a double-jumping large cardinal hypothesis. All the usual large cardinal hypotheses are single-jumping, in the sense that they each have a strong consistency strength, but the negations of the hypotheses do not. And yet, we know by theorem 3 that there are set-theoretic axioms with the double-jumping feature, so that both the axiom and its negation have strictly higher consistency strength. Is there a double-jumping large cardinal axiom? Is there a large cardinal axiom A, so that both A and $\neg A$ have a consistency strength that is strictly stronger than ZFC? Consider the following tentative large cardinal assertions:

Theorem 24. *Amongst the assertions of the form:*

> *"If this specific computational process halts, then there is an inaccessible cardinal"*

there are instances with double-jumping consistent strengths—the sentence and its negation have incomparable consistency strength over ZFC.

> *There is computable function f for which the assertions*

> > *"if $f(n)$ halts, then there is an inaccessible cardinal"*

have strongly independent incomparable consistency strengths, and each statement is double jumping in consistency stregnth.

Proof. Let f be the universal computable function defined relative to the theory ZFC $+ I +$ Con(ZFC $+ I$), and let ψ_n be the statement, "if $f(n)$ halts, then there is an inaccessible cardinal." If $n \neq m$, then there is a model M of ZFC $+ I +$ Con(ZFC $+ I$) in which $f(n)$ halts and $f(m)$ does not, provably so by ZFC in the model. Because M thinks ZFC$+I$ is consistent, it also thinks there is a model N of ZFC$+I+\neg$ Con(ZFC$+I$). The computation of f on n and m is the same in M and N, since it had already found the key decision step at a finite stage in M, whose

natural numbers are an initial segment of those in N. Because $f(n)$ halts in N, the statement ψ_n is equivalent to I, but N thinks $\neg\,\mathrm{Con}(\mathrm{ZFC}+I)$, so this is a model of $\neg\,\mathrm{Con}(\mathrm{ZFC}+\psi_n)$. But because $f(m)$ does not halt, the statement ψ_m is vacuously true, and provably so. So the consistency of ψ_m amounts to $\mathrm{Con}(\mathrm{ZFC})$, which holds as a consequence of I in N. So the statements are pairwise incomparable. By controlling any finite pattern of halting in f, we can similarly see that the statements are strongly independent.

To see that the statements ψ_n are all double-jumping, consider the model N as above, which satisfied $\neg\,\mathrm{Con}(\mathrm{ZFC}+\psi_n)$, but in light of I, it satisfies $\mathrm{Con}(\mathrm{ZFC}+\neg I)$ and consequently $\mathrm{Con}(\mathrm{ZFC}+\neg\psi_n)$, since $f(n)$ does halt there. Conversely, if we consider ψ_m in the same model, we've already observed $\mathrm{Con}(\mathrm{ZFC}+\psi_m)$ in N, and $\neg\,\mathrm{Con}(\mathrm{ZFC}+\neg\psi_m)$ holds there, since $f(m)$ provably does not halt. If we make this same analysis with n instead of m, we see that ψ_n and $\neg\psi_n$ are incomparable in consistency strength, and so ψ_n is double-jumping. $\qquad\square$

We could have used essentially any strong statement in the conclusion, instead of asserting the existence of inaccessible cardinals; for example, we could have said that if the computation halted, then there is a strong cardinal or a proper class of Woodin cardinals or what have you. In addition, essentially the same argument works with other forms of existential hypotheses. For example, statements of the form

> "If this specific diophantine equation $p(\vec{x})=0$ has a solution in the integers, then there is a supercompact cardinal"

or the form

> "If this specific set of polygonal tiles admits no tiling of the plane, then there is an almost huge cardinal"

will also respectively exhibit instances of double-jumping and strongly independent incomparability in consistency strength. Are these large cardinal axioms? They do seem to make at least provisional large cardinal assertions, and in this sense they might be regarded as large cardinal axioms. Such a provisional nature about consistency, I claim, is a necessary feature of any double-jumping hypothesis, because if a statement provably implies $\mathrm{Con}(ZFC)$, then it cannot be double-jumping, since the negation of the statement would follow from $\neg\,\mathrm{Con}(\mathrm{ZFC})$, which has no consistency strength at all. If a statement is to be double-jumping, therefore, then it must be provisional in this way about consistency.

5. Nonlinearity in the hierarchy of transitive-model existence

Let us consider an analogue of the hierarchy of consistency strength, but where one requires the existence of transitive models of the theory, rather than mere consistency. In the hierarchy of *transitive-model existence* strength, we say that theories are related $S \leq T$ if we can prove in ZFC that the existence of a transitive model of T implies the existence of a transitive model of S.

It turns out that many of the nonlinearity results carry over almost unchanged to this revised hierarchy.

Theorem 25. *There is a computable function f for which the assertions*

> *"there are $f(n)$ inaccessible cardinals"*

are incomparable and strongly independent in the hierarchy of transitive-model existence.

Proof. We use the same function f as in theorem 8. If $n \neq m$, then as before there is a model M with infinitely many inaccessible cardinals in which $f(n) < f(m)$, and both are defined. Inside this model M, let N be any $\in$-minimal transitive model of ZFC that has a transitive model with $f(n)$ many inaccessible cardinals. It follows by minimality that this model can have no transitive model with $f(n) + 1$ many inaccessible cardinals, and so N is a model that thinks "there are $f(n)$ inaccessible cardinals" has a transitive model, but "there are $f(m)$ inaccessible cardinals" does not. Therefore the transitive model existence statement for n does not imply the statement for m, and so these statements are all pairwise incomparable in that hierarchy. By considering finite patterns with the universal computable function, we similarly achieve that these statements are strongly independent. $\square$

6. NONLINEARITY IN THE LARGEST-NUMBER CONTEST

Allow me briefly to digress with a discussion of how some of these issues play out with the *largest-number* contest, in which contestants compete to describe the largest number subject to certain constraints of space and language; they write their submissions on a standard index card, with rules specifying which characters are allowed and how many. Perhaps one contestant fills their card naively with 9s in decimal notation, while another describes a much larger number with factorials `9!!!!!!!!` or with iterated exponentials `2^2^2^2^2^2`. On even a moderately sized card, we can describe some truly large numbers.

Although I did once supervise and judge an actual instance of the game with competitors from the audience of a large public lecture I gave in Shanghai [Ham13], nevertheless the largest-number contest is more often played in the imagination—it is more a thought experiment than an actual game. The reason is that the game leads quickly to difficult metamathematical matters, which begin to arise when one allows number descriptions going beyond mere primitive recursive terms. For starters, if one allows a free-form descriptive language, then one will immediately engage with Berry's paradox in submissions such as, "the largest number that can be described on 3×5 index card, plus 1," a description that itself fits easily on a 3×5 index card, but by doing so serves up a paradox.

One might hope to avoid paradox by restricting the language, allowing only precise definitions in a formal language, say, such as those of the form "the smallest number n such that $\varphi(n)$," where φ is a formula in the first-order language of arithmetic; or perhaps one allows submissions of the form, "the output of this specific Turing machine computation." These kinds of submission might initially seem tame, but how are we to know whether indeed there is such a number n for which $\varphi(n)$ or whether the submitted computational procedure will ever halt? Perhaps we should insist that submissions be accompanied by a proof that the submission does indeed define a number. For example, such a proof is easy to provide for definitions of the form, "the smallest number n such that $\varphi(n)$, if any, otherwise 17."

And yet, the judge will have difficulty to decide who has won. Even in the case where the language is restricted to primitive recursive terms of the form:

    googol plex bang bang stack
        or

`googol stack bang plex plex,`[1]

and so on, then it is an open question whether in the general case there is any feasible algorithm to determine the larger number; see [Ham17b].

There is no computable procedure at all to determine whether submissions of the form "the running time of this computation" are legitimate submissions, since the judge would have to solve these instances of the halting problem. The same holds for determining the winner for submissions of the form "the running time of this computation, if it halts, otherwise 17," for if another player submits 18, then the judge will have to know whether that computation halts to adjudicate the winner.

Apart from computable undecidability, however, there is the more subtle and profound issue of logical undecidability. We might suppose that the judge is presiding over the contest in the context of a fixed official background theory T, a consistent computably axiomatizable theory that she relies upon when adjudicating the comparative sizes of the number descriptions. Perhaps this theory is very strong—it might be ZFC plus an aspirational large cardinal hypothesis, amongst the strongest theories thought to be consistent. The problem, of course, is that no consistent computably axiomatizable theory will settle all the comparisons that might be put to it; there will be concrete number descriptions for which the theory does not provably settle their comparative size, and so the judge will not be able to declare the winner. One player might submit, "the output of this computation" and another "the output of that computation," but just as the theory does not settle the values of the universal computation, it will not be able to settle who has won.

A determined game-player might argue that we want the judge not to use a theory at all, but rather to use *actual arithmetic truth*, truth in the standard model $\mathbb{N}$. That is, we want to declare as the winner the number description that actually gives rise to the largest number, when this description is interpreted in the standard model $\mathbb{N}$, whether or not we are able to prove this. In ZFC we can prove after all that there is a definite truth predicate on the standard model $\mathbb{N}$, which can be used to adjudicate these comparisons.

But this proposal is much less helpful than one might expect. What good after all is using a complete truth predicate, if we don't know what the predicate's truth judgements are? Although ZFC and our other standard foundational theories prove that there is a definite arithmetic truth predicate, these theories do not tell us fully which arithmetic statements are true. So this proposal does not seem to enable the judge to determine the winner. Faced with two number descriptions, even if the judge might know that one of them describes the actually larger number, she may have no way to find out which one.

We can summarize the basic facts in the following theorem.

Theorem 26. *For any given consistent theory T, there are entries for the largest-number contest of the form, "the first n with this concrete property $\varphi(n)$," for which it is undecidable in T which player has won. Furthermore, the respective assertions that a given player wins are statements with incomparable consistency strengths.*

---

[1]Following pop-math practice, a *googol* is 10^{100}; the expression x `bang` refers to the factorial $x!$; the expression x `plex` means 10^x; and x `stack` means the x-iterated exponential $10^{10^{\cdot^{\cdot^{10}}}}$.

7. Illusory linearity and the confirmation bias argument

Let me discuss a few reasons for thinking that we might simply be mistaken about the linearity phenomenon. Perhaps it is illusory?

One observation tending to undermine the linearity evidence is simply the fact that many common large cardinal notions were constructed specifically to strengthen previous notions, often in well-ordered hierarchies. The progressions from inaccessible cardinals to hyperinaccessible and the hyperinaccessibility hierarchy and from Mahlo through the hyper-Mahloness hierarchy and ultimately to greatly Mahlo occurred exactly like this in the early days of set theory. Similarly with the progression from measurable cardinals to hypermeasurability and through the strongness hierarchy to strong cardinals. Solovay generalized the strongly compact to the supercompact cardinals with an embedding characterization that led naturally to many further strengthenings, among them the almost huge, huge, and superhuge cardinals. These conceptions altogether firmly established the paradigm of large cardinals as critical points of embeddings $j : V \to M$, which constitute the large-scale bones of the large cardinal hierarchy. One naturally strengthens these axioms simply by insisting upon progressively stronger closure properties on M. These strengthened notions are often therefore linearly related in a way that is completely unsurprising—they form a linear hierarchy precisely because that is how we created them.

Mirna Džamonja describes the situation like this:

> The linearity pretty obviously seems to be just a consequence of definitions mostly being variants of each other. Increase the target, increase the closure. . . clearly once we get to be more inventive we shall have no linearity. [Dža21]

This linear creation process occurs even at the very bottom of the hierarchy of consistency strength, with the tower of iterated consistency assertions, at each transfinite stage adding the consistency statement of the theory that has come before. Of course this creates a well-founded linear tower of consistency strength.

These observations show that huge parts of the consistency strength hierarchy exhibit linearity only for superficial, unsurprising reasons, which therefore cannot count as evidence of a broader or more fundamental linearity phenomenon. The unexpected instances of linearity are simply many fewer than one might have expected. Do those instances suffice to establish linearity as a genuine phenomenon?

Meanwhile, in a number of irritating instances, we don't yet actually know the natural hierarchy to be linear. For example, we don't yet know well exactly how the strongly compact cardinals fit into the hierarchy, even though this was one of the earliest large cardinal notions, arising from compactness properties of infinitary logic. This case is often strangely brushed off by the defenders of linearity, but I find it directly to undermine the claim of a sweeping linearity phenomenon for large cardinal notions originating in diverse parts of the subject. Similarly, we don't know how to compare one supercompact cardinal with two strongly compact cardinals; we don't know the strength of an indestructible weakly compact cardinal; we don't know how the proper forcing axiom fits into the hierarchy; and similarly in many other cases. As far as we know, there could be abundant nonlinearity in consistency strengths surrounding these notions.

One might reply to this criticism by pointing out that it is not only the large cardinal notions themselves that are linearly ordered by consistency strength, but all the other statements in mathematics that have been proved equiconsistent with various large cardinal notions. Work of Solovay and Shelah shows that the impossibility of removing the axiom of choice from Vitali's construction of a nonmeasurable set is equiconsistent with the existence of an inaccessible cardinal; dramatic work on determinacy establishes equiconsistency connections with Woodin cardinals; and so on in many other cases. From this point of view, it doesn't much matter that the large cardinal notions themselves are often linearly ordered in a trivial manner, if all these other statements from mathematics are proved to line up with that hierarchy—it would still establish a genuine linearity phenomenon.

But does this counterargument overplay its hand? We don't actually have *so* many instances of equiconsistency between large cardinals and mathematical principles arising outside logic and set theory. There are some very prominent cases, but are these truly sufficient to make the case for a genuine widespread linearity phenomenon? I'm not sure.

Let me next explain a subtler argument suggesting that we may be mistaken about linearity, namely, what I call the confirmation-bias argument. When engaging with the independence phenomenon in set theory and establishing relative consistency strengths, we generally begin with a model of one theory T and construct from it a model of another theory S. From any model of ZF, for example, we construct models of ZFC + CH or ZFC + ¬CH or Martin's axiom or what have you; from a model of ZFC with sufficient large cardinals, we construct models of ZF + DC with various determinacy axioms. When we begin with a model of T and construct a model of S, then we will have established $S \leq T$ in consistency strength, and when there is also a converse construction, then we will also know $T \leq S$ and consequently that the theories are equiconsistent, $S \equiv T$. When the model of S that we construct is a set model inside the model of T, then by the sufficiency condition I mentioned in section 1, we deduce the strict relation $S < T$.

In nearly all these arguments, the tools we use are forcing and transitive inner models (including set-sized models), often in sophisticated elaborate combination. Set theorists have become expert in using these tools to explore the vast range of set-theoretic possibility.

The confirmation-bias observation is that with this process and these tools by themselves, we shall never establish an instance of incomparability in consistency strength. The reason is that both methods preserve arithmetic truth—we cannot change the arithmetic of a model by forcing or by going to a transitive set or transitive class inner model. But in order to establish nonlinearity in consistency strength, it is necessary that we change the arithmetic truths of the models. To show that theories S and T are incomparable, after all, we need to provide a model of ZFC with Con(S) but not Con(T) and another with Con(T) but not Con(S), and this will require that these models have not only different natural number structures $\mathbb{N}$, but end-extension incomparable such natural number structures— and so they will also be interpretatively incomparable, for in any interpretation the inconsistency of the other model would still be present and we wouldn't achieve the incomparability situation.

The final conclusion of the confirmation-bias argument is that we shouldn't be surprised to observe only linearity, if our tools are incapable of observing nonlinearity. This is precisely what it means to suffer a confirmation-bias error.

8. Is self-reference disqualifying for naturality?

Many people object to the naturality of the arithmetic sentences mentioned in section 1 on the grounds that these sentences are self-referential. Perhaps it is thought that self-reference is a strange and unexpected feature in mathematics, and therefore it may perhaps be disqualifying for a self-referential assertion to be considered "natural." I should like to push back against this view.

I find diagonalization and self-reference to be at the very core of mathematical logic and set theory. Set theory as a subject essentially begins with Cantor's diagonal argument, establishing uncountability in an instance of diagonal self-application. Russell's refutation of the general comprehension axiom is explicitly self-referential, building the class of all non-self-membered sets. (And we don't take general comprehension to be fine otherwise, for "natural" assertions.) These arguments are surely amongst the founding central ideas of the subject, and the diagonalization idea is woven deeply throughout it. Furthermore, these diagonalizations are fundamentally the same as used to prove the fixed-point lemmas that lead to the Gödel and Rosser sentences. What can be the coherent philosophy of "natural" that counts the constructions of Cantor and Russell as natural, but not the fundamentally similar construction of the Gödel and Rosser sentences?

To my way thinking, the large cardinal axioms themselves engage in a kind of self-reference. To assert that there is an elementary embedding $j : V \to M$ of a certain kind is to assert that there is transformation of objects from our current world V to a new world M associating to every object x in our world a duplicate $j(x)$ in the new world, with all the same properties in that world as x had in the old world. So the axiom at bottom posits a system of duplicates $j(x)$, whose properties are stated by (self-)reference back to x. In this light, nearly every large cardinal axiom partakes of self-reference.

Or consider the proper forcing axiom PFA, a strong generalization of Martin's axiom MA from the case of c.c.c. forcing to proper forcing. This axiom is commonly included amongst the "natural" set-theoretic hypotheses, with most bets placing it as equiconsistent with (or very near to) the existence of a supercompact cardinal in consistency strength. The axiom makes the Martin's axiom claim about the existence of a filter meeting any given $\aleph_1$ many dense sets for a forcing notion $\mathbb{P}$, provided that it is proper. But to my way of thinking, the property of a forcing notion $\mathbb{P}$ being proper is at least as self-referential as the Rosser sentence or the other simple arithmetic sentences discussed earlier in this article (and an order of magnitude more difficult). Namely, $\mathbb{P}$ is proper essentially when $\mathbb{P}$ densely often contains conditions that are themselves generic or ensure genericity for the forcing with the version of $\mathbb{P}$ itself over suitable countable elementary substructures of sufficient fragments of the universe; put simply, $\mathbb{P}$ is proper when it consists of conditions that are generic for the forcing $\mathbb{P}$ itself, in countable simulacra over various countable universes. Is this any less self-referential than a sentence making an assertion about its own Gödel code? Why should the comparatively simple Rosser sentence be somehow beyond the pale, while properness counts as natural?

Diagonalization and self-reference are pervasive in logic and set theory, and to regard those features as automatically unnatural would have us declaring the entire subject unnatural.

9. Naturality

What does it mean to have a "natural" example in mathematics? Many mathematicians seem to adopt a know-it-when-you-see-it attitude to naturality, without giving a formal account. Does it matter that what counts as natural sometimes changes over time?

There is the connotation that natural examples are those occurring in practice—Koellner [Koe11] refers to examples that "arise in nature" and Steel [Ste14] says that natural set-theoretic hypotheses are those "considered by set theorists, because they had some set-theoretic idea behind them."

It would be the naturalist fallacy, of course, to imbue that conception of naturality with all the positive connotations one usually finds for this word in mathematics. Is mathematics rife with the naturalist fallacy? I don't believe so, because naturality as it is used in mathematics is usually not simply a reduction to actual practice. After all, one commonly hears naturality-based objections to examples offered in actual practice; mathematicians often criticize even a well-established argument or example, requesting a more natural one. Consider the subject of computability theory, for example, which has thousands of constructions using complex priority arguments to establish certain fundamental features in the hierarchy of Turing degrees. These are computability-theoretic constructions, introduced by computability theorists for computability-theoretic purposes, but the resulting degrees are nevertheless sometimes criticized as unnatural. One hears requests for "natural" Turing degrees with certain features—for example, is there a natural solution of Post's problem?

(Another instance, which I hesitate to mention: I am a set theorist who has introduced the theories and axioms of this paper for a set-theoretic purpose, defending them as natural and set-theoretic, while proving that they constitute instances of nonlinearity and illfoundedness. But are we to take this by itself as refuting Steel's conjecture, mentioned on page 2? I doubt that I shall succeed so easily.)

Let us also consider that, regrettably, the actual-practice conception of naturality can too easily be construed narrowly in a way that harms mathematical advance. Namely, in my experience concerns about naturality are sometimes raised in effect simply to reject unfamiliar ideas or constructions. In a few instances on MathOverflow, for example, a mathematician not from logic asked a mathematical question that I was able to answer using ideas from set theory and logic; perhaps my argument used transfinite recursion or some other standard logic method, completely natural to my way of thinking. But the solution was rejected as unnatural—the person was just not comfortable with those methods. Apparently what would have counted as natural in these cases was a solution using only ideas common in that subject's actual practice. These experiences have frankly predisposed me strongly against the free-and-easy but ultimately empty use of "natural" in mathematics. What a pity it would be if our conception of the natural simply leads us all into scattered knowledge silos.

A different conception of the arising-in-practice connotation would be that natural notions are those that might easily arise in an unrelated subject or practice.

On this view, examples in logic would count as natural if they might arise in graph theory, algebra or topology. An actual graph-theoretic decision problem would be seen as natural in computability theory. But this view, taken strictly, would seem to rule out many of the large cardinal hypotheses that Steel and Koellner want to see as natural, since other subjects generally give very little consideration to large cardinals at all. We need a different notion of the "natural."

There is a connotation that natural examples must be examples that have or at least could arise independently of whatever immediate application is currently being made of them. This would explain why many computability theoretic constructions are often seen as unnatural. Furthermore, this property would also be true whenever an example was not overly detailed or technical, since one can easily imagine that simple examples could arise independently of any particular motivation. Certainly in many instances, the natural examples seem to be those that are easily described or presented.

For such a view, however, one might recall Shelah's response to criticisms of his arguments as being overly "technical":

> The term technical is a red flag for me, as it is many times used not for the routine business of implementing ideas but for the parts, ideas and all, which are just hard to understand and many times contain the main novelties. [She93, Axis A]

One might respond similarly to claims of unnaturality. Would we want to hold that all difficult examples are unnatural? Earlier I had mentioned the proper forcing axiom PFA, which is surely difficult yet commonly seen as natural. And many large cardinal notions, including Woodin cardinals, remarkable cardinals, and many others, are commonly regarded as difficult.

Consider another naturality objection that one might imagine against the incomparability sentences of sections 2, 3, and 4. Namely, the differences between the sentences on offer are revealed only in ω-nonstandard models, which is an unnatural case. The differing natures of the computable function f used in theorem 8, for example, can be exhibited only in ω-nonstandard models; in standard models, the sentences will always agree.

The rebuttal to this objection is that consistency strength is *inherently* about ω-nonstandard models. If we only ever considered ω-standard models, they would all always agree about the consistency of any given theory whatsoever, and there would be no hierarchy to analyze. For one theory to be weaker than another in consistency strength, $S < T$, means exactly that there is a model of $\mathrm{Con}(S) + \neg\,\mathrm{Con}(T)$, and this model must be ω-nonstandard if T is consistent. Therefore any substantive treatment of the consistency strength hierarchy must ultimately be concerned with nonstandard models.

Let me next raise the distinction between a natural *kind* of problem and a natural problem instance. The polygonal tiling problem, for example, is surely a natural kind of problem, but does this mean that every individual tiling problem is natural? If not, this might speak against the success of my examples in corollary 14 as natural instances of nonlinearity. Similarly, perhaps the incomparability and ill-foundedness examples provided by theorems 8, 10, 11 and so on are merely instances of natural kinds of problems, rather than actually natural instances themselves.

The respective undecidability of the tiling problem, the diophantine problem, the halting problem, the Entsheidungsproblem, and many others holds profound

philosophical significance in identifying fundamental limitations on our ability to achieve mathematical knowledge mechanistically. The undecidability results show that we can have no uniform computable procedure to solve instances of these extremely natural kinds of problems. Yet, the undecidability results themselves are all proved ultimately by means of diagonal arguments that make use in each case of weird self-referential instances of those problems. Turing's argument for the undecidability of the halting problem, for example, invokes a program that asks about its own behavior when applied to itself as input (which is something we basically never do in practice). And this weird example is in effect copied into all the others, when one proves undecidability through a reduction of the halting problem.

But do we say that the halting problem is decidable for "natural" instances? No, even though for the programs arising in practice, those written for a clear computational purpose, we can indeed generally determine whether it will halt or not. (Indeed, Miasnikov and I proved [HM06] that there is a linear-time algorithm that correctly decides almost every instance of the halting problem, with respect to the asymptotic density measure—as the number of states n increases, the proportion of all n-state programs handled by our algorithm approaches 100%.) But in computability theory we do not generally say that the halting problem is essentially decidable; we do not highlight a "natural" halting-problem-decidability phenomenon, according to which we can decide halting in the natural cases that matter. Why do the set theorists claim that consistency strength is linear in the natural cases that matter?

Steel [Ste13, slide 26] advances the idea that the consistency strength order of natural theories aligns with containment of their arithmetic consequences. Specifically, the claim is that for natural theories for which $S < T$ in consistency strength, then any arithmetic consequence of S will also be provable in T. I pointed out in [Ham16a], however, that Steel's principle will require us to rule out the naturality of many theories that we might otherwise have been inclined to accept as natural. Specifically, if T is a natural large cardinal theory, then it seems to me that many set theorists would readily accept $\mathrm{Con}(T)$ as a natural assertion. Indeed, for many natural theories this is an assertion that has appeared as an explicit hypothesis in perhaps hundreds of published theorems in the set-theoretic research literature, and in this sense fulfills the actual-practice conception of naturality.

But to regard $\mathrm{ZFC} + \mathrm{Con}(T)$ as natural would in most cases contradict Steel's arithmetic containment principle. To see this, assume that T proves the existence of an inaccessible cardinal κ. It follows that T will prove all instances of the reflection scheme asserting, "if arithmetic statement ϕ is provable in ZFC, then it is true." The reason is that T will prove that ϕ holds V_κ, and since ϕ is arithmetic, it will be absolute to the full set-theoretic universe V. Meanwhile, the consistency-wise stronger theory $\mathrm{ZFC} + \mathrm{Con}(T)$ cannot prove all instances of this scheme, because this theory is consistent with the assertion $\neg\,\mathrm{Con}(\mathrm{ZFC} + \mathrm{Con}(T))$, and any model of this combined theory will think that $\neg\,\mathrm{Con}(T)$ is provable in ZFC but not true. Therefore, we have a natural theory T that is strictly weaker in consistency strength than another theory $\mathrm{ZFC} + \mathrm{Con}(T)$, yet the weaker theory proves some arithmetic statements that are not provable in the stronger theory. This would contradict the arithmetic containment principle, unless we regard $\mathrm{ZFC} + \mathrm{Con}(T)$ as unnatural. Therefore, if one is committed to the idea that consistency strength increases align

with containment of arithmetic consequences for natural theories, then we cannot allow $ZFC + Con(T)$ as a natural theory even when T is.

Steel will reply, of course, that indeed $ZFC + Con(T)$ is not a natural theory—this theory instantiates exactly what he describes as the *instrumentalist dodge* in set theory, described in [Fef+00, p. 423]. Namely, we don't want to assume merely that the large cardinals are consistent, but rather that they are actually true. What I have just argued here is that the arithmetic-containment principle for consistency strength requires this stance. In particular, according to this view the theory $ZFC + Con(ZFC + \exists \text{ inaccessible})$ is not natural.

10. ANALOGY BETWEEN SET THEORY AND COMPUTABILITY THEORY

Let me draw an analogy between set theory with its study of the hierarchy of consistency strength and computability theory with its study of the Turing degrees, a rich hierarchy of complexity that is surely as deep and complicated as the hierarchy of consistency strength, and also as philosophically significant—one can view the Turing degrees as the possible countable amounts of information. Russell Miller has described a "build it and they will come" philosophy in computability theory, according to which if one wants to exhibit a certain feature in the hierarchy of Turing degrees, then you simply have to get down to business and make it happen with a particular construction built for the purpose. Computability theorists seem quite commonly to embrace the chaotic scrappiness of the hierarchy of Turing degrees. I wonder whether such an attitude towards the hierarchy of consistency strength in set theory would lead us to discover fascinating new phenomenon in the degrees of consistency strength.

Meanwhile, computability theorists point to their own natural linearity phenomenon, namely, the "naturally arising" Turing degrees invariably arise in a linear, well-ordered part of the hierarchy of Turing degrees. The commonly arising definable sets of natural numbers, the ones we might be independently interested in, tend to have their Turing degrees landing precisely on one of the low-level iterated jumps:

$$0 < 0' < 0'' < 0''' < \cdots < 0^{(\omega)} < \cdots$$

In this sense the "natural" Turing degrees are well-ordered, and researchers seek a deeper explanation. Joseph Miller points to the research efforts around the Sacks question, asking for a degree-invariant solution of the generalized Post's problem, and the Martin conjecture, seeking to establish the Turing jump and its iterates as canonical for definable degree-invariant Borel actions, as part of the program to provide a deeper explanation of the linearity phenomenon in the Turing degrees. Antonio Montalbán [Mon19] explains the importance of Martin's conjecture like this:

> The [linear] hierarchy we were looking for seems to exist, but $\mathcal{D}$ [the Turing degrees] seems too chaotic to help us find it. The contrast between the general behavior in $\mathcal{D}$ and the behavior of the naturally occurring objects is so stark that there must be a deep reason behind it. We need to dig deeper.

Those conjectures in effect seek to replace naturality talk by identifying exactly the properties that are sought: degree-invariance and restrictions to Borel actions in place of arbitrary actions. One needn't refer any longer to the "natural" Turing degrees to engage with them.

11. A challenge for defenders of natural linearity

At the end of the day, my view is that almost all our talk of naturality in mathematics is thoroughly unsatisfactory. We have no coherent robust concept of what counts as natural, and empty naturality talk is too often used merely to reject the unfamiliar. For someone to declare a construction or idea "unnatural" is often little different from them saying, "I don't like it" or "it uses unexpected ideas." To my way of thinking, we may legitimately dismiss concerns about naturality as easily as they are raised, which is to say, effortlessly.

I should like to emphasize, however, that the rejection of examples or solutions as "unnatural" often indicates merely that a question was not well formulated, for the actual objections in play are ultimately something else—the solution lacked a certain unstated but presumed feature or exhibited a certain unanticipated feature. A more informed followup formulation of the question would clarify what was actually desired, and might lead to mathematical advance.

So let me close this article on a positive note with a challenge to the defenders of the natural linearity phenomenon. In light of the abundant counterexamples establishing pervasive nonlinearity and ill-foundedness in the hierarchy, I propose that we should abandon the empty naturality talk and instead get down to the work of identifying the attractive features we had sought in our notion of the natural. Can we give legs to a reified naturality notion that is sufficient to establish linearity in the consistency-strength hierarchy? For example, what is the set-theoretic analogue of the Martin conjecture for consistency strength? Might we ultimately hope to identify a broad class of assertions with welcome, attractive features—standing in for the so-called "natural" assertions—which provably align into a well ordered hierarchy of consistency strength? That would be how to do it.

Here is one small step in that direction, which I provided at [Ham21] in response to a question of Dmytro Taranovsky.

Theorem 27. *There is no monotone analogue of the independent Rosser sentence construction. That is, there is no assignment $\tau \mapsto \rho_\tau$ of sentences τ in the language of arithmetic to sentences ρ_τ with the properties:*

(1) *(Independence) If* $\mathrm{PA} + \tau$ *is consistent, then so are* $\mathrm{PA} + \tau + \rho_\tau$ *and* $\mathrm{PA} + \tau + \neg\rho_\tau$.
(2) *(Logical extensionality) If* $\mathrm{PA} \vdash \tau \leftrightarrow \sigma$, *then* $\mathrm{PA} \vdash \rho_\tau \leftrightarrow \rho_\sigma$.
(3) *(Monotonicity) If* $\mathrm{PA} \vdash \tau \rightarrow \sigma$, *then* $\mathrm{PA} \vdash \rho_\tau \rightarrow \rho_\sigma$.

The three properties are redundant, of course, since monotonicity is a strengthening of extensionality.

Proof. Since PA is consistent, it follows that $1 = 1$ is consistent with PA, and so $\rho_{1=1}$ is independent of PA. In particular, $\neg\rho_{1=1}$ is consistent with PA, and so $\mathrm{PA} + \neg\rho_{1=1} + \rho_{\neg\rho_{1=1}}$ is consistent. But PA trivially proves that $\neg\rho_{1=1} \rightarrow 1 = 1$, and so by monotonicity it follows that PA proves $\rho_{\neg\rho_{1=1}} \rightarrow \rho_{1=1}$, which contradicts the earlier stated consistency. $\square$

It follows that the Rosser sentence itself does not obey monotonicity. Can one weaken monotonicity to mere uniformity? Shavrukov and Visser [SV14] provide a uniform computable construction producing a sentence $F(A, B)$ that is strictly between A and B in the Lindenbaum algebra over PA, whenever A is strictly below B in that algebra, and which furthermore is extensional, in the sense that

if A and B are replaced with equivalent A' and B', then $F(A, B)$ is equivalent to $F(A', B')$. They use this construction to provide extensional Rosser constructions. The resulting sentences, however, are not Π_1^0. This seems to leave the following question open:

Question 28. *Is there a Π_1^0 formula $\rho(x)$ with the following properties?*

(1) *(Independence) If* $\mathrm{PA} + \tau$ *is consistent, then so are* $\mathrm{PA} + \tau + \rho(\ulcorner\tau\urcorner)$ *and* $\mathrm{PA} + \tau + \neg\rho(\ulcorner\tau\urcorner)$.

(2) *(Logical extensionality) If* $\mathrm{PA} \vdash \tau \leftrightarrow \sigma$, *then* $\mathrm{PA} \vdash \rho(\ulcorner\tau\urcorner) \leftrightarrow \rho(\ulcorner\sigma\urcorner)$.

A positive answer would provide a degree of nonlinear naturality, thereby undermining the linear naturality hypothesis, whereas a negative answer would support the hypothesis, in showing that independence must be sensitive to intensionality.

References

[AB04]   S.N. Artemov and L.D. Beklemishev. "Provability Logic". In: *Handbook of Philosophical Logic, 2nd ed.* Ed. by D. Gabbay and F. Guenthner. Vol. 13. Dordrecht: Springer, 2004, pp. 229–403.

[BE17]   Rasmus Blanck and Ali Enayat. "Marginalia on a theorem of Woodin". *J. Symb. Log.* 82.1 (2017), pp. 359–374. ISSN: 0022-4812. DOI: 10.1017/jsl.2016.8.

[Ber90]   Alessandro Berarducci. "The interpretability logic of Peano arithmetic". *J. Symbolic Logic* 55.3 (1990), pp. 1059–1089. ISSN: 0022-4812. DOI: 10.2307/2274474.

[Bla17]   Rasmus Blanck. "Contributions to the Metamathematics of Arithmetic. Fixed points, Independence, and Flexibility". PhD thesis. University of Gothenburg, 2017. ISBN: 978-91-7346-917-3. http://hdl.handle.net/2077/52271.

[Cai11]   Andrés E. Caicedo. *(Non?)-linearity of the consistency strength ordering in ZF.* MathOverflow answer. 2011. https://mathoverflow.net/q/59800.

[Dža21]   Mirna Džamonja. *On large cardinal linearity.* Twitter post. 2021. https://twitter.com/MirnaDzamonja/status/1356338350622388226 (version 01 February 2021).

[Fef+00]   Solomon Feferman, Harvey M. Friedman, Penelope Maddy, and John R. Steel. "Does Mathematics Need New Axioms?" *Bulletin of Symbolic Logic* 6.4 (2000), pp. 401–446. ISSN: 10798986. http://www.jstor.org/stable/420965.

[Fef60]   S. Feferman. "Arithmetization of metamathematics in a general setting". *Fundamenta Mathematicae* 49 (1960), pp. 35–92. ISSN: 0016-2736. DOI: 10.4064/fm-49-1-35-92.

[Fri98]   Harvey Friedman. *Proof-theoretic Degrees.* Foundations of Mathematics (FOM) mailing list. Post 20 in Friedman's FOM series on topics in the foundations of mathematics. 1998. https://cs.nyu.edu/pipermail/fom/1998-August/001929.html.

[Ham13]   Joel David Hamkins. *Largest-number contest: what is the largest number that you can describe on an index card?* 2013. http://jdh.hamkins.org/largest-number-contest/.

[Ham16a]   Joel David Hamkins. *Does the consistency strength hierarchy coincide with the "arithmetic consequence" hierarchy at ZF + Reinhardt?* MathOverflow answer. 2016. https://mathoverflow.net/q/236436.

[Ham16b]   Joel David Hamkins. *Every function can be computable!* Mathematics and Philosophy of the Infinite. 2016. http://jdh.hamkins.org/every-function-can-be-computable/ (version 25 March 2016).

[Ham17a]   Joel David Hamkins. *A program that accepts exactly any desired set, in the right universe.* Mathematics and Philosophy of the Infinite. 2017. http://jdh.hamkins.org/a-program-that-accepts-any-desired-finite-set-in-the-right-universe/ (version 06 May 2017).

[Ham17b]   Joel David Hamkins. *Help me put these enormous numbers in order: googol, googol-plex-bang, googol-stack and so on.* Mathematics Stack Exchange question. 2017. https://math.stackexchange.com/q/72646.

[Ham17c]   Joel David Hamkins. *The universal algorithm: a new simple proof of Woodin's theorem.* Mathematics and Philosophy of the Infinite. 2017. http://jdh.hamkins.org/the-universal-algorithm-a-new-simple-proof-of-woodins-theorem/ (version 28 May 2017).

[Ham18]   Joel David Hamkins. "The modal logic of arithmetic potentialism and the universal algorithm". *Mathematics ArXiv* (2018). Under review, pp. 1–35. arXiv:1801.04599[math.LO]. http://wp.me/p5M0LV-1Dh.

[Ham21]   Joel David Hamkins. *Uniform incomparable consistency strengths.* MathOverflow answer. 2021. https://mathoverflow.net/q/384050 (version 15 February 2021).

[HM06]   Joel David Hamkins and Alexei Miasnikov. "The halting problem is decidable on a set of asymptotic probability one". *Notre Dame Journal of Formal Logic* 47.4 (2006), pp. 515–524. ISSN: 0029-4527. DOI: 10.1305/ndjfl/1168352664. arXiv:math/0504351[math.LO]. http://jdh.hamkins.org/haltingproblemdecidable/.

[HW17]   Joel David Hamkins and W. Hugh Woodin. "The universal finite set". *Mathematics ArXiv* (2017). Manuscript under review, pp. 1–16. arXiv:1711.07952[math.LO]. http://jdh.hamkins.org/the-universal-finite-set.

[HW21]   Joel David Hamkins and Kameryn J. Williams. "The Σ_1-definable universal finite sequence". *Journal of Symbolic Logic* (2021). to appear. arXiv:1909.09100[math.LO].

[Jap94]   Giorgi Japaridze. "A simple proof of arithmetical completeness for Π_1-conservativity logic". *Notre Dame J. Formal Logic* 35.3 (1994), pp. 346–354. ISSN: 0029-4527. DOI: 10.1305/ndjfl/1040511342.

[Koe11]   Peter Koellner. "Independence and Large Cardinals". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2011. Metaphysics Research Lab, Stanford University, 2011. https://plato.stanford.edu/archives/sum2011/entries/independence-large-cardinals.

[Kri62]   Saul A. Kripke. "'Flexible' predicates of formal number theory". *Proc. Amer. Math. Soc.* 13 (1962), pp. 647–650. ISSN: 0002-9939. DOI: 10.2307/2034844.

[Lin03]   Per Lindström. *Aspects of incompleteness.* Second. Vol. 10. Lecture Notes in Logic. Association for Symbolic Logic, Urbana, IL; A K Peters, Ltd., Natick, MA, 2003, pp. x+163. ISBN: 1-56881-173-X.

[Mon19]   Antonio Montalbán. "Martin's conjecture: a classification of the naturally occurring Turing degrees". *Notices Amer. Math. Soc.* 66.8 (2019), pp. 1209–1215. ISSN: 0002-9920.

[Mos60]   A. Mostowski. "A generalization of the incompleteness theorem". *Fund. Math.* 49 (1960/1961), pp. 205–232. ISSN: 0016-2736. DOI: 10.4064/fm-49-2-205-232.

[She93]   Saharon Shelah. "The future of set theory". In: *Set Theory of the Reals (Ramat Gan, 1991)*. Vol. 6. Bar-Ilan University, 1993, pp. 1–12. arXiv:math/0211397[math.LO]. https://shelah.logic.at/v1/E16/E16.html.

[Sim09]   Stephen G. Simpson. "The Gödel Hierarchy and Reverse Mathematics" (2009). Manuscript published on the author's web page. http://www.personal.psu.edu/t20/papers/gh/.

[Ste13]   John R. Steel. *Gödel's program.* Slides for a talk given at the CSLI meeting, Stanford University. 2013. https://math.berkeley.edu/~steel/talks/stanford2013copy.pdf.

[Ste14]   John R. Steel. "Gödel's program". In: *Interpreting Gödel.* Cambridge Univ. Press, Cambridge, 2014, pp. 153–179.

[SV14]   V. Yu. Shavrukov and Albert Visser. "Uniform density in Lindenbaum algebras". *Notre Dame J. Form. Log.* 55.4 (2014), pp. 569–582. ISSN: 0029-4527. DOI: 10.1215/00294527-2798754. https://ezproxy-prd.bodleian.ox.ac.uk:2102/10.1215/00294527-2798754.

[Woo11]   W. Hugh Woodin. "A potential subtlety concerning the distinction between determinism and nondeterminism". In: *Infinity.* Cambridge University Press, 2011, pp. 119–129.

(Joel David Hamkins) PROFESSOR OF LOGIC, OXFORD UNIVERSITY & SIR PETER STRAWSON FELLOW, UNIVERSITY COLLEGE, OXFORD

Email address: joeldavid.hamkins@philosophy.ox.ac.uk

URL: http://jdh.hamkins.org