Journal for the Philosophy of Mathematics

OPEN ACCESS

Citation: MEADOWS, Toby. (2025). The Consistency Hierarchy Thesis. *Journal for the Philosophy of Mathematics*. 2: 107-142. doi: 10.36253/jpm-2971

Received: September 18, 2024

Accepted: January 13, 2025

Published: December 30, 2025

ORCiD
TM: 0000-0003-2741-7685

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Firenze University Press
www.fupress.com

The Consistency Hierarchy Thesis

TOBY MEADOWS

Department of Logic and Philosophy of Science, University of California-Irvine, US.
Email: meadowst@uci.edu

Abstract: Set theorists often claim that natural theories are well-ordered by their consistency strength. We call this claim the *Consistency Hierarchy Thesis.* The goal of this paper is to unpack the philosophical and mathematical significance of this thesis; and to develop an understanding of how it is defended and, more particularly, how one might refute it. We shall see that the thesis involves a curious admixture of mathematics and philosophy that makes it difficult to pin down. We investigate some intriguing attempts to refute the thesis that are hampered by the problem of understanding what makes a theory natural. We then develop a thought experiment exploring the idea of what the ideal scenario for refutation would look like. And we show that a counterexample is impossible if we insist that the counterexample uses respectable (i.e., transitive) models. Finally, we reflect on how these hurdles affect our understanding of the significance of the thesis by drawing a parallel with a more famous claim: the Church-Turing thesis.

Keywords: Set theory, Forcing, Inner Model Theory, Consistency, Incompleteness.

He must,
so to speak,
throw away the ladder
after he has climbed up on it.

Wittgenstein

The following pair of facts are well-known. ZFC provides a practically adequate foundation for mathematics as we know it today; and yet, if ZFC is consistent, then it cannot be complete. As such, there are a dizzying variety of extensions of ZFC many of which are incompatible with each other. In the face of such chaos, one might be tempted to take up a conservative attitude and thus, prefer to remain within the comforting confines of ZFC. At present, such a move makes little difference to one's ability to found ordinary mathematics, but the curious mind will still wonder what is out there in the big beyond and just how wild the jungle is. Many set theorists have offered a tantalizing answer to the latter question: natural theories extending ZFC are well-ordered by their consistency strength. We call this the *Consistency Hierarchy Thesis*. The purpose of this paper is to explain what this thesis really means and to argue that it involves such a strange mix of mathematics and philosophy that it is difficult to know how one could successfully defend or rebut it.

We shall start in Section 1 by providing a gentle introduction to the consistency strength relation and the claim that it forms a hierarchy. While this material is very elementary, our patient discussion is intended to draw out the mathematical and philosophical agendas that drive our interest in the problem. In Section 2, we consider how one might attempt to refute the thesis and argue that the prospects for a successful refutation seem bleak. In particular, we shall focus on some proposed counterexamples from Joel David Hamkins' recent paper on the topic (Hamkins, 2025). While we shall push back on these proposals, the emerging theme will not be so much that Hamkins is wrong, so much as that there is something odd about the question itself. Finally, in Section 3, we'll offer some explanation as to why the thesis is so difficult to rebut and reflect on how this should impact our understanding of it.

1. What is the thesis and why is it important?

1.1. What is relative consistency?

Let T be a theory in the language of set theory, $\mathcal{L}_{\in}$. Recall that a theory T is consistent if we cannot prove both φ and $\neg\varphi$ using assumptions from T. Using some form of Gödel coding we may formulate a statement, $Con(T)$, in the language of arithmetic that says T is consistent; or more formally, we have

$$T \text{ is consistent } \Leftrightarrow \mathbb{N} \models Con(T).$$

Further, if T can interpret PA as ZFC does, then such a statement can be reasonably formulated in $\mathcal{L}_{\in}$. Thus, we might naturally ask whether T can prove its own consistency. Of course, this was scotched by Gödel.

Proposition 1. *(Gödel)* $PA \nvdash Con(PA)$ *if* PA *is consistent.*

A quick perusal of the proof reveals that it continues to hold for any theory T that can interpret PA. It also gives us our first example of a relative consistency proof. It tells that if PA is consistent, then $PA + \neg Con(PA)$ is also consistent. We might say that $PA + \neg Con(PA)$ is *consistent relative* to PA.

That's the basic idea of relative consistency, but we need to take a little more care if we are to formulate an interesting mathematical relation. To illustrate this, note that we *do* think PA is consistent, and so we think that PA cannot prove $Con(PA)$. Thus, it seems reasonable then to think that $PA + Con(PA)$ should be stronger than PA and so, $PA + Con(PA)$ should not be consistent relative to PA. But as we've defined things so far, this is false. To see this note that the statement,

$$Con(PA) \rightarrow Con(PA + Con(PA))$$

is true simply because the consequent is true.[1] The upshot is that if we work in a background theory like ZFC, then every pair of theories for which ZFC can provide a model will be consistent relative to each other; i.e., equiconsistent. Or as a slogan: every pair of consistent theories would be equiconsistent with each other. Thus, the relative consistency relation is rendered trivial for all theories weaker than the background theory we are working in. This would be an uninteresting mathematical relation. The moral of this story is that we need to pay

[1] More specifically, assume standard mathematical conventions and work in ZFC. Then we can define a standard model $\mathbb{N}$ of arithmetic based on ω that satisfies PA. By soundness, this implies that $Con(PA)$ is true. Moreover, since $\mathbb{N}$ is standard it agrees with the universe on all arithmetic statements. Thus, $\mathbb{N} \models PA + Con(PA)$ as required.

attention to where we are standing when we prove these theorems. In particular, if we weaken our background theory to PA, then the statement above is no longer provable there. Or more formally,

$$PA \nvdash Con(PA) \rightarrow Con(PA + Con(PA))$$

since if it were provable an application of the deduction theorem would violate Gödel's second theorem. Thus, if we want an interesting mathematical relation we need to ensure that our background theory is weak enough not to trivialize it.

It's also worth noting the degree of metamathematics in the statement above. We aren't simply providing a counterexample to a conditional. To establish a failure of relative consistency, we need to prove that there is no proof in PA that if there is no proof of absurdity from PA, then there is no proof of absurdity from PA plus the statement that there is no proof of absurdity from PA. Obviously, the formal notation employed above makes it easier to articulate such statements, but we are – I think – far enough down the Gödelian rabbit hole that analyzing their philosophical significance becomes challenging. While it is certainly possible to reason accurately in these domains, it is not so obvious that our naive intuitions about provability come along for the ride.

In this paper, we shall be predominantly concerned with theories that extend ZFC. As such, we'll still have an interesting mathematical relation if we let ZFC be our background theory. With this in mind, we then offer the following definition:

Definition 2. For theories T and S extending ZFC in the language of set theory, let us say that T is *consistent relative to* S, abbreviated $T \leq_{Con} S$ if[2]

$$ZFC \vdash Con(S) \rightarrow Con(T).$$

If $T \leq_{Con} S$ and $S \leq_{Con} T$ we say that S and T are *equiconsistent*, abbreviated $\equiv_{Con}$. If $T \leq_{Con} S$ but $S \not\leq_{Con} T$, let us say that T is *properly consistent relative to* S, *abbreviated* $T <_{Con} S$.

We then observe that we have:

$$ZFC + \neg Con(ZFC) \equiv_{Con} ZFC <_{Con} ZFC + Con(ZFC).$$

Thus, we have a relation on theories that isn't trivial above ZFC. This raises a natural mathematical question: what kind of relation is $\leq_{Con}$? This is the kind of question a mathematician can get stuck into.

1.2. What is the thesis?

For the last sixty years, set theorists have been developing a better understanding of $\leq_{Con}$. The seminal results are from Gödel and Cohen.

Theorem 3. *(1, Gödel, 1940)* $ZFC + CH \leq_{Con} ZFC$*; and*
(2, Cohen, 1963) $ZFC + \neg CH \leq_{Con} ZFC$.

The first result is obtained by defining an *inner model* L of the universe in which $ZFC + CH$ holds. The second is obtained by taking a model of ZFC and using *forcing* to *generically*

[2]It's worth noting that little would change if we'd use PA as a base theory as almost all examples of relative consistency proofs in set theory can be carried out there. See Chapter VII.9 of (Kunen, 2006) for more discussion. One drawback with using PA is that we don't have the soundness and completeness theorems available, which tends to make proofs longer and more tedious.

extend and thus, obtain a model where ZFC is preserved but CH fails. Since the advent of these results, the techniques of inner model theory and forcing have developed substantially and become two of the mainstays of contemporary set theory. This work has lead to a much clearer understanding of the $\leq_{Con}$ relation and an intriguing answer to our question above. Moreover, this answer brings us to the headline of this paper: the *Consistency Hierarchy Thesis*. For a working version, we turn to John Steel:[3]

> If T is a natural extension of ZFC, then there is an extension H axiomatized by large cardinal hypothesis such that $T \equiv_{Con} H$. Moreover, $\leq_{Con}$ is a prewellorder of the natural extensions of ZFC. In particular, if T and U are natural extensions of ZFC, then either $T \leq_{Con} U$ or $U \leq_{Con} T$. (Steel, 2014)

Here we see the intriguing claim from the introduction of this paper. Instead of chaos, it is claimed that we have order. For ease of reference and uniformity of notation, let's break the quote above into its separate claims.

(CHT1) For all natural theories T extending ZFC, there is some extension LC_T of ZFC by a large cardinal axiom such that

$$T \equiv_{Con} LC_T;$$

(CHT2) $\leq_{Con}$ is a prewellordering on natural theories extending ZFC; and
(CHT3) For any pair S, T of natural theories extending ZFC either $T \leq_{Con} S$ or $T \leq_{Con} S$.

Clearly (CH3) follows from (CH2) and later we'll see that there is also a sense in which (CH2) follows from (CH1). Together, we'll call them the *Consistency Hierarchy Thesis* (CHT). Mathematically speaking, this is a surprising and interesting claim. (CH1) tells us that every natural theory is aligned with a large cardinal axiom. (CH2) then tells us that these theories are ordered in about as clean a way as one could want. Beyond being a pleasing arrangement, it also suggests that a serious mathematical idea is being chased, if it is true. But it is important to understand that CHT is not a theorem. The problem is not so much that we don't know whether it's true or not. The problem is just not stated with sufficient precision to even be amenable to proof or refutation.[4] This the first place where we see something extra-mathematical or even philosophical creeping into our discussion. The problem is that two of the terms used in CHT's formulation lack precise definitions: large cardinals; and natural theories.

When we speak of large cardinals, we generally think of them implying the existence of an elementary embedding from the universe into an inner model with certain closure properties.[5] For example, there is a measurable cardinal iff there is an elementary embedding $j : V \to M$ such that j moves at least one ordinal, the least of which is known as its critical point. Despite the availability of workable rules of thumb, there is – at least at present – no definition of large cardinal that captures all known large cardinals in a satisfying way that also leaves room for the future.[6] Nonetheless, we do currently have an enumeration of a large collection of large cardinal

[3] I should note that Steel calls this the *vague conjecture*.
[4] By proof here, I have in mind the kind of proofs that are written by mathematicians. One might also think that philosophers can deliver philosophical proofs, although we shall avoid that usage in this paper. This is pertinent to our discussion of Church's thesis below. An excellent article, which takes a different stance to that in this paper can be found in (Black, 2000).
[5] I'm ignoring smaller large cardinals like inaccessible and Mahlo cardinals here.
[6] Sometimes definitions of restricted class of large cardinal are useful. See for example (Woodin, 2001).

axioms that appear to be sufficient for at least our current purposes and which we don't know how to extend in a meaningful way.[7]

Natural theories, on the other hand, are a larger thorn in our side. They play a crucial role in CHT and they also lack a precise mathematical definition. In his discussion of CHT, Steel offers the following informal characterization:

> By "natural" we mean considered by set theorists, because they had some set-theoretic idea behind them. Here the standards are very liberal, as the many thousands of pages published by set theorists will testify. (Steel, 2014)

The general idea here is that a theory extending ZFC is *natural* if it is the sort of thing a set theorist might come up with when investigating some mathematical project. This is quite vague and indeed, there will certainly be problems at the borderline. Nonetheless, it's not difficult to identify some prototypical *in* and *out* cases.

On the *in*side, we have large cardinal axioms, forcing axioms, determinacy axioms, ultrapower axioms, generalized large cardinal axioms and perhaps dilator axioms (Goldberg, 2022; Kanamori, 2003; Lewis, 1998; Martin, nd; Todorcevic, 2014). Each of these is associated with a project that generalizes a mathematical question beyond the reach of ZFC and searches for axioms that address the problems in the expected way. For a classic example, it is well-known that Vitali sets provide examples of sets of reals that are not measurable. This raises questions about how and where this apparent pathology emerges. In the context of ZFC, the classical descriptive set theorist, Luzin, was able to show that every $\boldsymbol{\Sigma^1_1}$ set is Lebesgue measurable[8], but further progress was hampered by the limitations of ZFC. The use of determinacy axioms proved fruitful here. For example, Kechris and Martin showed that if every game on a $\boldsymbol{\Sigma^1_n}$ set of reals is determined, then every $\boldsymbol{\Sigma^1_{n+1}}$ set is Lebesgue measurable [9]. Thus, by extending ZFC with determinacy axioms a larger family of sets of reals could be tamed. Moreover, the addition of determinacy axioms seems to provide a very natural generalization beyond ZFC of the work carried out by classical descriptive set theorists in the mid twentieth century.[10] We might say that determinacy axioms provide archetypal examples of natural theories extending ZFC.

On the *out*side, typical examples of unnatural theories tend to involve what one might think of as metamathematical as opposed to combinatorial content. The axioms involved often make essential use of coding tricks and self-reference. For a classic example, we might consider the theory obtained by adding the statement $\neg Con(ZFC)$ to ZFC. This gives us a theory that talks about itself using coding and makes the bizarre claim that it is itself inconsistent. Beyond being a very odd thing to say, one might think of it as an unlikely thing for a mathematician to come up with when thinking about and working in set theory. There is a sense in which this theory doesn't talk about sets, but rather about the theory of sets. Riffing on Quine we might think that it is *mentioning* ZFC rather than *using* it. As such, we might think that it fails to satisfy Steel's criterion above. I should, however, say that I am quite skeptical about how robust the distinction between metamathematical and combinatorial content is. While anyone who has

[7] Of course, one can always take a large cardinal and move its successor. So roughly speaking by "meaningful" I mean extending by some means that we haven't already used in the large cardinals that come before it.

[8] See Theorem 29.7 in (Kechris, 1995).

[9] See Exercise 27.14 of (Kanamori, 2003).

[10] For a good discussion of the generalization of classical descriptive set theory see (Maddy, 2011) and (Martin, 1998).

thought about Gödel's incompleteness theorems will be aware that coding and self-reference seem to shift our focus away from the ordinary content of a theory, such content was always already there in the math. We just didn't see it until relatively recently. Moreover, anyone who has tried to understand how the logic of consistency statements works, will know that a new layer of combinatorial content emerges at this metamathematical level in the form of something like a modal logic.[11] Thus in this paper, we shall not take it that the use of metamathematical tools as sufficient for the identification of an unnatural theory.[12]

Despite these reservations, I still think it is often relatively easy to distinguish clear cases of metamathematical and combinatorial content.

Because of these ambiguities, our reasons for thinking that CHT is correct are based on what is essentially empirical evidence. Most of the natural theories that we have happened upon so far have been shown to be equiconsistent with a large cardinal extensions of ZFC. While many open questions remain, there are – as yet – no accepted counterexamples to CHT. As such, CHT has something in common with another famous thesis: the Church-Turing thesis. In that case, a major reason for thinking that it is true is that *all* of the models of informal computation we have come across so far have been shown to be equivalent to Turing machines. But there are also some important differences. While like the Church-Turing thesis, CHT has no accepted counterexamples, unlike the Church-Turing thesis, many equivalences remain unproven. For example, while the very educated guess is that the addition of a supercompact cardinal is equiconsistent with the addition of the proper forcing axiom to ZFC, this problem has remained open for over fifty years. So in contrast to the Church-Turing thesis, the picture with CHT is very far from complete, although there seems to be no compelling reason for pessimism.

All of this puts CHT in a remarkable position. I think it's clearly a claim that warrants mathematical attention. The $\leq_{Con}$ ordering is non-trivial and there is structure to be investigated. Indeed, the study of the $\leq_{Con}$ ordering has frequently provided some of the deepest results in set theory and inspired the development of new techniques. But CHT also makes use of extra-mathematical content in the form of large cardinals and natural theories. This certainly makes for an intriguing collection of set theoretic problems. But beyond this, is there any philosophical content to CHT? Or is this just another curio in a minority sport?

1.3. Why is it important?

Now that we've explained why CHT is a mathematically interesting problem, I want to shift our focus to its philosophical and foundational significance. I aim to demonstrate that CHT exerts a weighty influence on our understanding of set theory today. In particular, I will argue that: CHT provides a kind of evidence for the importance of large cardinals axioms; and CHT provides an explanation for the lack of disagreement between strong set theories with regard to concrete mathematical questions.

1.3.1. The importance of large cardinals

Recall that (CHT1) says that every natural theory is equiconsistent with a large cardinal extension of ZFC and that (CHT2) says $\leq_{Con}$ pre-well-orders the natural theories. The latter

[11] See, for example, (Boolos, 1984).

[12] This means that the discussion of putative counterexamples to CHT in Section 2 will need to be quite detailed and often formal. We shall need to demonstrate *why* certain metamathematical techniques give us unnatural theories.

of these is particularly significant since it would tell us that natural theories are ordered in a extremely tidy manner. But more than this, it would establish that our initial worries about chaos in the array of theories beyond ZFC were premature: we would have a much tamer realm than we could have reasonably hoped for. I take it that this insight is of philosophical and foundational importance for set theory and both mathematics and logic more generally. I would now like to demonstrate that there is a sense in which (CHT2) follows from (CHT1). Thus, we shall see that the purported tameness of natural theories is crucially dependent on our use of large cardinal axioms.

To see this, we start by observing that – unlike natural theories, in general – large cardinal extensions of ZFC tend to be very easily ordered by consistency strength. To illustrate this, we describe three grades of ordering that often occur in these relationships.[13] First, we note that very often a large cardinal that is stronger (in terms of consistency strength) than another will already satisfy the characteristic property of the weaker large cardinal. For example, measurable cardinals are stronger than inaccessible cardinals, and whenever κ is a measurable cardinal, κ is inaccessible. Thus, the consistency of ZFC plus a measurable cardinals implies the consistency of ZFC and an inaccessible cardinal. Second, we observe that even if the first condition doesn't hold, there will often be an example of the weaker large cardinal below the stronger one. For example, Woodin cardinals are stronger than measurable cardinals, but Woodin cardinals are not typically measurable. Nonetheless, there will be many measurable cardinals below a Woodin cardinal. This also gives the desired relative consistency fact. Third and finally, even when the former two situations don't occur, we may almost always use the stronger large cardinal to very easily define a model[14] where the weaker large cardinal axiom is satisfied. For example, Woodin cardinals are stronger than strong cardinals, but generally, there are no strong cardinals below a Woodin cardinal. Nonetheless, whenever δ is a Woodin cardinal, V_δ will think there are many strong cardinals. Thus again, we obtain the desired relative consistency claim.

There are, of course, examples that don't fit into any of the three grades, but such exceptions are comparatively rare.[15] The point that we are trying to emphasize is that the determination of consistency strength relationships between large cardinals is quite elementary. Moreover, it is easy to see that theories extending ZFC whose relative consistency can be ascertained by one of the three grades above will be pre-well-ordered by consistency strength. Thus, while there is some vagueness in the concept of a large cardinals, we can have much greater confidence that the strength of large cardinal axioms are pre-well-ordered, in contrast to the more difficult question regarding arbitrary natural theories. Now if we combine this observation with (CHT1), we see that (CHT2) follows immediately. If every natural theory is equiconsistent with a large cardinal extension of ZFC, then clearly the consistency strength of natural theories is also pre-well-ordered.

Thus, we see that large cardinals play a crucial supporting role in CHT, but in the – so to speak – other direction, CHT and the considerations above also tell us that large cardinals are a very special kind of axiom. They provide the underlying spine that brings order to the chaotic world opened by Gödel almost one hundred years ago. But before we fall prey to delusions of grandeur, I'd prefer to avoid pushing on and arguing that this somehow gives us evidence that

[13] We shall provide a more detailed sketch of one of these arguments in Section 3.2.1., but for now we'll content ourselves with a little vagueness.

[14] In particular, the model will be a rank initial segment of the universe satisfying the strong axiom.

[15] An obvious example occurs, if we admit that saying $0^\#$ exists is a large cardinal axiom. Then $0^\#$ is stronger than an inaccessible cardinal, but there are models where $0^\#$ exists that have no rank initial segment with an inaccessible cardinal. There are of course inner models, in particular, L that will have inaccessible cardinals, but I'm excluding this from the template since arguably L is a more sophisticated internal model to define than a rank initial segment.

large cardinal axioms are true. I think any attempt to answer that question would open a can of philosophical worms outside our interests here. But more importantly, I think addressing such questions could detract from what I take to be the surefooted lesson we can take from these observations. If one is interested in strong mathematical theories, then to the best of our current knowledge, large cardinal axioms are an essential tool in this investigation.

1.3.2. The absence of disagreement

For our next illustration of the foundational significance of CHT, we turn to the phenomenon of disagreement, or rather the lack thereof between strong natural theories regarding concrete mathematics. We've seen above that CHT provides order to the variety of theories extending ZFC, but we might still worry that the remaining options might affect ordinary mathematics regardless of whether these options are equiconsistent with a large cardinal axiom. For example, we know from Gödel that the theory of analysis is incomplete. And we know that ZFC and its extensions can fill in some of the gaps left by this incompleteness. Thus, there is a natural worry that the different theories extending ZFC will give different answers to problems in ordinary mathematics, in particular questions about the real numbers. It turns out that this is not the case. To explain this, we first need a refinement of (CHT1). In actual practice to date, we not only find that natural theories are equiconsistent with large cardinal extensions of ZFC, but that there is also a certain uniformity in the techniques used for the proofs of these facts. In particular, we generally find that:

(GE) Models of natural theories are obtained through *generic extension* of models of large cardinal extensions of ZFC; and
(IM) Models of natural theories deliver *inner models* of large cardinal extensions of ZFC.

This is a much stronger relationship than mere equiconsistency. For example, the models obtained by either generic extension or inner model theory retain exactly the same ordinals as the model we started with. This kind of structural preservation leads to theories that must agree with each other on substantial fragments of ordinary mathematics. For example, it can be argued that any pair of natural theories extending ZFC must agree on all Π^1_2 statements.[16] Moreover, as we consider strengthenings of ZFC by large cardinals, agreement on more mathematics can be obtained. Rather than describe a specific example of this phenomenon, we shall offer a general sketch of the proof strategy with the aim of highlighting the role played by CHT.

Theorem 4. *(Sketch[17]) Let T and S be natural theories extending ZFC; and let Γ be a fragment of the theory of analysis. Now suppose that there are large cardinal extensions LC_T and LC_S of ZFC, which are comparable by one of the three grades discussed above, such that:*

(i) LC_T interprets T via forcing that preserves Γ;[18]
(ii) T interprets LC_T via an inner model interpretation that preserves Γ;[19]
(iii) LC_S interprets S via forcing that preserves Γ; and

[16] See Steel's article in (Feferman et al., 2000; Maddy and Meadows, 2020; Meadows, 2021; Steel, 2014) for more detailed discussion of this.
[17] I've called this "theorem" a "sketch" since it is not stated with proper mathematical precision. Like the statement of CHT, it makes use of the imprecise terms: natural theory and large cardinal. Nonetheless, it seems it seems beneficial to draw out and highlight the underlying strategy in these arguments.
[18] By this we mean that LC_T can prove there is a poset $\mathbb{P}$ such that: $\Vdash_{\mathbb{P}} T$; and for all $\varphi \in \Gamma$, $\varphi \leftrightarrow \Vdash_{\mathbb{P}} \varphi$.
[19] By this we mean that T can prove there is a definable inner model N such that: $(LC_T)^N$; and for all $\varphi \in \Gamma$, $\varphi \leftrightarrow \varphi^N$.

(iv) S interprets LC_S via an inner model interpretation that preserves Γ.

Then T and S agree upon Γ; i.e., there is no sentence $\varphi \in \Gamma$ such that $T \vdash \varphi$ and $S \vdash \neg\varphi$, and there is no sentence $\varphi \in \Gamma$ such that $S \vdash \varphi$ while $T \vdash \neg\varphi$.[20]

Proof. (*Sketch*) Using CHT, we know that $LC_S \leq_{Con} LC_T$ or $LC_T \leq_{Con} LC_S$. We first the former and thus, that LC_T is stronger than LC_S. In this case, it will suffice to show that for all $\varphi \in \Gamma$ if $S \vdash \varphi$, then $T \vdash \varphi$ also. We proceed by contraposition and let $\mathcal{M}$ be a countable model satisfying $T \cup \{\neg\varphi\}$. Then let $N^{\mathcal{M}}$ be an inner model of $\mathcal{M}$ given by (2), so we $N^{\mathcal{M}}$ is a model of $LC_T \cup \{\neg\varphi\}$. Now since LC_T is stronger than LC_S as witnessed by one of our three grades of comparison, we may fix a model $V_\alpha^{N^{\mathcal{M}}}$ inside $N^{\mathcal{M}}$ satisfying LC_S.[21] Moreover, since $V_\alpha^{N^{\mathcal{M}}}$ is a rank initial segment of $N^{\mathcal{M}}$, they share the same theory of the reals, so $V_\alpha^{N^{\mathcal{M}}}$ also satisfies $\neg\varphi$. Finally, we use (3) to obtain a generic extension $V_\alpha^{N^{\mathcal{M}}}[G]$ of $V_\alpha^{N^{\mathcal{M}}}$ that satisfies $S \cup \{\neg\varphi\}$. Thus, we've established that $S \nvdash \varphi$. A similar argument works in the second case. □

For an example of this, let T be ZFC plus Δ^1_2-determinacy; and let S be ZFC plus the existence of a precipitous ideal. Then it can be shown via inner models and forcing that T is equiconsistent with LC_T where LC_T is the extension of ZFC with a Woodin cardinal.[22] By similar means it can be shown that S is equiconsistent with LC_S where LC_S is the extension of ZFC with a measurable cardinal. Moreover, it can be seen that these interpretations satisfy conditions (1) through (4) for Π^1_3 statements.[23] Finally, it is not difficult to see that $LC_S <_{Con} LC_T$ and so the argument strategy above tells us that T and S agree on Π^1_3 statements. The hard work in proving examples of this theorem sketch goes into establishing that the theories in question are sufficient to deliver conditions (1) to (4) for Γ. But the underlying structure of the proof is relatively simple: we move from models of natural theories to inner models of large cardinals in the spine and then slide down and generically extend outward to a model of another natural theory.

Thus, in addition to establishing the value of large cardinals as an instrument for understanding strong set theories, we see that CHT also provides a valuable explanation of why disagreement about concrete mathematics does not seem to occur between strong natural theories. Thus, we've now demonstrated that beyond being an interesting mathematical problem, CHT has deep philosophical implications for set theory and the way in which it provides a foundation for mathematics. Hopefully, it is clear that if a counterexample to CHT were discovered and accepted, then set theory as we know it would undergo a radical shift in perspective.

2. Attempts at refutation

Now that we have a clearer idea of what the Consistency Hierarchy Thesis is and why it is foundationally significant, we are going to shift our attention to the question of whether it is justified. We saw above that our reasons for believing it rest upon the fact that we have found many extensions of ZFC that satisfy it, and the fact that no counterexamples to it have been discovered and accepted by the set theory community. We likened CHT to the Church-Turing

[20] It might be more apropos to say that *S and T don't disagree on Γ*, but I've opted for the simpler slogan to avoid using too much negation.
[21] Here I'm abusing notation a little and allowing that α could be $Ord^{\mathcal{M}}$ for the case of the first grade.
[22] See Theorem 32.17 in (Kanamori, 2003).
[23] See Theorem 15.6 in (Kanamori, 2003).

thesis, but noted that in contrast the picture with CHT is much less complete. While many equivalences with large cardinal axioms have been established many also remain open. As such, a much wider flank is left exposed to the possibility of a counterexample. In this section, my goal is to consider the prospects for such a campaign. We shall begin by considering a counterexample that almost nobody thinks is significant. Then we shall consider a series of more serious attempts from a recent paper by Joel David Hamkins (2025). While we shall push back on Hamkins' intriguing examples, the real value of this work will be in illuminating more about just what kind of problem CHT presents.

2.1. A counterexample that nobody accepts

In this section, we give a short proof that there is a pair of theories whose consistency strengths are incomparable and consider what this means for CHT. The example is taken from Theorem 3 of (Hamkins, 2025), but it will benefit us to provide a quick sketch in order emphasize the technical nature of the reasoning involved and to illustrate how little purchase our ordinary intuitions about proof have in this arena. In particular, this will help us get a better understanding of why people are not impressed by it as a putative counterexample to CHT. Rather than providing a detailed discussion of notational conventions, I'll just note that I'm aiming to follow standard conventions as one might find in (Lindström, 2003).[24] Recall that if η is a Rosser sentence for the theory T, then η says that if d is (code of) a T-proof of η, then there is some $e < d$ that is a T-proof of $\neg\eta$.

Theorem 5. *Let η be the Rosser sentence for $ZFC + Con(ZFC)$ and let T denote this theory. Then supposing that $ZFC + Con(ZFC)$ is consistent, the consistency strengths of $Con(ZFC + \eta)$ and $Con(ZFC + \neg\eta)$ are incomparable; i.e.,*

(i) $ZFC \nvdash Con(ZFC + \neg\eta) \rightarrow Con(ZFC + \eta)$;
(ii) $ZFC \nvdash Con(ZFC + \eta) \rightarrow Con(ZFC + \neg\eta)$.

Proof. (*Sketch*) (1) Our plan is to show there is a model $\mathcal{M}$ of ZFC satisfying:

(i) $Con(ZFC + \neg\eta)$; and (ii) $\neg Con(ZFC + \eta)$.

For (i) we note since η is a Rosser sentence for T, we may fix a model $\mathcal{M}$ satisfying $T + \neg\eta$. Then unpacking the definition of $\neg\eta$, we see that $\mathcal{M}$ thinks there is a T-proof d of η and there is $e < d$ where e is a T-proof of $\neg\eta$. It can be seen that this statement is Σ^0_1 and so since ZFC is Σ^0_1-complete, we see that $\mathcal{M}$ thinks ZFC can prove it. Thus, $\mathcal{M}$ thinks that $ZFC \vdash \neg\eta$; i.e., $\mathcal{M} \models \neg Con(ZFC + \eta)$. For (ii), we simply observe that in $\mathcal{M}$ we have both $ZFC \nvdash \bot$ and $ZFC \vdash \neg\eta$ and so $ZFC \nvdash \eta$. Thus, $\mathcal{M} \models Con(ZFC + \neg\eta)$ as required.

(2) We show there is a model $\mathcal{M}$ of ZFC satisfying:

(i) $Con(ZFC + \eta)$; and (ii) $\neg Con(ZFC + \neg\eta)$.

[24] Of course, that book is focused on arithmetical theories rather than set theories but this makes little difference since ZFC is an essentially reflexive theory that interprets PA. I'll also note that when context requires it below, we shall have recourse to be more precise about our notational conventions.

For (i), we start by using Gödel's second theorem to obtain a model $\mathcal{M}$ satisfying:

$$[T + \eta] + \neg Con([T + \eta]).$$

Then with a little Rosser-style reasoning, it can be seen that $\mathcal{M}$ satisfies the statement: "there is a T-proof e of $\neg\eta$ and for all $d < e$, d is not a T-proof of η." And since this statement is Σ^0_1, we see that ZFC is able to prove it and so $\mathcal{M}$ thinks ZFC proves it too. With a little more work, it can be seen that ZFC proves that the statement implies η and so $\mathcal{M}$ thinks that $ZFC \vdash \eta$; i.e., $\mathcal{M} \models Con(ZFC + \neg\eta)$. For (ii), we now know $\mathcal{M}$ thinks that $ZFC \nvdash \bot$ and $ZFC \vdash \eta$. Thus, $\mathcal{M} \models Con(ZFC + \eta)$ as required. □

Thus, we see that the theories $ZFC + \eta$ and $ZFC + \neg\eta$ have consistency strengths that are incomparable. So we see very clearly that $\leq_{Con}$ is not a pre-well-ordering on all theories. What should we make of this? Why shouldn't we think of this as being a counterexample to CHT? I suppose for some people it might be, but the general consensus is that this counterexample does not succeed. The obvious culprit is the unnaturalness of the theories involved. But why should we think that $ZFC + \eta$ and $ZFC + \neg\eta$ are unnatural? First recall that if we put an intuitive gloss on things, η says something like

> If I'm provable, then I have already been refuted.

It's quite an odd sentence. It's conjured using coding and the diagonal lemma to achieve the effect of self-reference. Moreover, it's relatively obvious that the resulting theories were cooked up for the specific purpose of delivering a counterexample. We suggested above that these were indicators of unnaturalness and I think it's probably fair to say that the tools used above are not generally elements of the ordinary mathematician's toolkit. It's not the sort of theory that one would expect a mathematician doing set theory to come up while they were working with set theory when it is understood as a theory of sets. As such, I think we have reason to think that they do not satisfy Steel's criterion for being a natural theory.

But we can also make a more specific complaint in this case. Assuming we are comfortable accepting some large cardinals, it is easy to see that η is simply true.[25] But what do we learn from this? First, note that since any ω-model of ZFC will satisfy $Con(ZFC)$, we see that if η is false in a model $\mathcal{M}$ of ZFC then $\mathcal{M}$ cannot be an ω-model. I think it would be very strange to accept a theory extending ZFC that cannot be satisfied in a model with the genuine natural numbers. We'd have good reason to think that such a theory is defective. But more than this, it is also easy to see that $ZFC + \neg\eta$ is ω-inconsistent.[26] As is well-known, Gödel's original proof of the incompleteness theorem merely assumed that his target theory was ω-consistent rather than simply consistent (Gödel, 1986). While the result was later improved by Rosser, the assumption of ω-consistency was thought to be sufficient for the philosophical impact of his theorem to be felt. As such, it seems reasonable to exclude $ZFC + \neg\eta$ from the realm of natural theories.

Hopefully, this example illustrates the magnitude of the problem of refuting CHT. It is not sufficient to provide a pair of theories whose consistency strengths cannot be be compared. We must also show that those theories are natural. In the absence of any precise

[25] Very briefly, suppose there is an inaccessible cardinal κ and suppose toward a contradiction that η is false. Then there is a $(ZFC + Con(ZFC))$-proof of η and so η is true in every model of $ZFC + Con(ZFC)$. V_κ is such a model, so η would be true there, and since V_κ and the universe shares their natural numbers, η would be true, which contradicts our initial assumption.

[26] It is easy to see that $ZFC + \neg\eta$ must prove that there is some d that is a $(ZFC + Con(ZFC))$-proof of η; and yet for all $n \in \omega$, $ZFC + \neg\eta$ will also prove that n is not a $(ZFC + Con(ZFC))$-proof of η.

definition, a significant part of the counterexample challenge is philosophical, or at least, extra-mathematical. In particular, we need to mount an argument, as opposed to a proof, establishing that the theories in question are natural. This makes for quite a strange mathematical problem, and one that might not be amenable to a definitive resolution.

2.2. Some counterexamples that probably don't succeed

Our goal now is to consider some serious attempts to answer this refutation challenge. For this, we turn to Hamkins' recent paper on the topic (Hamkins, 2025). Among other things, this paper provides a helpful and elegant overview of literature regarding CHT. Hamkins then proposes a number of putative counterexamples to CHT including an impressive array of generalizations of the techniques involved.[27] Given the discussion of the previous section, we shall be most interested in the philosophical argumentation toward the naturalness of the theories involved, as we have seen that unnatural counterexamples are not difficult to find. As such, we'll focus on three relatively simple examples from Hamkins' paper that I think offer the most compelling philosophical defenses. The first is based on the idea of using some sort of calculation to find good axioms. The second is based on an alternative approach to enumerating our theories that takes care not to add dubious axioms. The third generalizes and addresses a problem with the first proposal by restricting our attention to the kinds of model that set theorists are generally more interested in. We shall be critical of each of these proposals and provide reasons why none of them is successful in establishing naturalness. However, I'd like to stress that despite the fact that Hamkins' paper is the apparent target of these criticisms, our real target sits in the background. There have been very few serious attempts to reject CHT, so Hamkins' paper should be recognized – at the least – for providing valuable clarification of just what is at stake. Our analysis of Hamkins' work will put us in a position to put CHT itself on trial in the final section of this paper.

2.2.1. Computing our axioms

Given that the essential objective of these counterexamples is arguably philosophical, it seems germane to begin with a kind of thought experiment. Suppose at some time in the not too distant future, most pure mathematicians have come to incorporate the use of computers into their mathematical practice. Even in the pursuit of most pure mathematics, many an hour is spent trudging through routine calculations that must be checked but yield little in the way of insight. The mathematicians of this future time have come some way to freeing themselves from these bonds. Now suppose that set theorists have come to consider a certain class of interesting mathematical problems; and they have figured out that for just about every problem in this class, there is a particular number $n \in \omega$ such that ZFC plus n inaccessible cardinals is, somehow, the optimal theory to address that problem. Supposing that the underlying class of problems is of independent mathematical interest, we would have reason to think that each of these extensions of ZFC is natural. Indeed, this move is arguably unnecessary since extensions of ZFC by inaccessible cardinals are almost universally regarded as natural theories. But suppose also that the calculation of the number of inaccessible cardinals appropriate to a particular problem is very tedious to calculate. Perhaps it involves a seemingly endless

[27] For another analysis and response to Hamkins' examples see (Grotenhuis, 2022). This dissertation takes up a Lakatosian approach that aims to use Hamkins' examples as a means to better isolate what a natural theory is. I am more pessimistic about the prospect of any satisfying analysis of natural theories, but this dissertation offers an intriguing point of view on this problem.

sequence of cases all based on a simple trick. The mathematicians of this fictional time knows what to do: they design a computer program that takes the statement of a problem from the relevant class and then after some time, returns the number of inaccessible cardinals required. This might result in a partial computable function from the naturals (as Gödel codes of statements of problems) to the naturals (as number of inaccessible cardinals to add to ZFC).[28] Moreover, the outputs of this partial function straightforwardly deliver natural extensions of ZFC. Let $\psi : \omega \rightharpoonup \omega$ denote such a partial computable function. Then for all $n \in \omega$

$$ZFC + \text{there are } \psi(n) \text{ inaccessible cardinals}$$

gives us a natural theory whenever $\psi(n)$ halts.

With our thought experiment complete, we now have an argument for the claim that we have a collection of natural theories. So far so good, but how to these theories play with CHT? While we have some comments to make in a moment, I think it is only fair to give Hamkins the mic dropping moment his intriguing result deserves.

Theorem 6. *(Hamkins, 2025) There is a partial computable function $\psi : \omega \rightharpoonup \omega$ such that the theories*

$$ZFC + \text{“there are } \psi(n) \text{ inaccessible cardinals.”}$$

for $n \in \omega$ have pairwise incomparable consistency strengths, assuming there are infinitely many inaccessible cardinals.[29]

Speaking loosely, this theorem appears to tell us that if we use the methodology suggested in the thought experiment above, then we could land in a situation were we have infinitely many natural theories, none of whose consistency strengths can be compared. As such, we have a much more serious challenge to CHT than we saw in Section 2.1. We seem to have both natural theories and nonlinearity.

I now want to push back against this putative counterexample from a couple of related angles. First, I'd like to consider how the theorem above is proved. Without getting too deep among the weeds, the essence of the proof is in finding the right partial computable function $\psi : \omega \rightharpoonup \omega$. The ψ required is very special: it is such that for any $m \neq n \in \omega$ there will be a model of ZFC where $\psi(m) < \psi(n)$ holds and another where $\psi(m) > \psi(n)$. Thus very informally, there are models of ZFC where one theory is stronger than the other and other models where the converse occurs. The rest of the argument is relatively straightforward. In order to obtain such a remarkable ψ, Hamkins makes an ingenious argument deploying the recursion theorem,[30] which states that for any total computable function $g : \omega \to \omega$ there is some $e \in \omega$ such that the partial computable function determined by e is the same as that determined by $g(e)$; or more formally, $\varphi_{g(e)} \simeq \varphi_e$. For the uninitiated, the recursion theorem is a close cousin of the notorious diagonal lemma used in proving Gödel's first incompleteness theorem. In both cases, we obtain a kind of fixed point through the use of a technical device simulating something like self-reference. The reader will recall that this was one of our superficial indicators of unnaturalness in a theory.

However, ψ has another remarkable feature that prompts further questions regarding its naturalness: the ψ used in the theorem never halts on any input in an ω-model! Indeed, this

[28] Recall that we merely assumed that *just about every* problem in the class had a number of inaccessible cardinals associated with it. Thus, the function could be properly partial rather than total.

[29] In fact, Hamkins proves a stronger results but this is more than enough for our purposes.

[30] For more details on the history of this kind of argument and some related theorems see (Hamkins, 2025).

fact is crucial for the proof of its special properties. Doesn't it seem a little odd to call a theory natural when – modestly assuming the actual natural numbers are well-founded – that theory is not properly defined? Moreover, if we return to the thought experiment that began this section, why would we select such a function for such a job? Hamkins is, of course, aware of this issue and offers an interesting response that I'll now try to motivate.[31] Suppose that in our thought experiment, our problems correspond to slightly different extensions of ZFC. Rather than saying that "there are $\psi(n)$ inaccessible cardinals," we might instead say "there are *at most* $\psi(n)$ inaccessible cardinals." So the natural theory returned puts a kind of bound on the number of inaccessible cardinals appropriate to the problem. Now when we feed ψ the code n of some problem, and the calculation of $\psi(n)$ does not terminate, an obvious interpretation suggests itself. The calculation doesn't terminate since there is no finite bound on the number of inaccessible cardinals appropriate to the problem at hand. Thus, when $\psi(n)$ doesn't terminate we could interpret this as determining the theory of ZFC with infinitely many inaccessible cardinals appended to it. I think this response is coherent and fits neatly with the philosophical motivations of these theories. So there is some reason to think that the naturalness of the initial thought experiment is preserved. But I also think this response makes such theories seem unnatural on other grounds. Suppose that given the considerations above, we countenance the use of partial computable functions in determining theories and further, we assume that theories are still determined even when that partial computable function does not halt. In the example above, we can prove that ψ never halts, so we know what to do. But as a general principle, this seems very strange. In general, this will lead us into positions where we are describing theories whose axioms cannot – even in principle – be determined by finitary means. Traditionally, foundational theories are prized for the simplicity of their axiomatization. While ZFC isn't finitely axiomatizable, it's arguably the next best thing. There's a finite set of very simple instructions that anyone can follow to determine whether *or not* some formula is an axiom of the theory. Without at least this, it's difficult to understand how we could – as finite beings – make use of such a theory. But this is what Hamkins is offering us here. The crucial point is that in a computable axiomatization, we use a total computable function, while Hamkins admits partial computable functions and assigns them values even when they do no halt. The only way for us to competently use such a theory would be through an impossible solution to the halting problem. I think this gives us a more substantive reason to doubt the naturalness of such theories.

My final quibble with this example is more subtle, but I also think more pressing. The issue this time is metamathematical and concerns how the theorem above should be stated in order to be in accord with our informal interpretation of it. The thinking here reminds me of the way in which a great painting often demands that its viewer think about where the artist was standing when they produced the image. Similarly in this case, we need to think about where we are standing when stating Theorem 6. In particular, I am concerned with where the computer, which figures out how many inaccessible cardinals to add to our theory, is being operated. Is it here with us, or in some other abstract context? My contention is that this computer is not being operated here, but rather in some abstract, nonstandard model of set theory that is quite confused about which computations terminate and which do not. The easiest way to see this is that in our world, the crucial function ψ never halts on any input that it is given. Thus, if the

[31] I do things a little differently to Hamkins in order to keep with the flow of this discussion, but I hope to have captured the spirit of his response.

calculation of $\psi(n)$ for $n \in \omega$ were undertaken here, then we'd just learn that for all $m < n \in \omega$

$$ZFC + \text{“there are } \leq \psi(m) \text{inacessibles”} \equiv_{Con} ZFC + \text{“there are } \leq \psi(n) \text{inacessibles”}$$

since neither $\psi(m)$ nor $\psi(n)$ are defined and so they both say that there are infinitely many inaccessibles.[32] I think this interpretation is in good accord with the thought experiment from the beginning of this section. If we want to check how many inaccessibles to add to ZFC in order to align with some problem, then it makes good sense to use a computer if the required calculation is long and tedious. But we want the computer to do essentially what we would have done, just without our having to do it; we don't want it to operate in some nonstandard environment and return a value when no such value is due. The theory we are interested in, is the theory that the calculation delivers in this world. But the crucial trick of Hamkins example relies on doing the exact opposite of this. We take the computer program and we run it in models where it does halt, but after a nonstandard number of steps. These are the only models where we get $\psi(m) < \psi(n)$ and $\psi(m) > \psi(n)$. But in such models

$$ZFC + \text{“there are } \leq \psi(m) \text{inacessibles”}$$

has a completely different meaning that it does here in our world. In those worlds, it says that there at most $\psi(m)$ inaccessible cardinals, while here it says there are infinitely many. More succinctly, we might put the problem as follows: we set out to compare two theories that interested us and ended up comparing two completely different theories instead.[33]

Before we move to the next example, I want to take a more strategic perspective on what we've seen above. While I've offered some criticisms of Hamkins' argument toward the naturalness of the theories he proposes, I don't want want to pretend that my criticisms are decisive refutations. This is, in part, because I have a congenial dialectical position in this debate. Even though it was quite vaguely defined, naturalness is a high bar for a theory to meet. As such, to push back against Hamkins' example it would probably suffice to just show that the water is muddy and that plausible controversy lurks at every turn. This would be enough to tarnish the credentials of a putative natural theory. While I think we have done more than this above, I think it is important to note which way the deck is stacked in this game. Indeed, I think this should have some effect on how we evaluate the significance of the Consistency Hierarchy Thesis.

2.2.2. Being very careful

For our next example, we again begin with a thought experiment to motivate the claim that a proposed extension of ZFC is natural. As we noted at the beginning of this paper, Gödel's

[32] Here we are using the fix for cases where ψ doesn't terminate. If we didn't use that, we wouldn't have any theories at all.

[33] It's important to note that this point is quite dependent on the thought experiment suggested above. As such, there could well be another way of motivating something like the theory above that is not exposed to this style of objection. To highlight the dependence, we note that ZFC itself is a computably axiomatizable theory that is interpreted quite differently in other worlds. For example, if we suppose that $ZFC \nvdash Con(ZFC)$, then by completeness we may fix a model $\mathcal{M}$ satisfying $ZFC \cup \{\neg Con(ZFC)\}$. However, the sentence $\neg Con(ZFC)$ is interpreted quite differently in $\mathcal{M}$ than in our background universe. This is because the statement $\neg Con(ZFC)$ requires the use of a formula to represent the codes of sentences from ZFC. Then since $\mathcal{M}$ satisfies $\neg Con(ZFC)$, $\mathcal{M}$ will think some largest natural number x such that the axioms of ZFC whose codes are below x are consistent. But the Reflection Theorem implies that every truly finite subset of ZFC is consistent in $\mathcal{M}$. This means that x must be a nonstandard element of $\omega^{\mathcal{M}}$. But then we we apply the Reflection Theorem inside $\mathcal{M}$ to the axioms whose codes are below x, we get a model $\mathcal{N}$ satisfying every genuine axiom of ZFC. So there is a sense in which $\mathcal{M}$ thinks ZFC is consistent after all. See the discussion on page 146 of (Kunen, 2006) for a more patient rendering of this argument. The upshot though is that the way in which theories dependent on computation are interpreted can be very dependent on context. Intuition certainly seems to struggle. This may give us a different reason to doubt that we are doing something natural and indeed later, in Section 3.2., this will give us some reason to reevaluate how we should interpret such theories.

incompleteness theorems can be understood as leaving open a chaotic realm beyond ZFC. While we saw that CHT goes some way toward taming this menagerie, Gödel's theorems also prompt another worry. To see this, first recall that ZFC is sometimes thought to delimit what we can assume, without remark, in a mathematical publication. If you can prove φ in ZFC, then φ is just a theorem in your paper; but if you also need a measurable cardinal, then your theorem is: φ, if there is a measurable cardinal. Roughly speaking, we might say that ZFC sets out the limits of consensus mathematical knowledge. But then by this standard, Gödel's second theorem would tell us that, for all we know, ZFC is inconsistent since this cannot be proved in ZFC, and we are all just wasting our time. Of course, there have been many attempts to argue on different grounds for the consistency of ZFC. Perhaps the best of them relies on the fact that – as of now – no proof of inconsistency has been found from the axioms of ZFC in over a century of use. But for those of a more anxious temperament, such empirical fodder could come as cold comfort. This is where Hamkins' next proposal enters our story.

> ... 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^o. In order to halt the enumeration we don't require an explicit contradiction in ZFC, but rather only a proof that there is such a contradiction (Hamkins, 2025).

Thus, rather than naively admitting all of ZFC into our mathematical knowledge base, we also perform an extra safety check with the goal of obtaining greater confidence in our theory. At face value, this extra move seems prudent and sensible. As such, if we can formulate a theory that achieves this effect it seems reasonable to call it natural. I now want to provide a more detailed exposition of how to understand the cautious enumeration. We will then be able to generalize this cautious position and obtain a collection of theories that form an infinite descending chain in the consistency hierarchy and thus, a further challenge to CHT.

We start with some general remarks about notation and enumeration of theories.[34] We shall dive a little deeper into the details than in the previous section. This is the most technical section of the paper and the reader may be best served by glossing it on a first reading. Following Lindström (2003), let us take it that ZFC denotes its axioms rather than, the more traditional closure of the axioms in first order logic. Suppose then that $\ulcorner\cdot\urcorner : \mathcal{L}_\in \to \omega$ is a Gödel coding; i.e., a computable injection from the formulae of set theory into the natural numbers. Now we know that there is a Σ^0_1 formula[35] $\tau(x)$ that enumerates the axioms of ZFC in such a way that for all $\varphi \in \mathcal{L}_\in$

$$\varphi \in ZFC \;\Leftrightarrow\; V_\omega \models \tau^{\ulcorner}\varphi^{\urcorner}$$

and further

$$\varphi \in ZFC \;\Rightarrow\; PA \vdash \tau^{\ulcorner}\varphi^{\urcorner}$$

and

$$\varphi \notin ZFC \;\Rightarrow\; PA \vdash \neg\tau^{\ulcorner}\varphi^{\urcorner}.$$

We have this since ZFC is computably axiomatizable and $\tau(x)$ witnesses this. Moreover, we shall suppose τ works by simply identifying the codes of the axioms which are not schema and

[34] In general, we aim to follow a slightly modernized version of the notation and terminology of (Lindström, 2003).

[35] Since we are working in set theory, we shall think of a Σ^0_1 formula a Σ_1 formula of the Lévy hierarchy that is then relativized to V_ω. See Kunen (2009) for a thorough treatment of this approach.

for Separation and Replacement it employs some very simple form of pattern recognition. We shall also say that $\tau(x)$ *binumerates* ZFC and we shall fix such a τ for the remainder of this section. For an arbitrary Σ_1^0 formula $\sigma(x)$, we shall let $B_\sigma{}^\ulcorner\varphi^\urcorner$ be a Σ_1^0 statement saying that there is a natural number coding a proof of φ using codes of formulae ψ such that $\sigma^\ulcorner\psi^\urcorner$ holds. We then write $Con(\sigma)$ to mean $\neg B_\sigma{}^\ulcorner\emptyset \neq \emptyset^\urcorner$; i.e, σ defines a set of axioms that are consistent. Thus, where we've written $Con(ZFC)$, we may now write $Con(\tau)$.[36] Let $Bew_\sigma(^\ulcorner\varphi^\urcorner, d)$ mean that d is the code of a derivation of φ from assumptions whose codes satisfy $\sigma(x)$. These technical considerations are important in this case since they are essential to our definition of an enumeration. For our purposes, $\tau(x)$ determines an enumeration of ZFC in the form of an ordering $\prec_\tau$ such that for all formula φ, ψ of $\mathcal{L}_\in$

$$\varphi \preceq_\tau \psi \;\Leftrightarrow\; \tau^\ulcorner\varphi^\urcorner \wedge \tau^\ulcorner\psi^\urcorner \wedge {}^\ulcorner\varphi^\urcorner \leq {}^\ulcorner\psi^\urcorner.$$

Thus informally, φ precedes ψ if they are both axioms of ZFC and the code of φ precedes that of ψ. Finally, we include a technical notation that will be helpful in articulating Hamkins' cautious enumeration. For arbitrary $\sigma(y)$ from $\mathcal{L}_\in$, let $(\sigma|x)(y)$ be the formula: $\tau(y) \wedge y \leq x$. Thus, $(\sigma|x)(y)$ defines the initial segment determined by $\preceq_\sigma$ below x.

With these remarks out of the way we are ready to provide a formalization of Hamkins' proposal by letting the cautious enumeration be determined by the formula $\tau^o(x)$, which says

$$\tau(x) \wedge \forall d < x\, \neg Bew_{\tau|x}(^\ulcorner\neg Con(ZFC)^\urcorner, d).$$

Less formally, $\tau^{o\ulcorner}\varphi^\urcorner$ holds just in case φ is an axiom of ZFC and "we have not yet found a proof" in the sense of there being some $d < x$ coding a proof of $\neg Con(ZFC)$ from axioms among those "we have enumerated so far" in the sense that they are from $\tau|x$. I think this interpretation gives a reasonably faithful reading of what Hamkins means when he speaks of "what we have enumerated so far" and what it means to "have not yet found a proof." Morally speaking, if we ever considered adding to our stock an axiom from which we can prove the inconsistency of ZFC, then we have reason to stop the enumeration. On this basis, we might argue for its naturalness.

Our next goal is to generalize this rendering of caution and obtain a failure of well-foundedness. But before we do this, we highlight the key result for this example. We first recall the well-known fact that ZFC is essentially reflexive.

Fact 7. *(ZFC) If T is a theory extending ZFC in $\mathcal{L}_\in$ then for all finite subsets Δ of T, $ZFC \vdash Con(\Delta)$.*

This is a straightforward consequence of the reflection theorem and it allows us to establish the following.

Theorem 8. *(Hamkins, 2025) $Con(\tau^o) <_{Con} Con(ZFC)$, supposing $Con(ZFC + Con(ZFC))$.*

Although this proof occurs in (Hamkins, 2025), we are going to make a few simple generalizations of it below so it will be valuable to have a suitable version of the proof available for inspection.

[36] I'm going to continue to write $Con(ZFC)$ in deference to standard conventions. However in contrast to Hamkins, I will use the $Con(\sigma)$ notation for the more exotic enumerations introduced by Hamkins and discussed below.

Proof. Since $\tau^o \subseteq ZFC$, it is obvious that $Con(\tau^o) \leq_{Con} Con(ZFC)$, so it suffices to show there is a model $\mathcal{M}$ satisfying $Con(\tau^0)$ but $\neg Con(ZFC)$. By our assumption and Gödel's second theorem, we may fix a model $\mathcal{N}$ satisfying

$$ZFC + Con(ZFC) + \neg Con(ZFC + Con(ZFC)).$$

Then in $\mathcal{N}$, we have $ZFC \vdash \neg Con(ZFC)$. Thus, we see that τ^o is a finite subset of ZFC according to $\mathcal{N}$. Now using Fact 7 inside $\mathcal{N}$, we see that $\mathcal{N}$ also thinks that $ZFC \vdash Con(\tau^o)$. Then since $Con(ZFC) \rightarrow Con(ZFC + \neg Con(ZFC))$ holds in $\mathcal{N}$, we see that $\mathcal{N}$ also thinks $Con(ZFC + \neg Con(ZFC))$ and so we may fix a model $\mathcal{M}$ in $\mathcal{N}$ satisfying $ZFC + \neg Con(ZFC)$. But since $\mathcal{N}$ also thinks $ZFC \vdash Con(\tau^o)$, we see that $\mathcal{M}$ satisfies $Con(\tau^o)$ and $\neg Con(ZFC)$ as required. □

Thus and perhaps surprisingly, we see that the cautious enumeration ends up having a consistency strength strictly below ZFC as it is ordinarily enumerated. With this, the infinite descent begins. For the next step, we adopt an even more cautious position and check that – so far – not only have we failed to find a proof that ZFC is inconsistent, we have also not found a proof that $ZFC + Con(ZFC)$ is inconsistent. More formally, we let $\tau^{oo}(x)$ be

$$\tau(x) \wedge \forall d_0 < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC) \urcorner, d_0) \wedge \forall d_1 < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC + Con(ZFC)) \urcorner, d_1).$$

Call this the *doubly cautious enumeration*. Essentially the same proof then gives us the following.

Theorem. *(Hamkins, 2025)* $Con(\tau^{oo}) <_{Con} Con(\tau^o)$*, supposing*

$$Con(ZFC + Con(ZFC) + Con(ZFC + Con(ZFC))).$$

And from here it is not difficult to see that the we can generalize this to form triply cautious enumerations, quadruply cautious enumerations and so on.[37] Then the proof of Theorem 8 can be easily adapted to obtain the following.

Theorem. *(Hamkins, 2025) For all* $n \in \omega$, $Con(\tau^{o(n+1)}) <_{Con} Con(\tau^{o(n)})$*, supposing for all* $n \in \omega$

$$Con(ZFC^{o(n)}).$$

Taking some stock, we started out by just wanting to be a more careful in the face of the ever-present risk of inconsistency associated with ZFC and its extensions. This led us to adopt a more conservative approach to the enumeration of ZFC with the goal of mitigating at least some of that risk. On the basis of the sensibleness of this worry, we have argued that the resulting theories are natural. But as we see above, they also form an infinitely descending chain of consistency strengths and thus a challenge to the CHT.

Such is Hamkins' second proposal. The mathematics is undeniable, so as in the previous section, we will now attempt to push back on the more philosophical claim that these theories are natural. We shall consider two objections. Our first objection begins with what might be

[37] More formally, Let $ZFC^{o(0)}$ be $\tau(x)$; i.e., ZFC. Let $ZFC^{o(n+1)}$ be $ZFC^{o(n)} + Con(ZFC^{o(n)})$. Let $\tau^{o(0)}(x)$ be $\tau^o(x)$ and let $\tau^{o(n+1)}(x)$ be

$$\tau^{o(n)}(x) \wedge \forall d_n < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC^{o(n+1)}) \urcorner, d_n).$$

better construed as a criticism of my formal interpretation of the cautious enumeration. While Hamkins' is informal, I characterized this enumeration using the formula $\tau^o(x)$ which says

$$\tau(x) \wedge \forall d < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC) \urcorner, d).$$

While we arguably have more precision, it also invites a couple of pointed questions:

(i) Why should we just consider proof codes that precede x according to the standard ordering of the naturals in this particular coding?
(ii) Why should we consider initial segments of τ determined by the standard ordering of the naturals?

I think there are easy ways to modify τ^o that address these questions and the ensuing discussion will serve to illustrate how our thought experiment above should be understood in more detail. With regard to (1), it does seem a little odd to just use proof codes that precede x according to $<$. Why use x? Why use $<$? With regard to using x, we use it simply because that's the only variable available. But perhaps we want to check more than just those proofs below x. To do this, we could simply take any computable, order-preserving map $f : \omega \to \omega$ and use $f(x)$ rather than x as the bound. With regard to $<$, this is the same problem underlying (2). Why should our enumeration be hostage to the standard ordering of the naturals as mediated by some Gödel coding? This is also easy to address by fixing a computable permutation $g : \omega \to \omega$ and using the order $\prec_g$ instead, where for all $m, n \in \omega$

$$m \prec_g n \;\Leftrightarrow\; g(n) < g(m).$$

If we we make these modifications let $\tau^o(x)$ say

$$\tau(x) \wedge \forall d \prec_g f(x) \, \neg Bew_{\tau|_g x}(\ulcorner \neg Con(ZFC) \urcorner, d)$$

where $(\tau|_g x)(y)$ says $\tau(y) \wedge y \prec_g x$. This again results in a computable axiomatization of the theory in question. None of this affects the results above. Moreover, this arguably helps us show how the cautious enumeration fits our motivating story a little more tightly. We might think of $\prec_g$ as giving the enumeration of the axioms as we come to use them. So if we come to use some instance of Replacement after an instance of Separation, then this could be reflected in the $\prec_g$ ordering.[38] Similarly, we might think of f as telling us how many proofs we worked out when we added the new axiom.[39] Arguably this describes something like the process of set theorists working today. Thus, while I think it is fair to say that my original formalization of τ^o was somewhat crude, it is not difficult to patch things to be in better accord with our underlying motivations.

But I think this discussion prompts a reasonable question that leads to an objection due to John Steel. The underlying idea behind the cautious enumeration is a prudential attitude in response to the epistemic hurdles erected by Gödel's second theorem. So what happens if the cautious enumeration gets stuck? Suppose we add a new axiom and do a few proofs and establish on the basis of what we have so far that $\neg Con(ZFC)$. What should we do? This is not an outright proof of inconsistency, so it doesn't quite tell us that what we have is trivial. Nonetheless it should give us reason for pause. And this is exactly what the cautious

[38] This assumes that there is a computable permutation giving the order of discovery, but that seems like a relatively plausible assumption.
[39] Strictly, I think it would be more faithful to introduce a further computable permutation $h : \omega \to \omega$ to deliver another ordering $\prec_h$ that records the order of discover of proofs.

enumeration tells us to do. But what next? Since we discovered the moment when the rot set in, perhaps we should retreat to what we had just before that occurred. If we did this, then we'd still in accord with τ^o so this seems to line up with our motivating story. But, and here is where the objection comes in, this doesn't seem to line up with what set theorists would actually do in such a situation. While we can agree that the proof of $\neg Con(ZFC)$ is a good reason to pause, the sensible response would be to undertake a much deeper kind of retreat. Steel puts it as follows:

> If we discovered a proof of $\neg Con(ZFC)$, we wouldn't retreat to "the amount of ZFC enumerated so far." We'd say the ideas were wrong, and drop back based on some analysis of how the ideas went wrong.[40]

In other words, the occurrence of such an event would shake us back to the foundations. Such a result would force us to see that there was something deeply wrong in our understanding of set theory and this misunderstanding would rightly prompt questions about our entire axiomatization. Given this, τ^o and its cousins might seen less natural than at first blush.

For our second objection, we focus on the role of finitude in the results above and reflect on how this injects a kind of strangeness into the τ^o enumeration. First recall the crucial role of ZFC's essential reflexivity (Fact 7) in the proof of Theorem 8. It allowed us – upon recognizing that τ^o had a finite extension – to fix a model satisfying τ^o. We also observed above that essential reflexivity was a simple corollary of the Reflection Theorem in set theory. We shall demonstrate below that there is a sense in which the theory enumerated by τ^o doesn't satisfy the Reflection Theorem, which seems like particularly unnatural feature in a theory given the technical and philosophical importance of that theorem.

The key point is that assuming ZFC is consistent, we cannot prove from the axioms enumerated by τ^o whether τ_0 determines a finite axiomatization or not. As such, we aren't in a position to know whether reflection is available or not. More formally, we have:

Proposition 9. $\neg B_{\tau^o}\ulcorner\neg\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))\urcorner$, *if* $Con(ZFC)$.

Less formally, this says that we cannot prove from axioms satisfying τ^o that: there is no finite initial segment of τ^o indexed by some x that is the entirety of τ^o, assuming ZFC is consistent.

Proof. Since we have $Con(ZFC)$, we may use Gödel's second theorem to fix a model $\mathcal{M}$ satisfying $ZFC + \neg Con(ZFC)$. Now as in the proof of Theorem 8, $\mathcal{M}$ will think that there are only finitely many code numbers satisfying τ^o; i.e., $\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))$. Thus since $\mathcal{M}$ satisfies the axioms that satisfy τ^o, we see by soundness that we cannot prove $\neg\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))$ from axioms satisfying τ^o. □

This in itself is an unusual feature of the τ^o enumeration. I think this tells us that any reasonable enumeration of ZFC should not be like this. Indeed, it is easily proven using the Reflection Theorem that ZFC cannot be generated from any finite subset of itself, provided that ZFC is consistent; i.e., i.e., ZFC is not finitely axiomatizable. Indeed one can prove this

[40] Quoted with permission from an email between John Steel and me.

in a theory of arithmetic like PA.[41] Or more formally, recalling our very simple enumeration τ, described at the beginning of Section 2.2.2., we have:

Proposition 10. *Suppose that ZFC is consistent. Then*

$$B_\tau{}^\ulcorner\neg\exists x\forall y((\tau|x)(y) \leftrightarrow \tau(y))^\urcorner.$$

Thus, if we use an enumeration like τ, then even a very weak theory can prove that τ determines an infinite theory. Moreover, it is easy to see that this result perseveres to stronger theories using the obvious modifications. The essential ingredient of this proof is, of course, the Reflection Theorem. As such, we now have some good reason to doubt that τ^o determines a theory with the Reflection theorem. We close this section by spelling this out in some detail.

We start with a way of defining theories that satisfy the Reflection Theorem. It is traditional to state the Reflection Theorem as a schema and then prove it as a kind of meta-theorem. But as our concerns here are particularly metamathematical, we add the fussy detail of a metatheory (PA) from which the theorem can proved.

Definition 11. Let us say that a formula $\sigma(x)$ of $\mathcal{L}_\in$ *enumerates a reflective theory* if PA proves that whenever $\sigma^\ulcorner\varphi_0{}^\urcorner \wedge \ldots \wedge \sigma^\ulcorner\varphi_n{}^\urcorner$ then[42]

$$B_\sigma{}^\ulcorner\forall\alpha\exists\beta > \alpha \bigwedge_{i<n} \varphi_i^{V_\beta}{}^\urcorner.$$

Thus, we have a reflective theory when PA can prove that any finite set of sentences satisfying $\sigma(x)$ are themselves satisfied in a rank initial segment of the universe. This is, of course, a well-known and weak version of the Reflection Theorem. A standard argument then shows that such theories cannot be finitely axiomatized.

Lemma 12. *If σ binumerates a consistent reflective theory, then PA proves that there is no x such that*[43]

$$\forall y(\sigma(y) \leftrightarrow (\sigma|x)(y)).$$

We can then put all this together to see that τ^o doesn't satisfy the Reflection Theorem.

Theorem 13. *τ^o does not enumerate a reflective theory, if ZFC is consistent.*

Proof. Work in PA and suppose toward a contradiction that τ^o does enumerate a reflective theory. Then Lemma 12 and Proposition 9 with $Con(ZFC)$ respectively imply that there is and there is not some x such that $\forall y(\sigma(y) \leftrightarrow (\sigma|x)(y))$, which is impossible. □

Summing up, we see that while Gödel's second incompleteness theorems reasonably prompt concerns about the consistency of strong theories like those extending ZFC, the effort to ameliorate these worries by modifying the way we enumerate axioms leads into a couple of arguments against the naturalness of such theories. First, we argued via Steel that the cautious

[41] In fact primitive recursive arithmetic would suffice. See Corollary IV.7.7 in (Kunen, 2006) for a proof and further discussion of the metamathematical setting. For the purposes of uniformity, we shall let PA be the set theory formed by: removing the axiom of infinity; adding its negation; and then replacing Foundation with Set Induction. The resultant theory is well-known to be bi-interpretable with the standard version of PA so no harm has been done (Kaye and Wong, 2007).

[42] This is essentially says that $\sigma(x)$ determines a theory satisfying a fussy restatement of Corollary IV.7.6 of (Kunen, 2006).

[43] A proof of this essentially follows that of Corollary IV.7.7 in (Kunen, 2006).

enumeration doesn't fit mathematical practice as well as it might seem as first. Second, we see that properties we have come to take for granted with regard to ZFC are no longer guaranteed in the context of unorthodox enumerations. Arguably the real lesson here is that we should be more careful in stating what we really want from an enumeration of a theory like ZFC. Computability is certainly a necessary condition, but each of the enumerations above are computable axiomatizations of ZFC and yet they behave very oddly.

Before we move to our final example, I want to stress again that I doubt that I've definitively argued that the enumerations discussed in this section are unnatural. As in the previous section, I wouldn't be surprised if one could push back. However, with every passing epicycle in such a debate, I think the case for unnaturalness looks stronger for the simple reason that a natural theory should be obviously so. This again is a testament to my dialectical position and should cast no shade on Hamkins' innovative examples.

2.2.3. Sticking to the good models

This final example is a variation of the one from Section 2.2.1., but I think it's important to discuss since it scratches at an itch that many set theorists will be feeling right now. While relative consistency proofs like Gödel's first incompleteness theorem are of great interest, relative consistency proofs in set theory tend to demonstrate a much tighter relationship between the theories in question. To see this, let's consider a well-known forcing example. Suppose we want to show that $ZFC + \neg CH \leq_{Con} ZFC$. To do this we might start with a countable transitive model M satisfying ZFC and then show that there is an M-generic set G such that the extension $M[G]$ of M safisfies ZFC and $\neg CH$. Thus, we have shown that if there is a transitive model of ZFC, then there is a transitive model of $ZFC + \neg CH$.[44] Let's introduce a little notation for this by writing $Con_\beta(T)$ to mean that there is a transitive model of T.[45] And let us write $T \models_\beta \varphi$ to mean that every transitive model of T also satisfies φ. Then we have shown that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + \neg CH).$$

Analogously, we may also show that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + CH)$$

using Gödel's inner model argument.[46] Now since we have restricted our attention to countable transitive models, we have not quite demonstrated that $ZFC + \neg CH \leq_{Con} ZFC$. There are a number of standard methods of patching up the proof to achieve this,[47] but this is the conceptual core of the argument. Moreover, this way of looking at things reveals a much closer connection between the theories in question. For example, the models of each theory have the same ordinal spine. By contrast, we cannot prove that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + \neg Con(ZFC))$$

even though $ZFC + \neg Con(ZFC) \leq_{Con} ZFC$. This is because if there is a countable transitive model of ZFC, then no model of $ZFC + \neg Con(ZFC)$ can be transitive since transitive models have the true set of natural numbers and so they must see that $Con(ZFC)$ is true. So we start

[44] Note that if there's a transitive model, then there is a countable transitive model too.
[45] The idea here is that T has a β-model, where a β-model of $\mathcal{L}_\in$ is well-founded. Then since we are working in ZFC as a background theory, we may assume that such a model is transitive without loss of generality.
[46] Of course, there are no issues regarding countable models in this case.
[47] See Section VII.9 of (Kunen, 2006) for an excellent overview.

with a good, well-founded model of ZFC and then obtain an ill-founded model whose ordinals have no meaningful relationship with the ordinals of the model we started with.

Thus, forcing and inner models seem to tell us much more about how the theories they discuss are related than a mere relative consistency argument. Moreover, as we saw in Section 2.2.1., these considerations give us reason to worry about the naturalness of theories extending ZFC that are not within the reach of forcing or inner model theory. At the least, we would hope that the existence of a transitive model of one natural theory should imply the existence of a transitive model of another. With this in mind, let us introduce a hopefully better strength relation between theories.

Definition 14. For theories T and S extending ZFC in the language of set theory, let us say that T is *transitively consistent relative to* S, abbreviated $T \leq_{trans} S$ if

$$ZFC \vdash Con_{\beta}(S) \to Con_{\beta}(T).$$

Given that we have just seen that $ZFC + \neg Con(ZFC) \not\leq_{trans} ZFC$, we have some reason to hope that the kinds of counterexamples that we have seen above will be eradicated and that CHT will be vindicated in this context. But intriguingly, Hamkins shows that this is not the case. Using essentially the same proof strategy as he employed for Theorem 6, he is able to show the following.[48]

Theorem 15. *(Hamkins, 2025) There is a partial computable function $\psi : \omega \rightharpoonup \omega$ such that the theories*

$$ZFC + \text{“}\textit{there are } \psi(n) \textit{ inaccessible cardinals.}\text{”}$$

for $n \in \omega$ are pairwise $\leq_{trans}$ incomparable, assuming there are infinitely many inaccessible cardinals.

Thus, even if we restrict our attention to transitive models, we have the mathematical bones of a counterexample to CHT. Of course, many of the worries raised in Section 2.2.1. are directly transferable to the current context: we are axiomatizing a theory using a partial computable function, which makes figuring out the axioms very difficult; and we seem overly concerned with the behavior of a partial computable function in nonstandard contexts where infinite natural numbers reside. As such, I still think we have good reason to be reserved about calling such a theory natural, but the generalization to transitive models is certainly remarkable and is a significant milestone for the discussion that is to come.

3. Explaining the difficulty

Thus far, we have explored the mathematical and philosophical significance of the Consistency Hierarchy Thesis and considered a number of putative counterexamples with regard to which we have expressed some reservations. In particular, we have raised doubts about the naturalness of the theories offered in these counterexamples. In this final section, I want to consider things from a different perspective. Rather than thinking about the actual examples that support the thesis and how some specific examples fail to defeat it, I want to consider the endgame of the problem itself. As such, I have two main goals. First, I want to think more seriously about what an uncontroversially successful counterexample to CHT would look like

[48] In fact, one might even say that the proof is a little easier.

and the likelihood of such a scenario taking place. Second, I want to pull together a few of the recurring threads of this paper and offer a kind of formal explanation as to why refuting CHT seems so improbable. My hope is to demonstrate that CHT is a somewhat idiosyncratic problem and that its peculiarities should be taken into account when we assess its significance.

3.1. The ideal counterexample scenario

In Section 2, we considered putative counterexamples to CHT that aimed to show that the consistency hierarchy for natural theories is not well-ordered by offering examples where comparability and well-foundedness fail, but we also saw some reasons to doubt that the examples really delivered natural theories. In this section, rather than attempt to provide a specific example, I want to take a step back and paint a picture of what I think the ideal counterexample to CHT would look like. My goal is to consider a scenario that would have maximal impact on the set theory community and also be accepted by them. I think there is value in exploring this idea as it makes clearer just how thorny this problem is.

We begin our story with young algebraic topologist, Gurt Ködel. Ködel has been working on a collection of seemingly natural problems in his field but has been unable to find any way to solve them. These problems seem like sensible generalizations of problems already considered by his mathematical community and yet they seem to lie beyond the reach of existing techniques. Naturally enough, he wonders whether these problems are perhaps independent of ZFC, but again, he is unable to get very far with this question. Nonetheless, Ködel does develop a new axiom (or perhaps collection thereof) which seems to add exactly the right ingredient to address the problems he has raised. Moreover, he discovers that this new axiom is able to address other problems, some of which were already in the literature and some of which are quite new. Ködel shares these results with his colleagues and goes on to gain a degree of fame by having opened the door to a hitherto unexplored region of mathematics. Other mathematicians go on to solve further problems with this axiom and this new field continues to flourish. One might imagine this taking place over a period of years or even decades.

This describes a prototypical scenario for the emergence of a natural theory. A new axiom is introduced for a clear mathematical purpose and even better, the development of the theory around this axiom garners wide interest in the mathematical community. Note that in contrast to the examples above, the story here is organic rather than contrived. Life is often easier in thought experiments.

So far so good. As the notoriety of this new field grows the set theory community gets wind of its success and ask the question: where does this new theory fit in the large cardinal hierarchy? What is its consistency strength? It is well-known that Ködel and his successors were unable to solve this problem and so it remained open. This raises further difficult philosophical questions. Perhaps the most important of these is: why is this a problem and who is it a problem for? On one hand, the set theory community has a pretty standard line on these matters: if you cannot prove that a theory is consistent relative to a large cardinal axiom, then we have no reason to think that theory is consistent. Perhaps the classic case of this is Quine's theory NF. While the jury might be out as to whether this really is a natural theory, there is not doubt that it currently lacks a consistency proof that has been agreed to be correct by the mathematical community.[49] As such, it is customary to regard the question of NF as open. However, if one were able

[49] Interestingly, during the time this paper was written NF now has a computer-verified consistency proof (Holmes and Wilshaw, 2024). This does not appear to be very well-known in the set-theory community yet, however, it seems reasonable to expect that there will now be consensus that the consistency

to show that NF was consistent relative to a large cardinal (or more likely much less) then we would regard the question of NF's consistency as having been resolved. It is important, however, to consider the significance of CHT in the set theorist's standard line. The success of CHT with regard to what we currently know is what gives us reason to think that consistency relative to a large cardinal axiom is – at least – a necessary condition for thinking that a theory is consistent. If CHT were to face serious doubts, then a failure to determine consistency relative to a large cardinal axiom would carry much less weight. But this is precisely the kind of pressure the natural theory imagined above would place on CHT. Our thought experiment is exploring a scenario where a theory governing an active area of mathematics cannot be understood as fitting into the large cardinal hierarchy.

Continuing our story, the consistency of Ködel's theory becomes one of the dominant open questions of set theory. Partial questions are developed and answered using forcing and inner model theory, but the main question remains unanswered. Much imaginative but failed work remains in notebooks and on blackboards never seeing the light of publication. The question becomes an albatross for CHT as work on Ködel's theory continues to flourish. And as is not atypical in mathematics, the question remains open for decades.

At this juncture in our imagined scenario, CHT looks much weaker than it does in our world. It hasn't been refuted, but the idea that the large cardinal hierarchy sets the milestones on the long road of theoretical strength is certainly tarnished. I think it's also worth noting that the kind of pressure placed on CHT in this example, is quite different to the examples considered in Section 2. In those cases, we considered examples of theories that refuted well-orderedness by breaking comparability and well-foundedness. Here, we are seeing a problem with a facet of well-orderedness that is arguably more significant for the philosophical and foundational consequences of CHT: that the consistency hierarchy might not be directed. Or more specifically, Ködel's theory challenges the idea that for any natural theory T there is a large cardinal axiom LC_T such that $T \leq_{Con} LC_T$. In other words, this would challenge the idea that the large cardinal axioms are cofinal in the hierarchy of consistency strengths. While breaking comparability or well-foundedness with natural theories, would certainly refute CHT as we have defined it above, one might easily design a fallback version of CHT that gives up on these features. While this would be disappointing, much of the philosophical and foundational significance of CHT described in Section 1 can still be obtained without them. Directedness, on the other hand, is a deal breaker. My contention is that if such a scenario were to arise, then our reasons for believing CHT, or anything like it, would be terminally weakened. Moreover, given the impeccable credentials for the naturalness of Ködel's theory, I think the set theory community would be forced to agree.

But there is another feature of this thought experiment that contrasts with the examples of Section 2 and that provides a little room for pushback. In Section 2, we offered theorems (via Hamkins) that delivered counterexamples to comparability and well-foundedness. In the thought experiment above, we have merely proposed a failure of discovery rather than a proof that there is no large cardinal axiom from which the consistency of Ködel's theory can be established. Given that we are in already in realm of mere fantasy, let us complete our ideal picture and consider what an extra move might look like. Let us suppose that some years after the advent and popularization of Ködel's theory a young logician, Caul Pohen studies the

of NF has been proved. I've opted to leave this little passage as it is, as it provides a pleasing illustration of exactly the kind of event that our fictional account describes.

theory and comes up with a remarkable theorem. We suppose that Pohen shows that there is no large cardinal axiom that we can add to ZFC such that we can start with a model satisfying that large cardinal and then either force or take an inner model to obtain a model of Ködel's theory. Given we have no formal definition of a large cardinal, it is difficult to see how this kind of thing could even be stated as a theorem. However, it is salient to note that there a no currently known examples that fit this template. Every instance of a natural theory that we currently know, even the proper forcing axiom, is consistent relative to a large cardinal axiom. I think it is clear that if Pohen were able to state and prove such a theorem, then the philosophical and foundational significance of the Consistency Hierarchy Thesis would be in tatters.

This is my thought experiment. Of course, it hasn't occurred and we have no good reason to think it will. But I also think the reasons we have to think it won't occur are very weak by conventional mathematical standards. What we do see is that while this ideal situation would be decisive against CHT, it also requires a lot of cards to land a certain way in order for it to take place. In particular, it seems very unlikely that an individual researcher could achieve it since the story has so many moving parts. First, we need a mathematical problem that seems beyond ZFC. Then we need a proposal for solving it that generates interest in the mathematical community. Finally, we need that problem to, at minimum, remain recalcitrant to analysis in the consistency strength hierarchy. Or better, that someone can just show that this theory is not consistent relative to large cardinals. It's a tall order.

3.2. Are we chasing a theorem?

We've now painted a relatively vivid picture emphasizing the challenges involved in attempting to refute the Consistency Hierarchy Thesis. In this section, I would like to offer another explanation as to why a refutation is unlikely to occur. This time, we shall focus shift our focus to the actual rather than counterfactual practice of set theorists and argue that what set theorists seem to expect from relative consistency proofs puts them in – or at least very close to – a position from which delivery of a counterexample is impossible.

3.2.1. Set theorists want more than relative consistency

We start by considering a classic pair of theories where one theory is strictly below the other in consistency strength. We shall then use this example to deliver a generalized version of consistency strength which is provably linear. While the following is standard – if not trivial, it will be hopefully shed a little light and warm up our intuitions.

Theorem 16. *$ZFC + \exists\kappa$ κ is strong $<_{Con}$ $ZFC + \exists\delta$ δ is Woodin, assuming that $ZFC + \exists\kappa$ κ is strong is consistent.*

Proof. Either ZFC with a Woodin cardinal is consistent or it is not. In the latter case, the theorem follows immediately. So we consider the case where it is consistent and fix a model $\mathcal{M}$ witnessing this. We then make the *crucial observation* that if δ is a Woodin cardinal, then there is some $\kappa < \delta$ such that[50]

$$V_\delta \models \kappa \text{ is strong.}$$

[50] A proof of this is essentially delivered in the proof of Proposition 26.13 of (Kanamori, 2003), however, it also follows trivially from the more modern definition of a Woodin cardinals as can be found in Theorem 26.14 of (Kanamori, 2003).

This establishes that ZFC with a Woodin cardinal proves that there is model of ZFC with a strong cardinal, so $V_\delta^{\mathcal{M}}$ thinks there is a strong cardinal, where δ is a Woodin cardinal of $\mathcal{M}$. Thus, we have

$$ZFC + \exists\kappa\ \kappa \text{ is strong} \leq_{Con} ZFC + \exists\delta\ \delta \text{ is Woodin.}$$

To get the strict, $<_{Con}$, we note that if we could prove the consistency of ZFC with a Woodin from the consistency of ZFC with a strong cardinal, then ZFC with a Woodin could prove that it was consistent, contradicting Gödel's second theorem. □

The engine of the proof above lies in its crucial observation; the rest is just metamathematical bookkeeping. But there is also a sense in which the theorem as stated dilutes what the proof actually delivers. If we use the completeness theorem to move from a syntactic setting to a semantic one, then theorem above just tells us that there is a model $\mathcal{M}$ of ZFC that satisfies:

$$\exists\mathcal{N}\ \mathcal{N} \models ZFC + \exists\kappa\ \kappa \text{ is strong} \ \wedge\ \neg\exists\mathcal{P}\ \mathcal{P} \models ZFC + \exists\delta\ \delta \text{ is Woodin.}$$

But this doesn't tell us that the models involved are any good. Are they well-founded? Are they correct? For all we may discern from the statement of the theorem, such $\mathcal{N}$ and $\mathcal{P}$ might disagree on the natural numbers. But the proof tells us more. Before we state this, let us also employ a common set-theoretic idiom and leave the large cardinal assumptions required implicit.[51] Then we have the following.

Theorem 17. *There is a transitive model M of ZFC satisfying that*

$$Con_\beta(ZFC + \exists\kappa\ \kappa \textit{ is strong}) \ \wedge\ \neg Con_\beta(ZFC + \exists\delta\ \delta \textit{ is Woodin}).$$

Proof. Let M be the shortest transitive model with a Woodin cardinal, δ. Then by the crucial assumption from the proof of Theorem 16, we see that V_δ^M thinks there is a strong cardinal. Thus, M satisfies $Con_\beta(ZFC + \exists\kappa\ \kappa \text{ is strong})$. But M cannot satisfy $Con_\beta(ZFC + \exists\delta\ \delta \text{ is Woodin})$ as M was the shortest of its kind. □

Stripped of syntactic detours, this seems to be a more informative theorem. It has an immediate connection to the crucial observation in that it is now essentially a corollary of it. The models employed here all meet a minimum standard of quality in that they are all transitive. I contend that the argument used above is more "set theoretic" in spirit and that it reveals a deeper connection between the theories being compared.

3.2.2. An unattainable target

What happens if we run a little further with the idea of the previous section? So rather than dealing with the syntactic thorniness of mere consistency with all of the counterintuitive features we've seen above, why not work with a collection of models that we have reason to think are good? Any natural set theory should have a well-founded model so why not restrict our attention to the case of well-founded models and their canonical representatives: transitive sets. With that in mind, we might consider a new consistency relation on set theories that better reflects the techniques and attitudes of working set theorists.

[51] In the case below, assuming the existence of a transitive model of ZFC plus a strong cardinal would suffice.

Definition 18. For theories T and S extending ZFC in the language of set theory, let us say that T is β-*consistent relative* to S, abbreviated $T \leq_\beta S$ if

$$ZFC \models_\beta Con_\beta(S) \rightarrow Con_\beta(T).$$

A quick flick back will reveal this is exactly the same as our definition of relative consistency[52] except that we have replaced every use of a consistency or consequence with β-consistency and β-consequence respectively. The underlying idea is that the use of transitive models provides – among other things – a better reflection of the preferences of set theorists. Moreover as we have discussed extensively above, they have good reason for this preference. Ill-founded models do not fit with our intended interpretations of set theory; and worse they often make errors in calculation.[53] Thus, we are investigating what happens when we rid ourselves of this hassle. With this definition in hand, we may then repackage Theorem 17, to obtain

$$ZFC + \exists\kappa\ \kappa \text{ is strong} <_\beta ZFC + \exists\delta\ \delta \text{ is Woodin}.$$

So we might argue that we now have a way of comparing theories that cuts away pathology and gets, at least, a little closer to the information that relative consistency proofs in set theory are able to reveal. I think it is clear that there is often more information still left out of reach in this analysis, but nonetheless, we have a certain kind of improvement on ordinary relative consistency. The obvious question then is: what happens to the Consistency Hierarchy Thesis if we articulate it in the context of β-consistency? The answer may be surprising: it turns it into a theorem!

Theorem 19. *Let T, S be computably axiomatizable theories extending ZFC. Then is must be the case that either: $S \leq_\beta T$ or $T \leq_\beta S$. Moreover, $\leq_\beta$ is a pre-well-ordering on such theories.*

Proof. We just establish the first claim here and leave the latter to an appendix. Nonetheless, the main trick will have already been revealed. Suppose toward a contradiction that T and S provide a counterexample. Then we may fix countable transitive models M_0 and M_1 of ZFC such that:

(i) $M_0 \models Con_\beta(T) \wedge \neg Con_\beta(S)$; and
(ii) $M_1 \models Con_\beta(S) \wedge \neg Con_\beta(T)$.

Using the Lévy-Shoenfield theorem, one can see that, without loss of generality, we may assume that M_0 and M_1 both satisfy $V = L$. And so by the condensation lemma, we may fix $\alpha_0, \alpha_1 < \omega_1$ such that $M_0 = L_{\alpha_0}$ and $M_1 = L_{\alpha_1}$.

Clearly, $\alpha_0 \leq \alpha_1$ or $\alpha_1 \leq \alpha_0$, so suppose the former. Then using (1), we see that there is some transitive $N \in L_{\alpha_0}$ that satisfies T. But then N is also an element of L_{α_1} which contradicts (2). Thus, $\alpha_1 > \alpha_0$. A similar argument then shows that $\alpha_0 > \alpha_1$, which is impossible. □

This simple theorem thus establishes that for any pair of computably axiomatizable theories, their β-consistency strengths are comparable, regardless of whether or not those theories are natural. Given that a natural theory must surely have a transitive model, we seem to have something very like CHT. I think this is quite striking and worthy of note. Of course, the proof is almost absurdly simple; the interest is rather in the set up. So what should we make of this?

[52] See Definition 2. Also see Section 2.2.3. for definitions regarding β-logic.
[53] For example, ill-founded models can think that ZFC is inconsistent.

First of all, $\leq_\beta$ is not $\leq_{Con}$, so we can rightly be accused of moving the goal posts.[54] But does that give us license to simply ignore the result above? I think that would also be too easy. Restricting our attention to transitive models is very much in line with ordinary set theoretic practice. But let us think a little more slowly about this. Recalling that $T \leq_\beta S$ if

$$ZFC \models_\beta Con_\beta(S) \to Con_\beta(T),$$

we see that there are two places where β-logic has entered the story. We are using it *internally* with the Con_β statements and also *externally* we demand that the relationship between the Con_β statements be itself a β-consequence of ZFC. We gave an argument for the internal uses of β-logic in Section 2.2.3., where we noted that in their actual practice theorists prove relative consistency statements using forcing and inner model theory. These arguments are such that if we start with a transitive model, then we also end up with one. Moreover, we have seen many examples above where the pathological behavior of ill-founded models causes counterintuitive and unnatural results. It almost seems obvious that an ill-founded model of set theory cannot be an intended model.

Some of this reasoning also carries over to the external uses of β-logic, but I think the external use is a little more exposed to reasonable doubt. Without it, we are somewhere near the situation of Section 2.2.3., where Hamkins has been able to propose a counterexample albeit without computable axiomatization.[55] With it, the CHT problem is dead. It's just a theorem. So a lot hangs on this move. As has so often been the case in this paper, I don't have anything definitive to say. Worse, I have now moved in the dialectical outsider position, so I don't really expect every reader to come on board. Nonetheless, I'd like to offer a couple of nudges to at least give the reader some pause.

First of all, one might push back against the external use of β-logic by noting that the set of β-logic consequence of ZFC strictly extends ZFC. For example, it is easy to see that $Con(ZFC)$ is β-consequence of ZFC.[56] So perhaps this theory is too strong. In response to this kind of worry, we note that it is easy to see that the β-consequences of ZFC are derivable from a modest, albeit global, large cardinal assumptions.

Proposition 20. *If there is a proper class of inaccessible cardinals, then for all sentences φ of set theory*

$$(ZFC \models_\beta \varphi) \Rightarrow \varphi.$$

Proof. Suppose ψ is true. It suffices to show that there is a transitive model of ZFC in which ψ holds. To see this we observe that since there are unboundedly many inaccessible cardinals, there is a club class of ordinals α such that V_α satisfies ZFC. Thus, using reflection[57] we may obtain such an α where $V_\alpha \models \psi$. □

Informally speaking, this tells us that in the context of a proper class of inaccessible cardinals, any β-consequence of ZFC is already true. Thus, given that ZFC β-consequences hold in

[54] For example, if we let T be ZFC and S be $ZFC + \neg Con(ZFC)$, we have $T \equiv_{Con} S$ and we also have $T <_\beta S$, assuming, say, an inaccessible cardinal.

[55] I believe the question of whether there is non-linearity in the β-consistency hierarchy when we use first order logic externally remains open. The work that seems most closely related to this question occurs in (Aguilera and Pakhomov, 2024). While this paper uses first order logic externally, it also makes use of theories that are not computably axiomatizable. Nonetheless, the proof theoretic techniques utilized there provide a helpful showcase for techniques beyond forcing and inner model theory.

[56] To see this, suppose that M is a transitive model of ZFC. Then clearly $Con(ZFC)$ is true and since every transitive model of ZFC has the correct version of ω, M must think that $Con(ZFC)$ holds.

[57] For a specific formulation, use Theorem IV.7.5 of (Kunen, 2006).

such a modest extension of ZFC, it seems reasonable to take them seriously. Moreover, the conclusion of the proposition is clearly a (very) generalized version of the kind of reflection principle considered by (Feferman, 1991). If we take ZFC seriously, then it seems like the kind of thing we should expect to be true. This doesn't definitively resolve the matter but will hopefully assuage some anxiety. It should also be noted that although Theorem 19 tells us that every pair of natural theories are have comparable β-consistency, it tells us nothing about the direction in which this is resolved or how we might go about figuring that out. As such, it reflects a minor incursion into second order logic where this problem manifests more starkly. In particular, second order logic might be said to decide the continuum hypothesis without giving us any indication of the direction in which it is resolved. Nonetheless, while Theorem 19 tells us nothing about the specific relations between particular theories, it does tell us that, in this rarefied context, the Consistency Hierarchy Thesis has been confirmed.

For a second nudge, it is also worth bearing in mind that what we have called relative β-consistency has periodically appeared in a different guises as a kind of folk ordering on theories. For example, in a footnote of (2014), Steel notes

> There are various ways to attach ordinals to set theories that correspond to the consistency strength order in the case of natural theories. One can look at the provably recursive ordinals, or the minimal ordinal height of a transitive model, for example.

Drawing the last of these out, we might let ρ be a map from the computably axiomatizable theories to the ordinals, such that: $\rho(T) = \alpha$ if α is least such that there is a transitive model M satisfying T where $\alpha = Ord^M$. From this, we can then obtain a relatively simple pre-well-ordering, $\preceq_{rank}$ on computably axiomatizable theories. I think it's fair to say that this definition has the feel of folklore and that it clearly puts everything we want in a sensible order. If you wanted to put theories into a pre-well-order, it's probably the most obvious way to do it. However, I think the philosophical and foundational significance of this ordering is less obvious. While it certainly works in the mathematical sense of providing an ordering, it's not so clear how this relates to CHT. It turns out that this obvious pre-well-ordering is generally equivalent to the β-consistency ordering.

Proposition 21. *Suppose T and S are computably axiomatizable theories in $\mathcal{L}_\in$ that extend ZFC, then*[58]

$$T \preceq_{rank} S \Leftrightarrow T \leq_\beta S.$$

Again, this is hardly definitive evidence that relative β-consistency is the right way to go. But the fact that there are at least a couple of ways one might bump into this ordering suggests that something significant is occurring here.

So much for my nudges. Where does this leave us? Since we've changed the subject, we certainly haven't confirmed CHT. Nonetheless, I do think we can take a couple of lessons away from this experience. The first of these is that Theorem 19 highlights the fact that ordinary set theoretic methods cannot be used to obtain a counterexample to CHT. A successful refutation of CHT would require, for example, a pair of theories with regard to which we *cannot prove* that

[58] We save the proof for the appendix.

one theory is consistent relative to the other. As such, it requires a proof that there is no proof; i.e., a consistency proof is required. But in contrast to the techniques of ordinary set theory, such consistency proof cannot be obtained by forcing or inner model theory. As we have seen, these proofs take us from transitive models to transitive models and are thus, hostage to Theorem 19 where no counterexamples are available. Of course, the fact that techniques from outside the usual ken of set theory – like computability theory and proof theory – may be required to address this problem has no impact on CHT as a mathematical problem. But – and this brings us to the second lesson – it may have some affect on how we evaluate the foundational significance of the problem. If the only way to obtain a counterexample is to work within the pathological realms of nonstandard models, then perhaps we don't have to take a counterexample to CHT very seriously at all. Indeed and more provocatively, perhaps a pair of theories providing a counterexample to CHT is – in itself – evidence that at least one of those theories isn't natural.

3.3. Making sense of it all

Let's close this paper by trying to pull its many threads together. We started by defining what it means for one theory to be consistent relative to another and from there we were able to describe what we called the Consistency Hierarchy Thesis. We were able to get an understanding of what makes consistency strength interesting from a mathematical perspective. We then noted that CHT provides a tantalizing answer to the question of what lies beyond ZFC: a surprising amount of order. Moreover, we noted the fundamental role that CHT plays in our arguments for the existence of large cardinals and our understanding of set theory's influence on ordinary mathematics. Nonetheless, we also observed that CHT is not a genuine mathematical problem. In particular, the notion of "natural theory" has no mathematically precise formulation and yet it plays a crucial role in the articulation of CHT. As such, we started to wonder about the prospects for refuting CHT. With this in mind, we then explored a number of putative counterexamples proposed by Joel David Hamkins. As we suspected, in each of these cases, the claim for the naturalness of the theories involved was exposed and vulnerable to pushback. So despite the impressive ingenuity inherent in each of these examples, we were not optimistic that the broader set theory community would embrace them as successful. We did not, however, find this to be a cause for simple celebration on behalf of CHT. Rather, we worried about the dialectical advantage possessed by those aiming to defend CHT against putative counterexamples. The bar for naturalness among theories is set so very high. This then prompted a different tack. Instead of considering real examples, we asked what an ideal counterexample to CHT might look like. We ended up with a fictional account that – were it to occur – would likely succeed in refuting CHT and gain acceptance from the mathematical community that it had. But the story had a long wish list. A number of clever people had to do a number of brilliant things over a probable period of decades. Nothing like it has occurred and nothing like it seems to be in the process of occurring right now. This prompted our final shift in perspective. Rather than taking CHT literally, we considered the kinds of argument that set theorists use in establishing relative consistency claims. In particular, forcing and inner model theory always take us from one transitive model to another. When we reformulated CHT by immersing it in the context of transitive models, our problem simply disappeared since CHT became a theorem.

We seem to have painted a picture in which CHT appears almost immune to refutation. At the least, we've seen that a successful counterexample will be quite strange in that the standard

techniques of forcing and inner model theory cannot deliver it. Perhaps the strangeness of the tools used to establish such a result would – while delivering a minor mathematical miracle – only offer us philosophical crumbs in return. Part of appeal of CHT lies in the fact that the interpretations yielding its instances preserve a great deal of meaningful structure. If this were absent, then the foundational significance of such an example would be difficult to evaluate. For reasons like these, we may end up in a situation where any proposed counterexample to CHT is rejected since theories that are not connected by forcing or inner model theory are – almost by fiat – unnatural. While we are not currently in such a situation, this could seem unappealing or even question begging. But I think it actually gets us closer to the right way to understand CHT. As we suggested above, it bears an interesting relationship with the Church-Turing thesis, although the CHT story is far less complete. However, if natural theories continue to be regimented by the large cardinal hierarchy using forcing and inner model theory, then we may well come to see things from the other side of the table: we might see natural theories as being – in an informal sense – exactly those that can be accessed from large cardinal axioms by forcing and inner models. There would be a long way to go along that road before we get to such a position, but it is important to remember that – if this road were to be traveled – we would not only have a longer list of elegant equiconsistency proofs, the tools developed in producing those proofs would also give us insight into what is making those theories seem natural. The point I'm trying to make is that this would not be a victory via mere enumeration, greater understanding will also need to come along for the ride. In this spirit, I'd like to end this paper with a shoutout to two projects that show great promise in developing such understanding and which go far beyond the relatively superficial analysis of this paper. The longest running program in this area takes place within inner model theory. John Steel describes three closely related hierarchies: the consistency strength hierarchy of this paper; the Wadge hierarchy on homogeneously Suslin sets of reals; and the mouse order on canonical inner models for large cardinals. For Steel, the final ordering on mice provides the foundation for the other two (Steel, 2024).[59] Thus, the consistency strength hierarchy is explained through a deeper hierarchy of models that instantiate natural theories. More recently, James Walsh has taken a different angle on this problem of understanding by asking very directly: what makes a theory natural? This work draws on techniques from ordinal analysis in proof theory and analogies with Martin's Conjecture in computability theory (Montalbán and Walsh, 2019; Walsh, 2025). While a lot of work remains to be done and many deep questions remain open, the progress toward understanding is plain to see.

Acknowledgements. I'd like to thank John Steel for some helpful discussions about this material. I'd also like to thank Pen Maddy for some helpful comments on an earlier version of this paper. Thanks also to the Irvine LPS Logic Seminar for letting me present this material. And thanks to the anonymous referees who helped me improve this paper.

References

Aguilera, J. P. and Pakhomov, F. (2024). Non-linearities in the analytical hierarchy. Preprint.

Black, R. (2000). Proving church's thesis. *Philosophia Mathematica*, 8(3):244–258. DOI: https://doi.org/10.1093/philmat/8.3.244.

[59] For an overview on the state of the art on this topic see the introduction to (Steel, 2022).

Boolos, G. (1984). The logic of provability. *The American Mathematical Monthly*, 91:470–480. DOI: https://doi.org/10.1080/00029890.1984.11971467.
Cohen, P. (1963). The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences of the USA*, 50(6):1143–1148.
Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, 56(1):1–49. DOI: https://doi.org/10.2307/2274902.
Feferman, S., Friedman, H. M., Maddy, P., and Steel, J. R. (2000). Does mathematics need new axioms? *The Bulletin of Symbolic Logic*, 6(4):401–446. DOI: https://doi.org/10.2307/420965.
Gödel, K. (1940). *Consistency of the Continuum Hypothesis*. Princeton University Press.
Gödel, K. (1986). On formally undecidable propositions of principia mathematica and related systems. In Feferman, S., editor, *Kurt Gödel Collected Works Volume I: Publications 1929-1936*, volume 1. Oxford University Press, New York. DOI: https://doi.org/10.1093/oso/9780195147209.003.0014.
Goldberg, G. (2022). *The Ultrapower Axiom*. De Gruyter, Berlin, Boston. DOI: https://doi.org/10.1515/9783110719734.
Grotenhuis, L. (2022). Natural axiomatic theories and consistency strength: A lakatosian approach to the linearity conjecture. Report.
Hamkins, J. D. (2025). Nonlinearity and illfoundedness in the hierarchy of large cardinal strength. *Monatshefte für Mathematik*. DOI: https://doi.org/10.1007/s00605-025-02082-1.
Holmes, M. R. and Wilshaw, S. (2024). Nf is consistent. https://arxiv.org/abs/1503.01406.
Kanamori, A. (2003). *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*. Springer. DOI: https://doi.org/10.1007/978-3-540-88867-3.
Kaye, R. and Wong, T. L. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, 48(4):497–510. DOI: https://doi.org/10.1305/ndjfl/1193667707.
Kechris, A. S. (1995). *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer. DOI: https://doi.org/10.1007/978-1-4612-4190-4.
Kunen, K. (2006). *Set Theory: An Introduction to Independence Proofs*. Elsevier, Sydney.
Kunen, K. (2009). *The Foundations of Mathematics*. Mathematical Logic and Foundations. College Publications.
Lewis, A. (1998). Large cardinals and large dilators. *The Journal of Symbolic Logic*, 63(4):1496–1510. DOI: https://doi.org/10.2307/2586663.
Lindström, P. (2003). *Aspects of Incompleteness: Lecture Notes in Logic 10*. Lecture Notes in Logic. Taylor & Francis.
Maddy, P. (2011). *Defending the Axioms: On the Philosophical Foundations of Set Theory*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199596188.001.0001.
Maddy, P. and Meadows, T. (2020). A reconstruction of steel's multiverse project. *Bulletin of Symbolic Logic*, 26(2):118–169. DOI: https://doi.org/10.1017/bsl.2020.5.
Martin, D. A. (1998). Mathematical evidence. In Dales, H. G. and Oliveri, G., editors, *Truth in Mathematics*. Clarendon Press. DOI: https://doi.org/10.1093/oso/9780198514763.003.0012.
Martin, D. A. (n.d.). *Determinacy*. Unpublished book manuscript.
Meadows, T. (2021). Two arguments against the generic multiverse. *Review of Symbolic Logic*, 14(2):347–379. DOI: https://doi.org/10.1017/S1755020319000327.
Montalbán, A. and Walsh, J. (2019). On the inevitability of the consistency operator. *Journal of Symbolic Logic*, 84(1):205–225. DOI: https://doi.org/10.1017/jsl.2018.65.
Steel, J. (2024). The comparison lemma: Young set theory workshop 2024.

Steel, J. R. (2014). Gödel's program. In Kennedy, J., editor, *Interpreting Gödel: Critical Essays*, pages 153–179. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511756306.012.

Steel, J. R. (2022). *A Comparison Process for Mouse Pairs*. Lecture Notes in Logic. Cambridge University Press. DOI: https://doi.org/10.1017/9781108886840.

Todorcevic, S. (2014). *Notes on Forcing Axioms*. World Scientific. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore: Volume 26. Edited by Chitat Chong, Feng Qi, Theodore A. Slaman, W. Hugh Woodin and Yue Yang. DOI: https://doi.org/10.1142/9013.

Walsh, J. (2025). On the hierarchy of natural theories. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2106.05794.

Woodin, W. (2001). The continuum hypothesis, part II. *Notices of the AMS*, 48(7):681–690.

A. Appendix

We show here – among other things – that the $\leq_\beta$ ordering is the same as the ordering obtained by considering the ordinals of the least transitive model of a theory. We start with a simple technical fact.

Lemma 22. *Suppose $x \in L_\alpha \cap \mathbb{R}$ and β is the second admissible ordinal greater than α. Then if $\exists y \in \mathbb{R}\ \varphi(y, x)$ is true statement where $\varphi(y, x)$ is Σ^0_1, then there is some $y \in L_\beta$ such that $\varphi(y, x)$.*

Proof. First note that there is tree T on ω^2 with $T \in L_\alpha$ such that for all $y \in \mathbb{R}$

$$\varphi(y, x) \Leftrightarrow y \in [T(x)].$$

Then, if γ is the first admissible above α, $T(x)$ is ill-founded iff L_γ thinks there is no order preserving map from $T(x, y)$ into the ordinals. And then this holds iff L_β thinks there is a branch $y \in \mathbb{R}$ through $T(x)$. Thus, there is some $y \in L_\beta$ such that $\varphi(x, y)$. □

The following lemma establishes that if T has a transitive model, then it has a model of minimal height in L.

Lemma 23. *Suppose T is a computable theory with a transitive model and that $\alpha = Ord^M$ where M is the shortest of these models. Then T has a model N in L with $Ord^N = \alpha$. In fact, such a model will be an element of L_β for any $\beta > \alpha$ where $L_\beta \models ZFC^-$.*

Proof. Let T and α be as described and let M be a transitive model of T with $Ord^M = \alpha$. Clearly $\alpha < \omega_1$ so we may fix $x_\alpha \in \mathbb{R}$ that codes a well-ordering of length α. First we claim that such an x_α exists in L. To see this suppose not. Then we must have $\omega_1^L < \alpha < \omega_1$. Now note that the statement

> There is countable transitive model of T.

is Σ^1_2 and so is true in L by the Lévy-Shoenfield theorem. Thus, we may fix a countable transitive model M^* of T with $M^* \in L$ and $|M^*|^L < \omega_1^L$. But then $Ord^{M^*} < \omega_1^L < \alpha$ contradicting the minimality of α.

Now with $x_\alpha \in \mathbb{R} \cap L$ coding α in hand, we observe that our assumptions imply that:

> There is some $m \in \mathbb{R}$ coding a model satisfying T and such that the ordinals of that model are isomorphic to the well-ordering encoded by x_α.

This statement is Σ^1_1 in x. Thus using Lemma 22, we may obtain some $m \in L_\beta \cap \mathbb{R}$ witnessing its truth, where $\beta > \alpha$ and $L_\beta \models ZFC^-$. The real m can then be decoded and collapsed in L_β to obtain a model N with $Ord^N = \alpha$. □

We now formalize the definition of our ranking function on theories.

Definition 24. Let ρ be a function from a computably axiomatizable theories to the ordinals be such that $\rho(T) = \alpha$ if α is the least α such that there is a transitive model M of T with $Ord^M = \alpha$; and let $\rho(T) = *$ if there is no such model, where $*$ is some non-ordinal from V_ω. Let us say that $T \preceq_{rank} S$ if $\rho(T) \leq_* \rho(S)$, where $\leq_* \subseteq (Ord \cup \{*\})^2$ is the usual ordering on the ordinals with $*$ added to the end.

Informally speaking, when we say $T \preceq_{rank} S$, we are saying that the shortest model of T, if there is one, is shorter than the shortest model of S.

Theorem 25. *$T \preceq_{rank} S$ is Π^1_2.*

Proof. It is easy to see that Lemma 23 implies $T \preceq_{rank} S$ iff

> For all $\alpha < \omega_1$, if $L_\alpha \models ZFC^-$, then $L_\alpha \models T \preceq_{rank} S$.

Then since this statement is Π^1_2, we are done.

□

Theorem 26. *If $T \preceq_{rank} S$, then $T \leq_\beta S$.*

Proof. Suppose $T \not\leq_\beta S$. Then there is a transitive model M of ZFC^- satisfying $Con_\beta(T)$ but not $Con_\beta(S)$. Let $\alpha = Ord^M$. Then by Lemma 23, we see that L_α thinks that $\rho(T) < \alpha$ and $\rho(S) = *$. Thus, $L_\alpha \models T \preceq_{rank} S$, which suffices by the proof of Theorem 25. □

We then note that the converse can fail if we consider theories that don't extend ZFC. Here is a somewhat artificial example demonstrating this.

Proposition 27. *There are computably axiomatizable T and S such that $T \leq_\beta S$ but $T \not\preceq_{rank} S$, supposing there is a transitive model of ZFC.*

Proof. Let S be ZFC and let T be KP plus the statement that there is a transitive model of ZFC. Let L_α be such that α is least such that L_α satisfies ZFC. Let L_β be the next admissible ordinal after α. Then β is least such that $L_\beta \models T$. So clearly we have $S \preceq_{rank} T$ and $T \not\preceq_{rank} S$. We now show that $T \leq_\beta S$. To do this, first let γ be the least ordinal greater than β satisfying ZFC^-. Note that $\gamma > \beta > \alpha$. Then let N be an arbitrary model of ZFC^- and suppose that N satisfies $Con_\beta(S)$. Then we must have $L_\alpha \in N$ and indeed $L_\gamma \in N$. Thus, N also satisfies $Con_\beta(T)$ as required. □

However, if we restrict our attention to theories extending ZFC in $\mathcal{L}_\in$ this glitch disappears.

Problem A.1. Suppose T and S are computably axiomatizable theories in $\mathcal{L}_\in$ that extend ZFC, then

$$T \preceq_{rank} S \Leftrightarrow T \leq_\beta S.$$

Proof. We just need the $\Leftarrow$ direction here, so suppose $T \not\preceq_{rank} S$. Then $\rho(T) >_* \rho(S)$ and so S must have a transitive model. If there are no transitive models of T, we trivially get $T \not\leq_\beta S$. So suppose that T also has a transitive model. Using Lemma 23, we may fix M satisfying S with $Ord^M = \rho(S)$ and fix β least such that L_β satisfies ZFC^- and $M \in L_\beta$. Similarly, we may fix N satisfying T with $Ord^N = \rho(T) > \rho(S)$. Now suppose toward a contradiction that $Ord^N < \beta$. But then $L^N = L_\gamma$ for some $\gamma < \beta$ and by Lemma 23, we must have $M \in L_\gamma$. But since L_γ satisfies ZFC this means β cannot have been the least model of ZFC^- with $M \in L_\beta$, so we have a contradiction. □