This is a ``preproof'' accepted article for *The Bulletin of Symbolic Logic*.
This version may be subject to change during the production process.
DOI: 10.1017/bsl.2025.10093

# OF SHEEP AND WOLVES

## EQUIVALENCE & DISAGREEMENT IN SET THEORY

TOBY MEADOWS

Abstract. This paper offers a philosophical overview and investigation of the problem of incompleteness in set theory and what this entails for the ensuing debates about proposed extensions of $ZFC$. The incompleteness of $ZFC$ is well-known and leaves us with a rich array of competing extensions. What should we make of disagreements between them? We start by considering second order logic and its categoricity theorems and how they might be used to compare different set theories. We then aim to use interpretability as a way of understanding that some of these debates are insubstantial. This culminates in some discussion of the relationship between interpretability and the generic multiverse. The second half of the paper then takes up a more modest goal: we search for common ground and settle for partial agreement between set theories in much the same way that physicists are often content with empirical agreement. We then aim to describe a natural bound on the amount of agreement that we can expect to obtain between reasonable extensions of $ZFC$.

Set theory purports to provide a foundation for mathematics by providing a framework into which every mathematical subject can be absorbed and understood. However, the subject matter of set theory itself arguably remains a vexed question. It is well-known that our best set theories are incomplete. Naturally enough, this incompleteness prompts debate over the propositions that are left undecided. Without any prospect of deciding such questions, doubts about the fixedness of the content of set theory are bound to arise and indeed have [Mostowski, 1967, Hamkins, 2012, Steel, 2014]. Problems of this kind put the project of set theory on a strange footing. While we happily embed arithmetic and analysis into the lower rungs of the cumulative hierarchy, we seem hampered in our efforts to understand set theory itself. When one person says $\varphi$ and another says $\neg\varphi$, who are we to believe? The development of frameworks for posing and settling these questions are fundamental to the set theoretic enterprise. Nonetheless, it's arguable that the incompleteness encountered in set theory runs so deep that the methodologies for settling such questions end up at least as controversial as the questions themselves.

My goal in this paper is to provide a metamathematical and philosophical investigation and overview of the logical mechanics of disagreement in contemporary set theory. On might think of this as a piece of mathematical philosophy of mathematics. In particular, I will argue that in many cases disagreement in set theory is not as serious as it first might seem. I will offer two perspectives for thinking this that divide the paper essentially in two. First, we'll investigate the idea that the problem is one of mere semantics. Perhaps these are are just definition debates and we are better understood as *arguing past each other*. This is the topic of Section 2. Second, we'll investigate the idea that there is sufficient

I'm extremely grateful to Juliette Kennedy and Jouko Väänänen for inviting me to speak on this material at their workshop on internal categoricity in Helsinki. They were even generous enough to give me two sessions to cover everything! I'm also very grateful to the attendees of the Irvine Logic Seminar who were patient enough to sit through an early version of this material. Finally, I'd like to thank the kind, anonymous referee who took the time to write such a helpful report.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*common ground* between the interlocutors of such debates that the stakes are lower than one might expect. This is the topic of Sections 3 and 4. In both cases, our main investigative tool will be relative interpretability. However, we shall begin our discussion in Section 1 by considering an important kind of argument for the opposing position that most debates about set theoretic questions are substantive. Besides setting the scene for our investigation, it will also introduce us to an intriguing and important recurring character in our story: second order logic.

In more detail, the paper is set out as follows. We being in Section 1 with second order logic and its categoricity theorems. We discuss the prospects for these and related tools for understanding disagreement and we see that this perspective pushes the idea that one party to a disagreement must be right while the other is wrong. We observe that this attitude seems to contrast with those of contemporary set theorists, who sometimes take the models delivered by forcing to be just as good as their ground models. Second order logic and forcing are constant lietmotifs in this paper; the former will push us toward rigidity while the latter points toward a more flexible point of view. We then introduce the formal tools of interpretation in Section 2. Our primary goal here is to take seriously the idea the parties to disagreement about set theory could be merely arguing past each other and so, in some sense, both of them could be right. The investigation here takes the form of a Goldilocks search whereby we rule out a number of equivalence relations that are either too strong or too weak to plausibly model the idea of a merely verbal dispute. This will lead us to an interpretative understanding of the generic multiverse. In Section 3, we'll shift our perspective away from disagreement regarding entire theories to those parts that seem particularly important: the common ground. The underlying motivation for this is an analogy with the idea of empirical equivalence between theories in physics. We'll show how the tools of interpretability can be naturally generalized to work in partial contexts and illustrate some limitations of their applicability. Finally, in Section 4 we'll ask just how much common ground we can reasonably expect to find between strong set theories. Again we undertake a Goldilocks search that leads us on the long road from agreement on the logic of transitive models to a logic that brings together our recurring characters, second order logic and forcing: Ω-logic.

Given the length and density of this paper, it also seems apropos to provide a few – hopefully helpful – remarks for the reader. First and regarding the philosophical side of things, my goal below has been to deliver a somewhat opinionated overview rather than a comprehensive exploration of the rich dialectics that I believe the problems discussed below pose. I do this, in part, since this paper is long enough already, but also because I think these are philosophical discussions that are beginning to attract more attention. As such, I would prefer to say a few things that are perhaps a little incautious in the hope that others will come to refine and correct me in the future. Nonetheless, whenever I have felt that I'm stepping into controversy or taking a less trodden path, I have tried to highlight this to the reader so that they might not be misled. This seemed to me like the right strategy for a paper of this kind. On the mathematical side of things, the reader will note that I have included quite a lot of proofs, sometimes of quite well-known propositions. I have done this because I believe that proofs are very much part of the content of this discussion. Nonetheless, I have aimed to merely sketch the more familiar claims, while being more comprehensive with material less known to a general audience. Moreover, I have also tried to design the document so that it might be read by those who would prefer to pass these proofs by – even if only on a first reading. As such, a number of theorems are stated in

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

a sub-optimal form in order to better fit with the discursive thread that is being traced at the time. When this occurs, we'll use the footnotes to point the reader to their canonical statements.

## 1. Comparison & Categoricity

The problem of incompleteness is, of course, hardly peculiar to set theory. With regard to arithmetic, we often appeal to Dedekind's categoricity theorem in this regard. A pleasing argument shows that any two models of arithmetic in the full semantics of second order logic must be isomorphic to each other.[1] Thus given any arithmetic statement, either it or its negation will be true in such a model. This result doesn't tell us which one is the right one, but it does seem to give us reason to be confident that there is a right answer. This strategy is not without controversy. For example, its use of second order logic in its full semantics appears to ask us to assume the fixedness of something much more complicated than original problem.[2] Nonetheless, we may still wonder whether this strategy can be generalized to the problem of set theory.

**Theorem 1.** *[Zermelo, 1976] Suppose $\mathcal{M}$ and $\mathcal{N}$ are models of second order $ZFC$ in the full semantics.*[3] *Then either $\mathcal{M}$ or $\mathcal{N}$ is isomorphic to a rank initial segment of the other.*

Following [Shepherdson, 1951], we may prove this by taking the transitive collapses of $\mathcal{M}$ and $\mathcal{N}$, noting that since we are using the full semantics both $\mathcal{M}$ and $\mathcal{N}$ are well-founded. An inductive argument then reveals that the respective collapses must be $V_\alpha$ and $V_\beta$ for some ordinals $\alpha$ and $\beta$. Thus, one must be an initial segment of the other.

We have here a nice uniqueness result: models of second order $ZFC$ in the full second order semantics are unique up to a rank initial segment embedding. However, it doesn't give us a full isomorphism and so we don't learn that $\mathcal{M}$ and $\mathcal{N}$ decide the sentences of $\mathcal{L}_\in$ in the same way. The strategy above is thus blocked. Moreover, we might again worry about the use of the full semantics of second order logic.[4] From this point of view this weak categoricity result doesn't seem so surprising. Indeed, we might think it on the verge of begging the question. Fortunately, Jouko Väänänen has provided an extension of Zermelo's theorem that makes no appeal to second order logic and even better, provides a genuine isomorphism. The statement of the theorem requires a change to our standard practice of doing set theory in the language $\mathcal{L}_\in$, which consists of just one two-place relation symbol, $\in$. Rather, we work in a language $\mathcal{L}_{\in,\in^*}$ with a pair of two-place relation symbols, $\in$ and $\in^*$. We then use a theory $ZFC(\in,\in^*)$ which has the usual axioms of $ZFC$ for both $\in$ and $\in^*$. For example, for any sets $x$ and $y$, there is both an $\in$-pair and an $\in^*$-pair of $x$ and $y$. Then with regard to the Separation and Replacement schemas for $\in$ and $\in^*$, we permit arbitrary formulae of the combined language $\mathcal{L}_{\in,\in^*}$.

**Theorem 2.** *[Väänänen, 2019] $ZFC(\in,\in^*)$ proves that $\in$ and $\in^*$ determine isomorphic structures.*

[1] A model of second order in its *full* semantics logic includes every subset of the its object domain in its second order domain. See Chapter 4.2 of [Shapiro, 1991] for a detailed description of the full semantics; it is called standard semantics there. Dedekind's theorem can be found as Theorem 4.8 in [Shapiro, 1991].

[2] See [Meadows, 2013] and [Maddy and Väänänen, 2023] for some discussion of this controversy.

[3] See Section 5.2 of [Shapiro, 1991] for a description of second order $ZFC$. Essentially, we just replace the Separation and Replacement schemas with their natural second order counterparts.

[4] Speaking a little imprecisely, a model of second order logic in its full semantics is a model that is closed under subsets: whenever $x$ is an element of $\mathcal{M}$, so is every one of its subsets. More precisely, whenever $\mathcal{M} = \langle M, \in_\mathcal{M} \rangle$ and $x \in \mathcal{M}$ and $y \subseteq \{z \in M \mid z \in_\mathcal{M} x\}$, then there is some $w \in M$ such that for all $z$, $z \in_\mathcal{M} w$ iff $z \in y$.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*Proof.* (Sketch[5]) First, we let $\mathcal{U} = \langle V, \in \rangle$ and $\mathcal{U}^* = \langle V, \in^* \rangle$ be the structures associated with each membership relation and note that they share the same domain. We show that $\mathcal{U}$ and $\mathcal{U}^*$ are isomorphic. Since both membership relations are permitted in the Replacement Schemas, we may mutually collapse $\mathcal{U}$ and $\mathcal{U}^*$ into each other. A little more precisely, we have collapse functions:

$$\pi^* : \mathcal{U}^* \cong \mathcal{M}^* \subseteq \mathcal{U} \text{ and}$$
$$\pi : \mathcal{U} \cong \mathcal{M} \subseteq \mathcal{U}^*.$$

With some work (that we omit for simplicity), it can be shown that these collapses yield ordinals of the same length; i.e., that $Ord^{\mathcal{M}^*} = Ord^{\mathcal{U}}$ and $Ord^{\mathcal{M}} = Ord^{\mathcal{U}^*}$. Now observe that since $\mathcal{M}$ is a submodel of $\mathcal{U}^*$, $\pi^*$ also collapses $\mathcal{M}$ to some $\mathcal{N}$ . We use this to make a squeezing argument to obtain the isomorphism. More specifically, we we have it that

$$\pi^* \restriction \mathcal{M} \cong \mathcal{N} \subseteq \mathcal{M}^* \subseteq \mathcal{U}.$$

Since $\mathcal{N}$ and $\mathcal{U}$ are transitive, isomorphic structures with respect to the $\in$-membership relation, we must have $\mathcal{N} = \mathcal{U}$ and so

$$\mathcal{U} \subseteq \mathcal{M}^* \subseteq \mathcal{U},$$

which means that $\mathcal{M}^* = \mathcal{U}$. Thus, we have $\pi^* : \mathcal{U}^* \cong \mathcal{U}$ as required. Furthermore, the isomorphism is definable in $\mathcal{U}^*$ and a similar argument shows that such an isomorphism is also definable in $\mathcal{U}$. □

Clearly, the position from which this theorem is articulated is quite different to the standard $ZFC$ position employed in Zermelo's theorem. What should we make of this? The most obvious distinction is that we are not merely talking about models, which are sets, but rather a pair of relations that mutually perceive each other as proper classes. A little informally, we might say that our position has shifted; we are no longer on the outside looking in, but rather, we are part of the picture comparing one membership relation to another. We might say that we have shifted from an external perspective to an internal one. For this reason, Väänänen's theorem is sometimes said to establish *internal categoricity* in contrast to Zermelo's external perspective.[6] But why would this matter? I think the simplest way to convey the potential philosophical value of this move is through a kind of thought experiment.

Recall that we began this discussion by observing the problem of incompleteness and its impact on our perception of the subject matter of set theory. I raised the concern that, with respect to some undecided statement $\varphi$, we might have one person who proposes $\varphi$ while another proposes that $\neg\varphi$. Suppose that we are those two people. Given our disagreement, we might wonder whether we can be talking about the same subject matter. In this situation, it seems quite natural for me to augment my set theory with an extra membership relation $\in^*$ to accommodate your possibly distinct notion. Then assuming we both agree that $ZFC$ provides a sound basis on which to theorize about sets, Väänänen's theorem tell us that the universe associated with my $\in$-relation is isomorphic to that of

[5] My goal is to convey the main idea of the proof rather than nail down all the details. This is because I want to highlight that these kinds of moves were present in Shepherdson's argument and because we shall see them again later. I should also note that the axiom of choice is not used in this proof, so it still holds if we use $ZF(\in, \in^*)$ instead. I've used $ZFC$ above just because it is the standard foundation for mathematics and thus, fits the discussion a little better. Similar remarks apply later to Enayat's Theorem 16 and Corollary 19. I thank Ali Enayat for prompting to note this.

[6] More detailed discussion of internal categoricity and its philosophical significance can be found in [Maddy and Väänänen, 2023] and [Button and Walsh, 2018].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

your $\in^*$-relation. This seems like a big step forward. We are now on a similar footing to that we had with regard to arithmetic. We haven't decided the question, but the isomorphism tells us that there is a right answer.

At least that's one way of telling the story. I'd now like to spend some time drawing out some of the more controversial assumptions that underlie this little argument.

(1) The domain for both $\in$ and $\in^*$ should be the same.
(2) Both membership relations should be permitted into formulae used in the Separation and Replacement schemas.
(3) We only need to agree on $ZFC$.

We shall address each of these assumptions in turn below.

*(1) The domain for both $\in$ and $\in^*$ should be the same.* In the setup of $ZFC(\in, \in^*)$, our quantifiers range over a single domain objects. This assumption is crucial for us to be able to obtain the genuine isomorphism between the $\in$ and $\in^*$ relations rather than a mere embedding of one relation into a rank initial segment of the other. I'm not aware of a convincing philosophical argument that this is either a good or a bad assumption, however, I'll now record a couple of observations that may give some reason for caution. First, we might think that some simple consequences of this theory are counterintuitive. To see this, let's consider an obvious example. In $ZFC(\in, \in^*)$ there will be an emptyset, $\emptyset$, associated with $\in$ and an emptyset, $\emptyset^*$, associated with $\in^*$. In general, they will be distinct and in these cases, there could be some $x$ such that $x \in^* \emptyset$ and $y$ where $y \in \emptyset^*$. Thus, we have objects that are empty according to one relation and nonempty according to another. Of course, this is not a mathematical issue: Väänänen's extension of Zermelo's theorem is both elegant and intriguing. However, this issue may give us cause to wonder how well it fits with our thought experiment above.[7]

This brings us to the second observation: perhaps there is a theory better suited to our argument than $ZFC(\in, \in^*)$. We motivated our thought experiment by imagining that you and I were concerned that our discourse about set theory might not – as we might have assumed – have been focused on a particular subject matter. In an attempt to formalize this situation, we proposed to expand our languages to accommodate two membership relations. Väänänen's theorem then appeared to resolve the situation. However, we also noted that the use of a shared domain led to some counterintuitive consequences. With that in mind, we might think that a more natural theory for our thought experiment would be one that separated the domains associated with the two membership relations. Thus it might be more natural to incorporate predicate symbols $V$ and $V^*$ that we treat as domains for $\in$ and $\in^*$ in the obvious way. We might then add an axiom demanding that $V$ and $V^*$ partition the domain of quantification. This will remove the counterintuitive features we've just discussed, but it will also mean that the isomorphism is lost. As with Zermelo's Theorem 1, we'll only get an embedding of one structure into an initial segment of the other. So perhaps less ground has been gained than appeared at first sight.

[7] To be very clear, I am not suggesting that Väänänen intended to apply his theorem in this way. Moreover, I do not wish to claim that this thought experiment exhausts the opportunities for philosophical application of this theorem.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*(2) Both membership relations should be permitted into formulae used in the Separation and Replacement schemas.* $ZFC(\in, \in^*)$ has two Separation schemas; one for $\in$ and one for $\in^*$. For example on the $\in$-side, we have it that for any $x$ there is a set $y$ such that for all $z$

$$z \in y \leftrightarrow z \in x \wedge \varphi(z)$$

where crucially $\varphi(z)$ is any formula of $\mathcal{L}_{\in,\in^*}$. Thus, we can use both $\in$ and $\in^*$ to provide conditions for separation. This is a powerful assumption that is also crucial to obtaining Väänänen's result. In particular, we use the shared version of Replacement to obtain a definable collapse of one membership relation within the other. But what reason do we have for thinking that this assumption is a reasonable one within the context of our thought experiment?

This question has a more canonical response that originated in the context of the induction schema in arithmetic [Parsons, 1990, Field, 1999]. To apply this to set theory, rather than merely taking the axioms of $ZFC$ at face value, we are implored to consider the intentions behind them. With regard to the Replacement Schema, it literally says the following:[8] Suppose $x$ is a set and that $\boldsymbol{F}$ is a (class) function that is definable in $ZFC$ by a formula of $\mathcal{L}_\in$. Then the pointwise image of $x$ via $\boldsymbol{F}$ is itself a set; i.e., $\boldsymbol{F}``x$ is a set. We might argue that the restriction to (class) functions that are expressible in $\mathcal{L}_\in$ is merely forced upon us by our choice of language: this is as much as we can do with the tools we have available. But really the motivation for Replacement is that given any set $x$ and any (class) function $\boldsymbol{G}$ whatsoever, $\boldsymbol{G}``x$ should also be a set regardless of whether or not $\boldsymbol{G}$ is definable. So the story goes. To articulate this *enhanced* version of Replacement, we need to move to a second order version of set theory or one that allows quantification over classes. We then regard the Replacement schema of $ZFC$ as a mere approximation (among many) of thetrue second order version.[9] But for our purposes in $ZFC(\in, \in^*)$ this is not required. In order to stay faithful to this intention, the natural thing do to is to admit – not only those (class) functions definable using a single membership relation – but rather (class) functions that can be defined using both of them. If we are taking seriously the idea that any (class) function should enjoy Replacement, then those defined using $\mathcal{L}_{\in,\in^*}$ are certainly within our scope.

*(3) We only need to agree on $ZFC$.* Our use of Väänänen's theorem in the thought experiment aimed to show that if you and I both agree on $ZFC$ then we have good reason to think that we are talking about the same subject matter. However, our motivation coming into that example was a little more ambitious. We were concerned about the possibility that you think $\varphi$ is true while I do not, where $\varphi$ is some sentence that is not decidable by $ZFC$. But this puts some pressure on our decision to use $ZFC(\in, \in^*)$ to model our debate. To see this, let's make things more specific and suppose that you think that the continuum hypothesis, $CH$, is true while I think it is not. Then it would seem natural to represent this situation by letting, say, $\in$ have $CH$ satisfied while $\in^*$ has $\neg CH$. We might record these statements as $CH^\in$ and $\neg CH^{\in^*}$ respectively. Following through on this, we should then use the theory

$$T = ZFC(\in, \in^*) + CH^\in + \neg CH^{\in^*}$$

[8] We just discuss Replacement here. A similar story can be told for Separation.

[9] Alternatively one might invoke a story about the indefinite extensibility of such classes, rather than referring to a definite totality of them.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

to model our debate. If we do this, then Väänänen's theorem still holds since it just requires $ZFC(\in, \in^*)$. Thus, we get an isomorphism between the $\in$ and $\in^*$ relations and this entails that they satisfy the same statements. But this is, of course, impossible since $\in$ and $\in^*$ disagree on $CH$. So $T$ is inconsistent. What should we make of this? The first and perhaps most obvious response is that this reveals that one of us is wrong. Indeed, this little argument is really just another way of saying – as we did with arithmetic – that there is a right answer. It gives us no clues regarding how to identify it, but we are given reason to believe that there is one.

In this paper, I want to open the door to a second response and provide a critical investigation of its prospects and boundaries. Rather than taking the road above, I would like to retain some doubts around the assumptions that motivate our thought experiment and see where this can lead us. For one reason to explore this territory, we note that a non-trivial section of the set-theory community appears to retain doubts as to the fixedness of the subject matter of set theory. One example of this occurs in the debate between John Steel and Hugh Woodin regarding what is known as the generic multiverse [Steel, 2014, 2004, Woodin, 2012, 2011a, 2004]. Far too briefly, this is a view inspired by the ubiquity of undecidability in set theory.[10] Rather than accepting incompleteness as a problem to be solve, we are encouraged to think of it as evidence of a failure to pose these questions correctly. In particular, we take it that generic extensions of models of set theory are, so to speak, just as good as their ground models.[11] Other examples of roughly this attitude can be found in Joel David Hamkins' multiverse approach to set theory and Saharon Shehah's probabilistic attitude to set theory [Hamkins, 2012, Shelah, 1991]. In the sections that follow we are going to investigate this territory by exploring two perspectives through which one might not take all debates in set theory seriously: that some debates are merely verbal; and that there is sufficient common ground that we need not worry.

## 2. Arguing Past Each Other

The *phenomenon of arguing past on another* should be familiar one. Two disputants appear to be vigorously engaged in a deadlocked debate when it is revealed that the disputants do not understand the meaning of some crucial term or claim in the same way. With this revealed and after a little more discussion, they realize that once they understand each other's meaning, the debate is resolved and their disagreement has been translated away. I aim to explore the idea that many debates about axioms in the foundations of mathematics can be seen as instances of disputants arguing past each other. The idea then is that these disputants are really saying the same thing, but just employing language in divergent ways to achieve this. To put this idea on a more precise footing, we shall use the theory of interpretation. We start with a brief overview of this theory as it will provide us with a mathematical framework for analyzing disputes between alternative foundational theories[12] . As intimated above, the fundamental instrument of interpretation is translation. Unless stated otherwise, we shall restrict our attention to the language of set theory, $\mathcal{L}_\in$.

[10] I'd like to be clear that I'm not claiming that either Steel or Woodin believe the view described; rather that this debate has been conducted in response to a folk attitude that has been prevalent since the arrival of forcing. Their work has brought much needed precision to what was once a murky disagreement often debated on the basis of dubious metaphysical assumptions.

[11] We'll revisit the generic multiverse later in this paper, however, detailed analysis of its goals can be found in [Maddy and Meadows, 2020, Meadows, 2021].

[12] A more comprehensive overview can be found in [Visser, 2006].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Definition 3.** [Tarski, 1953] Let $T$ and $S$ be theories in $\mathcal{L}_\in$. We say $T$ *interprets* $S$ if there are formulae $\delta_t(x)$ and $\varepsilon_t(x,y)$ recursively determining a translation $t : \mathcal{L}_\in \to \mathcal{L}_\in$ such that:[13]

$$\begin{aligned} t(x = y) &:= x = y \\ t(x \in y) &:= \varepsilon_t(x,y) \\ t(\neg\varphi) &:= \neg t(\varphi) \\ t(\varphi \wedge \psi) &:= t(\varphi) \wedge t(\psi) \\ t(\forall x \varphi) &:= \forall x(\delta_t(x) \to t(\varphi)) \end{aligned}$$

and for all $\varphi \in \mathcal{L}_\in$

$$S \vdash \varphi \Rightarrow T \vdash t(\varphi).$$

Informally speaking, the $T$ user provides a translation of the $S$ user's $\in$-relation and defines a domain on which it is interpreted. Then anything that $S$ can prove can be proven in $T$ through the lens of this translation. Beyond this syntactic perspective, the following fundamental theorem provides a more picturesque understanding of the mechanics of interpretation.

**Theorem 4.** *Suppose $T$ interprets $S$ via $t$. Then $t$ determines a function*

$$t^* : mod(T) \to mod(S)$$

*such that for all $\mathcal{M} \models T$, $m_0, ..., m_n \in M$ and $\varphi(\bar{x}) \in \mathcal{L}_S$*

$$\mathcal{M} \models t(\varphi)(m_0, ..., m_n) \;\Leftrightarrow\; t^*(\mathcal{M}) \models \varphi(m_0, ..., m_n)$$

*when for all $i \leq n$, $\mathcal{M} \models \delta_t(m_i)$.*[14]

We call such a $t^*$ a *mod-functor* and we'll frequently omit the $^*$ below as it should not cause any confusion. By way of explanation, this theorem tells us that whenever $T$ interprets $S$ we obtain a function taking a model of $T$ and returning the internal model of $S$ defined within it. Then using the soundness and completeness theorems, we see that whenever $T$ interprets $S$, then $S$ is consistent relative to $T$. We now consider a classic example of an interpretation.

**Example 5.** Let $ZFC + MC$ be $ZFC$ plus the assumption there is a measurable cardinal and let $ZFC + V = L$ be $ZFC$ plus the assumption that every set is constructible (i.e., that $V = L$). Dana Scott showed that $ZFC + MC$ implies that $V \neq L$, so both theories cannot be true at the same time. However, it is possible for $ZFC + MC$ to interpret $ZFC + V = L$. We define our interpretation $t$ by letting $\delta_t(x)$ say that $x$ is constructible (i.e., $x \in L$); and we let $\varepsilon_t(x,y)$ simply say that $x \in y$. Thus we preserve the interpretation of the membership relation but restrict the domain of the interpretation

[13] Note that the = clause is somewhat unusual in that we are demanding that identity, unlike membership, is left alone by the translation. Since we will almost always be working in extensions of $ZFC$ this doesn't make much difference since $ZFC$ is Morita complete in the language of Barrett and Halvorson [2016] and Meadows [2023b]. In the language of Visser and Friedman [2014], all the interpretations considered below are *identity preserving*. In places where results are sensitive to this issue, we shall take care to highlight this to the reader using the footnotes.

[14] This useful theorem seems to be part of the folklore although it is arguably implicit in Section 2(d) of [Feferman, 1960]. Extensive use and generalization of it can be found in [Visser, 2006]. A sketch of the proof can be found for Theorem 3.2 of [Meadows, 2023b].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

to the constructible hierarchy. Thus $t(\varphi)$ is just $\varphi^L$. It can then be seen that for all axioms $\psi$ of $ZFC$

$$ZFC + V = L \vdash \varphi \;\Rightarrow\; ZFC + MC \vdash t(\varphi)$$

and $ZFC + MC \vdash (V = L)^L$ essentially by definition.[15]

Returning to our motivating philosophy, what should we make of this example? Does it tell us that a debate between the $ZFC + MC$ user and the $ZFC + V = L$ user would be an instance of arguing past each other? Can the debate be deflated using the interpretation $t$? I don't think that would be right. However, something significant is certainly happening. The $ZFC + MC$ user is able to translate everything that the $ZFC + V = L$ user says and prove it themselves. Under this interpretation, it is simply not possible for them to disagree.[16] Nonetheless there are a number of features of this interpretation that should make us hesitant to suggest that the substance of the debate has been removed and they are really just talking past each other.

In the remainder of this section, we are going to work our way toward an interpretability relationship that plausibly captures our idea of a debate based on misinterpretation. We shall do this by considering a sequence of stronger and stronger relationships noting at each point their shortcomings with respect to our underlying motivation. They will all be too weak for our purposes, but this will be resolved in the following section.

2.1. **Too weak.** Let us start by observing that mere interpretability – as instantiated in Example 5 – is generally asymmetric. While one theory can interpret another, this does not mean that the other can interpret the original. For example, it is easy to see that $ZFC + V = L$ cannot interpret $ZFC + MC$. This is because $ZFC + MC$ says there is a measurable cardinal $\kappa$, and so it can prove that *there is* a model of $ZFC + V = L$. This means that $ZFC + MC$ proves *outright* – and not merely relatively – that $ZFC + V = L$ is consistent. Thus, if $ZFC + MC$ were also consistent relative to $ZFC + V = L$, then $ZFC + MC$ could prove its own consistency, thus, contradicting Gödel's second incompleteness theorem. Thus, there is an interpretative asymmetry.

But what does this tell us about our debate scenario? While $ZFC + MC$ is in a position to interpret $ZFC + V = L$ in such a way that it can provide a complete simulation of $ZFC + V = L$ via translation, $ZFC + V = L$ cannot do the same. Only one party to the dispute is able to make a move that could put them in position to translate the debate away. Indeed, the debate ensuing out of Example 5 is in an even worse position. Not only is the $ZFC + V = L$ user unable to interpret $ZFC + MC$ and thus, simulate $ZFC + MC$ via translation; the $ZFC + V = L$ user proves that there are no measurable cardinals. As such, I think it could be reasonable for the $ZFC + V = L$ user to conclude that $ZFC + MC$ is just plain false. Perhaps $ZFC + MC$ does not have an interpretation since it simply doesn't deserve one. On the other side of this dialectic, the $ZFC + MC$ user may draw different conclusions. The $ZFC + MC$ user may see the $ZFC + V = L$ user's inability to simulate $ZFC + MC$ as indicative of a pathological weakness in $ZFC + V = L$. They might claim that $ZFC + V = L$ is

[15] For more details see the opening chapters of [Devlin, 1984].

[16] More specifically, there will be no sentence $\varphi \in \mathcal{L}_\in$ such that the $ZFC + MC$ user asserts – or better proves – $t(\varphi)$ while the $ZFC + V = L$ user proves $\neg\varphi$.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

unduly *restrictive* as theory purporting to provide a foundation for mathematics.[17] Regardless of who is right, it seems clear that they are not merely arguing past each other. This asymmetry motivates an obvious improvement of mere interpretation.

**Definition 6.** $T$ and $S$ are *mutually interpretable* if there are translations, $t$ and $s$ such that

$$t : mod(T) \leftrightarrow mod(S) : s.$$

It is no longer enough that $T$ interprets $S$, they must be able to interpret each other.

**Example 7.** $ZFC$ and $ZFC + V = L$ are mutually interpretable. The interpretation, $t$, described in Example 5 also works for $ZFC$ to interpret $ZFC + V = L$. Given a model of $ZFC$ we just use its version of $L$ to get a model of $ZFC + V = L$. Note that, unlike $ZFC + MC$, $ZFC$ alone cannot prove the consistency of $ZFC + V = L$ since it lacks the measurable cardinal or any other extra assumption that would permit such a proof to go through. Indeed, $ZFC$ is – strictly speaking – a weakening of $ZFC + V = L$. Thus to interpret $ZFC$ in $ZFC + V = L$, we can just use the *identity translation* $i$ that neither restricts the domain nor alters the membership relation. This works simply because a model of $ZFC + V = L$ already is a model of $ZFC$.

Mutual interpretability is clearly an equivalence relation on theories. Does it give us a more plausible relation to represent our motivating idea of arguing past each other? Clearly, the asymmetry of mere interpretability has been removed. However, I now want to discuss three reasons to be reserved about whether mutual interpretability is really sufficient to deflate debates between alternative foundations of mathematics.

(1) Practical triviality
(2) Poor fit
(3) Loss in translation

2.1.1. *Practical triviality.* One reason to pause is the comparative ease with which mutual interpretability results can be obtained. The following theorem of Per Linström gives a stark illustration of this. Recall that a theory $T$ is *essentially reflexive* if $T$ can prove the consistency of each of its finite fragments.

**Theorem 8.** *[Guaspari, 1979, Lindström, 1979] For essentially reflective theories $T$ and $S$ that can interpret $PA$, the following are equivalent:*

(1) *$T$ interprets $S$; and*
(2) *Every $\Pi^0_1$ theorem of $S$ is a theorem of $T$.*

The essential idea of the proof is to work within $T$ and use the completeness theorem to define a model of $S$.[18] The assumption around $\Pi^0_1$-statements ensures that the construction does not get stuck. This then leads to the following – perhaps counterintuitive – result.

**Fact 9.** *$ZFC$ and $ZFC + \neg Con(ZFC)$ are mutually interpretable.*

[17] For a more thorough discussion of restrictiveness in foundations see [Feferman et al., 2000, Maddy, 1997, Meadows, 2022].

[18] See Theorem 6.6 in [Lindström, 2003] for more details.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*Proof.* (Sketch) Clearly $ZFC + \neg Con(ZFC)$ can interpret $ZFC$ using the identity interpretation. In the other direction, we note that the reflection theorem ensures that both theories are essentially reflexive. Moreover, it can be seen that $ZFC + \neg Con(ZFC)$ has no more $\Pi^0_1$ consequences than $ZFC$. Thus, $ZFC + \neg Con(ZFC)$ can also interpret $ZFC$. □

Morally speaking, $ZFC$ and $ZFC + \neg Con(ZFC)$ are very different theories. The latter theory augments $ZFC$ with a statement that is in about as bad faith as one could imagine with its base: it says that it's inconsistent. Assuming – as we do – that $ZFC$ is actually consistent, then the only models of this theory will be ones where the natural numbers are nonstandard and their putative witnesses to the inconsistency of $ZFC$ will be among them. It seems very unlikely that a hypothetical debate between users of these theories is one in which the participants are merely arguing past each other. I think the right lesson to take from this is that mutual interpretability simply sets the bar on equivalence too low. It certainly tells us something of value, but it doesn't tell us that arguments between proponents of mutually interpretable theories aren't engaging in a substantive debate.

2.1.2. *Poor fit.* Another reason for reservation, is the weakness of the simulation offered by an interpretation. When $T$ interprets $S$ via some translation $t$, we know that every theorem of $S$ is provable – modulo the translation – in $T$. We might say that the simulation is *complete*. But is it also *sound*? Returning to the characters of Example 5, we find a situation where this is not the case. Recall that we had a translation $t$ such that for all sentences $\varphi$ of $\mathcal{L}_\in$

$$ZFC + V = L \vdash \varphi \;\Rightarrow\; ZFC + MC \vdash t(\varphi).$$

However, the converse fails. For example, $ZFC + MC$, proves that there is a countable transitive model satisfying $ZFC$ and the statement that $0^\#$ exists, which we abbreviate as $0^\#\exists$. The existence of such a model is a $\Sigma^1_2$ proposition and thus, by the Lévy-Shoenfield theorem,[19] it doesn't change its meaning when its quantifiers are restricted to $L$. Thus, $ZFC + MC$ also proves that there is a transitive model in $L$ that satisfies $ZFC + 0^\#\exists$. But from Gödel's second theorem, we know that $ZFC + V = L$ cannot prove its own consistency let alone that there is a transitive model of $ZFC$, so the arrow above cannot go in both directions. Regarding soundness, we see that $ZFC + MC$ proves more than $ZFC + V = L$ even through the translation. Moreover, the excess statements are ones with regard to which a reasonable advocate of $ZFC + V = L$ might demur. Of course, this example involves interpretative asymmetry, however, this problem also occurs between $ZFC$ and $ZFC + \neg Con(ZFC)$. In particular, when we interpret $ZFC$ in $ZFC + \neg Con(ZFC)$ using the identity translation, we end up with $\neg Con(ZFC)$ being an obvious theorem of $ZFC + \neg Con(ZFC)$ but not of $ZFC$.

What should we make of this? In the context of our thought experiment, we see that mere interpretability doesn't provide a perfect simulation. When a theory $T$ interprets a theory $S$, we see that rather than proving *exactly* what the $S$ proves, $T$ proves everything $S$ proves and possibly *more*. Suppose I am the $S$ user and you are using $T$ to interpret me via some interpretation $t$ that overshoots in this way. You might claim that you are providing a simulation that deflates any potential debate, but I may object that you are proving too much. I might think that among your extra theorems there is a false statement. My reason for using $T$ could be that it represents the very limit of my present

[19] See Theorem 13.15 of [Kanamori, 2003] for a statement and proof.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

knowledge and therefore that the excess given by $T$ goes too far. This lack of fit motivates our next equivalence relation, which is based on a refinement of ordinary interpretation.

**Definition 10.** Let us say that $T$ *faithfully interprets* $S$ if there is a translation $t$ such that

$$S \vdash \varphi \ \Leftrightarrow\ T \vdash t(\varphi).$$

We thus solve the fit problem by fiat. We demand that the translation proves exactly the translations of the theorems of $S$. Such an interpretation seems to have a better claim to providing a genuine simulation. Faithful interpretation also has a pleasing semantic version: $T$ faithfully interprets $S$ via $t$ if and only if the associated mod-functor $t : \mathit{mod}(T) \to \mathit{mod}(S)$ is a surjection up to elementary equivalence.[20] The use of faithful interpretation gives a more plausible claim to providing a proper simulation and thus we might wonder if it is a suitable model for debates where parties argue past each other. Disappointingly, we see that for a wide class of theories, faithful interpretation is almost as easy to obtain as ordinary interpretation.

**Theorem 11.** *[Lindström, 1984] Let $T$ and $S$ be essentially reflective theories that can interpret $PA$. Then the following are equivalent:*

(1) *$T$ faithfully interprets $S$; and*
(2) *Every $\Pi^0_1$ theorem of $S$ is a theorem of $T$ and every $\Sigma^0_1$ theorem of $T$ is a theorem of $S$.*

Roughly speaking, this is a generalization of Theorem 8 in which we use the completeness theorem in $T$ to obtain a model of *every* complete theory extending $S$. This then gives us every model of $S$ up to elementary equivalence. To do this, we use the tree of different completions of $S$ as given by Henkin's completeness theorem.[21] We then use some inherent incompleteness in $T$ to chose a path through that tree depending on how certain undecidable sentences turn out in a model.[22] The conditions of (2) in the theorem ensure this construction is not derailed. With this in hand, we can easily obtain the following result.

**Fact 12.** *$ZFC$ and $ZFC$ plus the statement "there is no $\omega$-model of $ZFC$" are mutually faithfully interpretable.*

This follows since these theories have exactly the same $\Sigma^0_1$ and $\Pi^0_1$ consequences; and indeed, this is the necessary and sufficient condition for mutual faithful interpretation. The requirements are obviously stricter than those for mere mutual interpretation, but as the Fact above shows, we still seem to end up with something that is far too easy to obtain.[23] In particular, the claim that there is no $\omega$-model of $ZFC$ seems like a substantive point of disagreement.

It's also possible to obtain a little insight into why faithful interpretation doesn't deliver what we want. In the proof of Theorem 11, we need to make a bunch of random decisions to ensure that we capture

[20] In other words, for all models $\mathcal{N}$ of $S$ there is a model $\mathcal{M}$ of $T$ such that $t(\mathcal{M}) \equiv \mathcal{N}$. The claim is proven using a simple compactness argument.

[21] See Theorem 6.14 in [Lindström, 2003] for more details.

[22] In $ZFC$, we can use the continuum pattern for this. See Theorem 2.10 of [Lindström, 2003] for a way of doing this in arithmetic.

[23] For an example where faithful interpretation is ruled out, note that $ZFC + \neg Con(ZFC)$ cannot faithfully interpret (on any interpretation) $ZFC$ since it has an extra $\Sigma^0_1$ consequence in the form of $\neg Con(ZFC)$.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

all the models of the interpreted theory up to elementary equivalence. To do this, we make use of the undecidability of the interpreting theory. This allows us to make infinitely many choices through the tree of theory completions.[24] However, it is not difficult to see that this proof strategy is not revealing any kind of deep connection between the theories. Rather, we are using a clever trick to connect the incompleteness of one theory to that of another. This does the job, but it is not telling us that these theories provide two different ways to talk about the same subject matter.

2.1.3. *Loss in translation.* If we are to think of our interlocutors as arguing past each other, then we want to have some sense that they are really talking about the same subject matter and the source of disagreement is merely a matter of semantics. If we're really using language differently to talk about the same content, then we should expect that if I translate my language into your language and then back again, I should end up saying exactly what I said in the first place. Or from the semantic perspective of Theorem 4, we should expect that if we start with a model of my theory transform and it into one of yours and then back into mine, I should end up exactly where I started. This does not occur with our example of mutual interpretability above. To see this, let

$$i : mod(ZFC + V = L) \leftrightarrow mod(ZFC) : t$$

be the mod-functors determined by the interpretations described in Example 7. Now suppose that $\mathcal{M}$ is a countable model of $ZFC$ and let $\mathcal{M}[c]$ be a generic extension of $\mathcal{M}$ by a Cohen real.[25] Then $\mathcal{M}[c]$ satisfies $ZFC$ and that $V \neq L$. Then note that $i \circ t(\mathcal{M}[c])$ is just $\mathcal{M}[c]$'s version of $L$ and so it satisfies $V = L$. Thus $i \circ t(\mathcal{M}[c]) \neq \mathcal{M}[c]$.

Thus, the mutual interpretations do not return us to exactly where we started: information has been lost in translation and it cannot be recovered. In particular, if we have a model of $ZFC$ with non-constructible sets, then once those sets are gone we have no way to get them back through interpretation. Thus, it seems wrong to think of a dispute over whether to include $V = L$ is one where the participants are merely arguing past each other. We need a stronger relation.

2.2. **Too strong.** We now have number of failed approaches to modeling a debate between users of alternative foundations that are practically identical. We started from mere interpretation and worked our way up to mutual faithful interpretation. However, in each case we found reason to think that the fit between the equivalence relation and the idea of merely verbal dispute was flawed. In this section, we are going to start from the other end of the spectrum and work our way down. In particular, we are going to begin with the strongest natural equivalence relation on theories aside from identity. We start from the semantic perspective.

**Definition 13.** Let $T$ and $S$ be theories articulated in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. We say that $T$ and $S$ are *definitionally equivalent* if they are mutually interpretable as witnessed by mod-functors

$$t : mod(T) \leftrightarrow mod(S) : s$$

in such a way that:

[24] See [Meadows, 2022] for examples and generalization of this kind of argument.
[25] The reader will lose little if they think of $\mathcal{M}$ being a countable transitive model. See the appendices of [Maddy and Meadows, 2020] for discussion of how to force over possibly ill-founded models.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

(1) whenever $\mathcal{M} \models T$, then $\mathcal{M} = s \circ t(\mathcal{M})$; and
(2) whenever $\mathcal{N} \models S$, then $\mathcal{N} = t \circ s(\mathcal{N})$.

Intuitively speaking, this tells us that if we start with a model of $T$ and then transform it into a model of $S$ and back again using $t$ and $s$, then we end up exactly where started again. Similarly, if we start with a model of $S$. Thus, these translations do not suffer any of the information loss we discussed in Section 2.1.3.

For a trivial example, let $sLO$ be the theory of stricty linear orders using the $<$ reation and let $nsLO$ be the theory of non-strict linear orders using the $\leq$ relation. We then translate $x < y$ as $x \leq y$ and $x \neq y$; and in the other direction, we translate $x \leq y$ as $x < y$ or $x = y$. We make no alterations to the domains.[26] It is then easy to see that if we start with a model $\mathcal{A}$ of $sLO$ and use these translations to make a model of $nsLO$ and then a model of $sLO$ again, we get back to $\mathcal{A}$. Similarly, if we start with a model of $nsLO$.[27] If we imagined a debate between an $sLO$ user and an $nsLO$ user, then intuitively we'd probably think of this as being a pointless debate. The fact that these theories are definitionally equivalent reveals why. It gives us a precise way of understanding the way in which the parties to such a debate would be merely arguing past each other.

For a more interesting example, recall Aczel's proposal for a theory of sets that intentionally admits ill-founded sets. The theory is formulated by removing Foundation and replacing it with Aczel's anti-foundation axiom.[28] Let us call this $AZFC$. This theory had been proposed as a means of addressing an apparent weakness of ordinary $ZFC$: it's inability to deal with an ill-founded membership relation. For example, according to $ZFC$ there can be no set $x$ such that $x$ is itself its only member. $AFZC$ on the other hand, proves that such sets exists. Nonetheless, we have the following result.

**Theorem 14.** *[Visser and Friedman, 2014] $ZFC$ and $AZFC$ are definitionally equivalent.*

If we brought this to bear on our debate between alternative theories, then this result shows that $ZFC$ and $AZFC$ can be understood as different ways of talking about the same subject matter. If we imagined a debate between the $ZFC$ user and the $AZFC$ user, then this result tells that there is not so much at stake. Or perhaps better, while we haven't ruled out the possibility that the $ZFC$ and $AZFC$ user are talking about two different abstract structures, it is clear that $ZFC$ and $AZFC$ could be used to talk about the same abstract structure. This doesn't mean that there cannot still be some kind of disagreement. In particular, the $AZFC$ user might claim that – although the $ZFC$ user can simulate them perfectly – $AZFC$ is simply a more natural tool for the investigation of ill-founded structures. Such a debate seems comparable to debates between programming languages when all such languages are Turing complete. There's something to debate, but it's on a very different level to the discussion in this paper.

[26] Indeed, definitional equivalence cannot permit an interpretation that restricts the domain to a proper subset.

[27] Note that the translations do not take us forth and back to the same formula. For example, we get $x < y$ ends up being translated back to $(x < y \vee x = y) \wedge x \neq y$. Nonetheless, these formulae are equivalent over $sLO$.

[28] The anti-foundation axiom says that for an $\mathcal{A} = \langle A, R \rangle$ where $R \subseteq A^2$ there is a unique surjective, near-embedding $f : \mathcal{A} \to \langle X, \in \rangle$ where $X$ is transitive. Here a near-embedding is a homomorphism such that for all $x, y \in A$, $xRy \leftrightarrow f(x) \in f(y)$; the $\leftarrow$ direction takes us beyond an ordinary homomorphism. This definition is described in Theorem 1.6 of [Forti and Honsell, 1983]. Alternative formulations and further discussion can be found in [Aczel, 1988] and [Barwise and Moss, 1996]. See [Maddy, 1997] for philosophical discussion of the place of Aczel's set theory in the foundations of mathematics.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

While I believe the semantic perspective on definitional equivalence gives us a very concrete way of understanding what is happening, we might worry that it relies too much on toy models and so we might only be able to make use of this relationship from a – so to speak – external perspective. It turns out that an internal approach can be offered which is quite similar in spirit to Theorem 2. To motivate this, first recall the given a theory $T$ articulated in $\mathcal{L}_T$, we may form a *simple definitional expansion* $T^*$ of $T$ by adding a new relation symbol $P$ to $\mathcal{L}_S$ and an axiom of the form

$$\forall \bar{x}(P\bar{x} \leftrightarrow \varphi(\bar{x}))$$

where $\varphi(\bar{x})$ is a formula of $\mathcal{L}_T$.[29] In effect, we are defining the extension of $P$ using the formula $\varphi(\bar{x})$. This means that the new symbol, $P$, is redundant in the sense that it can always be replaced by $\varphi(\bar{x})$ wherever it occurs.[30] Let us say that a *definitional expansion* of $T$ is formed by a series of simple definitional expansions. We then have the following result linking definitional equivalence with a more internal perspective.

**Fact 15.** *Let $S$ and $T$ be theories articulated in $\mathcal{L}_S$ and $\mathcal{L}_T$ respectively where $\mathcal{L}_S$ and $\mathcal{L}_T$ share no vocabulary.*[31] *Then the following are equivalent:*

(1) *$S$ and $T$ are definitionally equivalent; and*
(2) *There are definitional expansions $S^*$ and $T^*$ of $S$ and $T$ respectively into $\mathcal{L}_S \cup \mathcal{L}_T$ such that $T^*$ and $S^*$ are logically equivalent (i.e., they are the same theory).*

Informally speaking, this tells us that $S$ and $T$ can each be definitionally expanded to the same language $\mathcal{L}_S \cup \mathcal{L}_T$ to give the same theory. Let's apply this to our comparison between $ZFC$ and $AZFC$ by supposing that you think we should use $ZFC$ while I think we should use $AZFC$. Both theories are articulated in $\mathcal{L}_\in$, but say different things about the membership relation. Following the same course as we did in relation to Väänänen's theorem, it seems appropriate to move to a language with two membership relations, $\mathcal{L}_{\in,\in^*}$. We let $ZFC$ use $\in$ and $AZFC$ us $\in^*$. Then the you can use $ZFC$ in $\mathcal{L}_{\in,\in^*}$ to provide a definition for $\in^*$. And similarly the I can use $AZFC$ in $\mathcal{L}_{\in,\in^*}$ to provide a definition for $\in$. The resultant theories then have exactly the same consequences. As in our discussion of Väänänen's theorem, we move to a shared language and obtain a kind of collapse between the theories in question.

Definitional equivalence addresses all the problems of the previous section and gives us a plausible way to model disputes where participants are arguing past each other. But how well does it work on our original project to compare alternative extensions of $ZFC$? Very interestingly, we see something like categoricity rears its head once more.

[29] Constant symbols and functions symbols can also be added by a definition, however, the underlying theory must establish the required uniqueness for such objects.
[30] Of course, this doesn't mean that a definitional expansion is without value. For example, the practical use of $ZFC$ is littered with an array of defined vocabulary without which it would likely be unusable by humans.
[31] This annoying technical restriction is required for the equivalence but it is also easy to accommodate in our philosophical story. We are simply taking care to distinguish our language from that of our interlocutor.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Theorem 16.** *[Enayat, 2016] If $S$ and $T$ are theories extending $ZFC$ that are definitionally equivalent, then $S$ and $T$ are the same theory.*[32]

*Proof.* Suppose that we have mod-functors $t : mod(T) \leftrightarrow mod(S) : s$ witnessing the definitional equivalence of $T$ and $S$. We claim that $T$ and $S$ have the same consequences. To see this, let $T^*$ be a definitional expansion of $T$ into $\mathcal{L}_{\in,\in^*}$ by adding the axiom

$$\forall x \forall y (x \in^* y \leftrightarrow \varepsilon_t(x, y))$$

defining $\in^*$ using $\varepsilon_t$. Then note that since $t$ and $s$ witness definitional equivalence we also see that $T^*$ implies that $\in$ can be defined in terms of $\in^*$; i.e., we have

$$\forall x \forall y (x \in y \leftrightarrow \varepsilon_s(x, y)^{\in^*})$$

where $\varepsilon_s(x, y)^{\in^*}$is the same formula as $\varepsilon_s(x, y)$ except that instances of $\in$ have been replaced by $\in^*$. Given that $T$ and $S$ both extend $ZFC$, this means that $T^*$ is an extension of $ZFC(\in, \in^*)$ and so by Väänänen's theorem $\in$ and $\in^*$ are always isomorphic and so $T$ and $S$ must have the same logical consequences. □

This seems like bad news. Our goal has been to develop a precise means of modeling disputes between adherents of distinct set theories extending $ZFC$ in such a way that we could think of the parties as arguing past each other. We'd hoped to translate the disagreement away and having investigated a number of equivalence relations that were too weak, we used this discussion to motivate the relation of definitional equivalence. This provided the, so to speak, gold standard of equivalence through which no information is lost. However, in the case of set theories extending $ZFC$, we see that definitional equivalence is trivial in that it reduces straight back to identity. For some, this could be the end of the road. An argument very like the categority argument we started with also appears to block the use of interpretability for our project. This much is certainly true: we'll need to lower our standards if we are to retain our goal of adequately modeling such disputes as instances of arguing past each other.

2.2.1. *The interpretability hierarchy.* Fortunately, there are a number of lower standards that remain philosophically appealing. So with the problems of the previous section in mind, we now introduce a series of natural weakenings of definitional equivalence and investigate their prospects for deflating debates between strong set theories.

**Definition 17.** Let $T$ and $S$ be theories articulated in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. Suppose that $T$ and $S$ are mutually interpretable as witnessed by mod-functors

$$t : mod(T) \leftrightarrow mod(S) : s.$$

We say that $t$ and $s$ witness that $T$ and $S$ are:

- *Bi-intepretable* if there are definable functions $f$ and $g$ over $T$ and $S$:
  (1) whenever $\mathcal{M} \models T$, then $f^{\mathcal{M}} : \mathcal{M} \cong s \circ t(\mathcal{M})$; and
  (2) whenever $\mathcal{N} \models S$, then $g^{\mathcal{M}} : \mathcal{N} \cong t \circ s(\mathcal{N})$.

[32] This is actually a weakening of Enayat's theorem, but this version fits better with our current discussion. A stronger version of the theorem will be discussed later. We also note that a similar and sharper observation about the relationship between internal categoricity and bi-intepretablity is made in [Freire and Hamkins, 2020].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

- *Iso-congruent* if:
  (1) whenever $\mathcal{M} \models T$, then $\mathcal{M} \cong s \circ t(\mathcal{M})$; and
  (2) whenever $\mathcal{N} \models S$, then $\mathcal{N} \cong t \circ s(\mathcal{N})$.
- *Sententially equivalent* if:
  (1) whenever $\mathcal{M} \models T$, then $\mathcal{M} \equiv s \circ t(\mathcal{M})$; and
  (2) whenever $\mathcal{N} \models S$, then $\mathcal{N} \equiv t \circ s(\mathcal{N})$.

So less formally, we see that definitional equivalence takes us back and forth to exactly where we started. Sentential equivalence takes us back and forth to a model with the same complete theory. With isocongruence we return to an isomorphic structure. And with bi-interpretability, we return to an isomorphic structure and we can actually define the isomorphism. Taking little stock we see that we have the following chain of implications among the equivalence relations we've discussed so far:

> Definitional equivalence → Bi-interpretability → Iso-congruence → Sentential equivalence → Mutual faithful interpretability → Mutual interpretability → Equiconsistency.

I don't believe it's known whether the first two arrows reverse, however the rest of the sequence is a strict hierarchy.[33]

We've seen some reason to think that definitionally equivalent theories just provide two different ways of talking about the same thing. But given that the new relations are weaker we might wonder whether they could adequately model disputes about alternative set theories. We can provide some philosophical motivations for using these relations for this purpose. For example, suppose that we adhere to some form of structuralism about mathematics. There are many ways to formulate such positions, however, at their core is the maxim that we should not be concerned with what a structure is made from, we should only be concerned with the structure itself.[34] More succinctly, we should only care about a mathematical structure up to isomorphism. Thus in general, given two isomorphic structures, we think there is no good reason to prefer on over the other. Then both bi-interpretability and iso-congruence seem to provide a good fit with this position. Both of these relations tells us that if we go back and forth we end up with a model that is – by structuralist lights – just as good as the one we started with. Bi-interpretability provides an extra feature in that we are – so to speak – able to see the isomorphism, while this is not required for mere isocongruence. We might say that bi-interpretability gives an *internal* perspective to the structuralism involved, while mere iso-congruence provides an *external* point of view.[35] With respect to sentential equivalence we might turn to some species of formalism in which the enterprise of mathematics is understood to be – at its root – a syntactic process of proving theorems from axioms. Our talk about infinite mathematical structures is to be thought of as a fiction or heuristic crutch that support the real work that is fundamentally

[33] It should be noted that if we allow a more liberal definition of bi-interpretability that allows us to treat definable equivalence classes as objects, then there is a pair of theories that are bi-interpretable but not definitionally equivalent. The first example of this was provided by Visser and Friedman [2014], however, a much more natural example can be found in Theorem 16 of [Enayat and Łełyk, 2024]. Moreover, if we restrict our attention to countable models, then there is a pair of theories that are iso-congruent but not bi-interpretable. See [Meadows, 2023b] for more details.

[34] There is an extensive literature on structuralism in philosophy of mathematics. However, for our purposes the modest attitude based approach on [Hamkins, 2020] is sufficient and very general.

[35] This distinction between internal and external perspective is motivated by a similar distinction in epistemology. See Section 3 of [Balaguer, 1995] for a related discussion. It is not clear to me whether or how this relates to the distinction between internal and external perspectives with regard to categoricity.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

syntactic. From this perspective, our talk of moving back and forth between models is merely a congenial way of getting at the important fact, the back and forth move preserves the real syntactic information since it preserves the complete theory.

Returning to our motivating problem, we must now ask whether going lower in the hierarchy of equivalence relations helps. The first step down is disappointing. Visser and Friedman have shown that for theories that are able to provide a reasonable foundation for mathematics, if they are bi-interpretable, then they are definitionally equivalent. Recall that a *sequential theory* is a theory that interprets a very weak set theory that is intended to capture the minimal requirements for coding sequences and thus, syntax.[36]

**Theorem 18.** *[Visser and Friedman, 2014] If $T$ and $S$ are bi-interpretable theories that are sequential, then $T$ and $S$ are definitionally equivalent.*[37]

The proof takes a few steps, however, a crucial move is to take the internally defined isomorphisms to deliver an internal version of the Cantor-Bernstein theorem which allows us to turn the injections into a bijection. It is then easy to see that $ZFC$ and its extensions are sequential theories and so we have the following result.[38]

**Corollary 19.** *[Enayat, 2016] If $S$ and $T$ are bi-interpretable extensions of $ZFC$, then $S$ and $T$ are the same theory.*[39]

Thus, it seems that our internal structuralist perspective cannot support the possibility of disagreement between adherents of mutually inconsistent extensions of $ZFC$. This could be another stopping off point for some readers. It tells us that the only times we can translate away apparent disagreement by translating back and forth to an isomorphic structure – where that isomorphism is visible – the disagreement was illusory in the first place.

2.3. **The middle ground.** We appear to be running out of room in our hierarchy with only iso-congruence and sentential equivalence remaining. These are relative newcomers and as such, less is known about them. Nonetheless, the *space* between between iso-congruence and bi-interpretability seems very small. The only example distinguishing them that I know relies on restricting our attention to countable models to ensure certain isomorphisms exist but remain undefinable. More work is

[36] More specifically, a sequential theory must interpret *Adjunctive Set Theory* via a direct interpretation. Adjunctive set theory consists of two axioms: first, there is an empty set; and second, for any pair of sets $x$ and $y$ there is some $z$ containing exactly $y$ and the members of $x$. A direct interpretation is one that preserves identity and doesn't restrict the domain. All interpretations considered in this paper preserve identity. I should also mention that in fact, Visser and Friedman really show that we only need one of the theories to be able to interpret a somewhat weaker class theory. See [Visser and Friedman, 2014] for more details.

[37] It's important to note that the interpretations witnessing bi-interpretability must preserve identity. In the context of htis paper, all interpretations are identity preserving: see the first clause of Definition 3. In the context of $ZFC$ this doesn't matter too much, but it can be very important in weaker theories. For a significant example, see Theorem 16 of [Enayat and Łełyk, 2024].

[38] This kind of collapse result appears to be peculiar to strong theories like $ZF$ and $ZFC$. For an excellent account of this phenomenon in $ZF$ and related theories see [Enayat and Łełyk, 2024].

[39] This is an easy consequence of Theorem 16, Theorem 18 and the fact that $ZFC$ is Morita complete. In particular, in $ZFC$ we can represent definable equivalence classes with definable objects using Scott's trick. In model theory such theories are said to *eliminate imaginaries*: see Section 4.4 [Hodges, 1997] for more details. I thank Ali Enayat for drawing my attention to this.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

required here, but given this we shall also leave iso-congruence aside. Fortunately, it is possible to separate sentential equivalence from iso-congruence. This also gives us a good example of sententially equivalent extensions of $ZFC$.

**Theorem 20.** *[Meadows, 2022] Let $ZFC + V = L[c]$ be the theory consisting of $ZFC$ augmented by the statement that the universe has been constructed from a Cohen real over $L$. Then $ZFC + V = L$ and $ZFC + V = L[c]$ are sententially equivalent but they are not isocongruent.*

Morally speaking, it's not difficult to see why these theories aren't isocongruent. If we start with a transitive model of $ZFC+V = L$ and define a model of $ZFC+V = L[c]$ it will want to be well-founded have the same order type for its ordinals if it has any hope of defining a model of $ZFC = V = L$ that is isomorphic to the original. But then it can be collapsed and so – with a little work – we end up having a Cohen real in the original model, which is impossible.

To show sentential equivalence, the obvious thing to do here is take a countable model $M$ of $ZFC+V = L$ and then add a Cohen real to obtain a model of $ZFC+V = L[c]$. And in the other direction taking a model $N$ of $ZFC+V = L[c]$, we can simply define $N$'s version of $L$ to obtain a model of $ZFC+V = L$. The problem with this strategy is that the first step isn't a genuine interpretation. Interpretations always return models whose domain is a definable subset of the original model. Generic extensions always return models that are proper supersets of the original model. This can be addressed by using a Boolean valued ultrapower where we build a model using an ordinary ultrafilter over a complete Boolean algebra rather than a generic ultrafilter.[40] This means that the Boolean ultrapower can be defined within the ground model, so we are able to define an internal model of $ZFC + V = L[c]$ within any model of $ZFC + V = L$.[41] Exploiting the homogeneity of the forcing used to obtain Cohen reals, it can then be shown that regardless of which ultrafilter we use, the resultant model will have the same complete theory. Putting this together, we can show that the respective back and forth transitions between models of these theories always give us a model that is elementary equivalent to the one we started with.

This gives us our best prospect, so far, of an equivalence relation modeling disputants arguing past each other. It doesn't appear to be practically trivial like mutual interpretability or its faithful improvement; and it doesn't collapse into the triviality of identity like definitional equivalence and bi-interpretability. If you believed that every set is constructible and I believed that every set is constructible from a Cohen real over $L$, then we might use the interpretations above to translate this disagreement away by noting that the back and forth interpretations preserve the complete theory of the original models. I think many set theorists would agree that sententially equivalent theories like those above are close enough as foundations as to be practically the same. Anything you can do in one theory you can do in the other, via the translations. There is also a syntactic rendering of sentential equivalence.

**Proposition 21.** *$T$ and $S$ are sententially equivalent iff there exist translations $t : \mathcal{L}_S \to \mathcal{L}_T$ and $s : \mathcal{L}_T \to \mathcal{L}_S$ witnessing mutual interpretability such that:*

(1) *$T \vdash \varphi \leftrightarrow t \circ s(\varphi)$, for all sentences $\varphi$ of $\mathcal{L}_T$; and*

[40] See [Hamkins and Seabold, 2012] for a detailed account of Boolean valued ultrapowers. Another way of doing this is to define a notion of a *generic interpretation*. This is explored in [Meadows, 2022].

[41] See [Hamkins and Seabold, 2012] or [Meadows, 2023a] for more details about Boolean valued ultrapowers.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

(2) $S \vdash \psi \leftrightarrow s \circ t(\psi)$, *for all sentences* $\psi$ *of* $\mathcal{L}_S$.

Here we see that the translation functions return us sentence-by-sentence to exactly where we started according to the theory in question. We might see this as offering a more *internal* perspective on sentential equivalence. If you are a $S$ user and I am a $T$ user, then these translations reveal that I can translate any sentence of my theory into yours and then return to obtain a sentence that arguably has the same meaning.[42]

This looks like good news, however, we should also take into account what is lost in the move below iso-congruence and bi-intepretability. In general, the back and forth translations do not return us to a structure that is isomorphic to the one we started with: we only get elementary equivalence. As such, we might wonder whether this relationship is strong enough to satisfy the mathematical structuralist. I think the structuralist can reasonably complain that information has been lost in translation. For example, when forcing is involved our use of Boolean valued ultrapowers means that the structure we end up with is typically ill-founded from the perspective of the original model. If we like transitive models, then the model we end up with seems obviously pathological. Nonetheless, if we consider the example above, I think it's not so surprising that this kind of information loss occurs. In a nutshell, the information contained in ultra or generic filters is, so to speak, too random to be recovered by a definition or interpretation. Essentially, this is because generics and ultrafilters are only obtained by choice principles. The consequence of this is that when we use an interpretation that forgets a generic set – as in the example above – there is simply no way of defining it back. Given this, if we want forcing arguments to be in our toolkit for theory comparison, it seems that isocongruence is too much to demand and that sentential equivalence provides a reasonable way of modeling the idea that distinct theories are saying the same thing.

2.3.1. *Pointwise equivalence.* We've now finally found in sentential equivalence an equivalence relation that avoids the practical triviality of mutual interpretation and the collapse into identity of bi-interpretability. I think it is a good candidate for theoretical equivalence between strong set theories. Nonetheless, it also suffers a practical limitation in that examples of sententially equivalent theories extending $ZFC$, tend to make use of very artificial theories. For example, while $ZFC + V = L$ is a natural albeit restrictive theory, the theory $ZFC + V = L[c]$ is not commonly used and certainly never proposed as a serious foundation for mathematics. Indeed, there is a sense in which this theory was cooked to provide a simple pair of sententially equivalent theories.[43] It would be better if we had an equivalence relation between theories that worked between theories more commonly used by set theorists. More particularly, it would be good to identify a non-trivial equivalence relation that was able to enjoin theories of the form $ZFC + \varphi$ and $ZFC + \neg\varphi$. In this little section, we'll consider a significant weakening of sentential equivalence that will provide many more examples of equivalent theories. We'll then discuss an equivalence relation from contemporary set theory that has a similar

[42] It is, however, worth noting that in contrast to the internal characterization of definitional equivalence (as in Fact 15) we are using two different theory contexts $T$ and $S$ rather than one. We shall see that the ability to provide satisfying internal syntactic counterparts to our external semantic equivalence seems to become more difficult the more we weaken the equivalences.

[43] A more natural example of sentential equivalence occurs between an extension of $ZFC$ and Steel's multiverse theory $MV$. See Theorem 32 and [Meadows, 2021] for more details. However, this is not an equivalence between extensions of $ZFC$ since $MV$ does not (globally) satisfy the powerset axiom.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

extension: the generic multiverse. As this is the first place in this paper where forcing enters in a very general fashion, I should make a quick remark about nomenclature. Whenever I talk about forcing or generic extension, I shall mean that the underlying partial order *is a set* unless I state otherwise.

Our plan is to weaken sentential equivalence in two stages. For the first stage, we remove demand that we always have to use the same translations to get between a model of one theory and another. Rather, we may vary which translation we use in response to the model we are considering.

**Definition 22.** Let $T$ and $S$ be theories extending $ZFC$. We say that $T$ and $S$ are *pointwise sententially equivalent* if:

- For all $\mathcal{M} \models T$ there is a model $\mathcal{N}$ of $S$ and mod-functors $t, s$ such that

$$t(\mathcal{M}) \equiv \mathcal{N} \;\&\; s(\mathcal{N}) \equiv \mathcal{M}.$$

- For all $\mathcal{N} \models S$ there is a model $\mathcal{M}$ of $T$ and mod-functors $s, t$ such that

$$s(\mathcal{N}) \equiv \mathcal{M} \;\&\; t(\mathcal{M}) \equiv \mathcal{N}.$$

The salient difference is that the translations are no longer deployed *uniformly* across the models of $T$ and $S$. We can use different translations from model to model. This opens up the playing field substantially.[44]

To throw a little philosophical light on this equivalence relation, we return to our motivating example of a debate about competing foundational theories. Suppose that you are using theory $S$ while I'm using theory $T$, where $T$ and $S$ are pointwise sententially equivalent theories extending $ZFC$. We see that any model of my theory $T$ can define a model of $S$ and that model can in turn define a model of $T$ that is elementary equivalent to the one we started with. This means that if you prove a theorem in $S$, there is a sense in which I can always access that by using the translation to move to a model of $S$ where that theorem is true and further there is a sense in which this model of $S$ is equivalent to the one I started with in the sense that it can define a model that is elementary equivalent to it. Similar remarks show that you, as an $S$ user, can access theorems of $T$ by moving to an equivalent model. Moreover, if some statement $\varphi$ is merely satisfiable according to $S$ then I know there is a model of my theory $T$ that can be defined in a model of $S$ where $\varphi$ is true and that is equivalent in the sense that this model of $T$ can define a model of $S$ that has the same complete theory as the original model of $S$. Again, similar remarks apply when we start from a sentence $\psi$ that is satisfiable according to $T$.

I think these remarks speak of a tight connection between pointwise sententially equivalent theories, however, unlike our previous examples there does not appear to be a syntactic or internal counterpart to the pointwise equivalence relation. Since the pointwise equivalence allows us to vary the interpretations used as we move between models, the talk of models becomes an essential part of the definition and as such, its philosophical interpretation. Given that we noted that sentential equivalence may be more appealing to those with formalist tendencies, this talk of models makes it more challenging to provide a satisfying philosophical account of who should find pointwise sentential equivalence to be sufficient for us to say that we have equivalent ways of talking about the same subject. It seems to be that the easier we make it to find equivalences, the more difficult it is to explain the underlying philosophy.

[44] We note that if $T$ and $S$ are pointwise sententially equivalent then they are equiconsistent with each other over a weak base theory. I do not know whether such $T$ and $S$ are mutually interpretable.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

The second stage of our weakening compounds this phenomenon. Given that the pointwise relation focuses on models, this opens up the possibility of using arbitrary parameters in the definitions of our interpretations. Then just as we allowed ourselves to vary the interpretation from model to model, we might also allow ourselves to vary the parameters used in those interpretations. To set this up, we first modify our interpretative mechanism to incorporate the use of parameters. Given that we are focused on models, we just give the generalized version of mod-functor.

**Definition 23.** Let us say that $t(\cdot). : \sum_{\mathcal{M} \in mod(T)} M \to mod(\emptyset)$ is a *parametric mod-functor* determined by the formulae $\delta_t(x, z)$ and $\varepsilon_t(x, y, z)$ if all ordered pairs of a model $\mathcal{M} = \langle M, \in_{\mathcal{M}} \rangle$ of $T$ and element $m \in M$, we have $t(\mathcal{M})_m = \langle N, \in_{\mathcal{N}} \rangle$ where:[45]

(1) $N = \{x \in M \mid \mathcal{M} \models \delta_t(x, m)\}$; and
(2) $\in_{\mathcal{N}} = \{\langle x, y \rangle \in M^2 \mid \mathcal{M} \models \varepsilon_t(x, y, m)$.

Thus, the mod-functor definition remains the same except that we now allow the use of a parameter, $m \in M$, to define the interpretations of domain and membership relation. With this in hand, we can provide our final weakening of sentential equivalence.

**Definition 24.** We say that $S$ and $T$ are *pointwise, parametrically sententially equivalent* if

- For all $\mathcal{M} \models T$ there exist $\mathcal{N} \models S$, parametric mod-functors $t, s$, and sets $m \in M$ and $n \in N$ such that
$$t(\mathcal{M})_m \equiv \mathcal{N} \;\&\; s(\mathcal{N})_n \equiv \mathcal{M}.$$
- For all $\mathcal{N} \models S$ there exist $\mathcal{M} \models T$, parametric mod-functors $s, t$, and sets $m \in M$ and $n \in N$ such that
$$s(\mathcal{N})_n \equiv \mathcal{M} \;\&\; t(\mathcal{M})_m \equiv \mathcal{N}.$$

Once again, we are allowed to vary the interpretation we use from model to model. However in addition to this, we are also allowed to pick parameters from the respective models to get the interpretations to do the required work. This opens up the playing field further and again complicates the philosophical story. While sentential equivalence seems more appealing to those with formalist tendencies, we're now not only talking about models but making essential use of parameters within them. Nonetheless, it is arguably at this level that we see a more practical level of intervention by obtaining equivalences via forcing of distinct natural theories extending $ZFC$. To demonstrate this we first recall that following very useful theorem.

**Fact 25.** *(Laver, 2007; Woodin) If $V$ is a generic extension of an inner model $M$, then $M$ can be defined in $V$ using a parameter.*[46]

Informally speaking, this tells us that if we are in a generic extension, then we can define – using parameters – any of the ground models from which our universe eventuated. This fact, in combination with Boolean valued ultrapowers, allows us to show a striking equivalence.

[45] Note that $mod(\emptyset)$ is intended to denote models of the empty theory; i.e., arbitrary models of $\mathcal{L}_\in$. Also note that we are abusing notation a little in the hopes of leaving less clutter on the page and making the relationship with ordinary mod-functors more obvious to the reader. In particular, while we should write $t(\langle \mathcal{M}, m \rangle)$, we are writing $t(\mathcal{M})_m$ instead. We hope this causes no confusion.

[46] Note that this theorem cannot be extended to work for class forcing: see Theorem 26 of [Antos, 2018].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Example 26.** Consider the theories $ZFC + CH$ and $ZFC + \neg CH$. We claim that they are parametrically, pointwise, sententially equivalent and provide a sketch of a proof. We start with a model $\mathcal{M}$ of $ZFC + CH$. As is well known, in any such model, there is a complete Boolean algebra $\mathbb{B}$ such that

$$(\Vdash_{\mathbb{B}} \neg CH)^{\mathcal{M}}.$$

Let $U$ be a ultrafilter over $\mathbb{B}$ from $\mathcal{M}$. Then the Boolean ultrapower $\mathcal{N}$ of $\mathcal{M}$ by $U$ can be defined in $\mathcal{M}$ using the parameter $U$. Then it can be seen that $\mathcal{N}$ has an inner model $\mathcal{M}^*$ that is a generic refinement of $\mathcal{N}$ and is elementary equivalent to $\mathcal{M}$. By Fact 25, $\mathcal{M}^*$ is definable using a parameter in $\mathcal{N}$. Thus we have moved from a model of $ZFC + CH$ to a model of $ZFC + \neg CH$ and back to a model of $ZFC + CH$ that is elementary equivalent to the one we started with.

Next we start with a model $\mathcal{M}$ of $ZFC + \neg CH$. Rather than going to $L^{\mathcal{M}}$, we may force to collapse all the cardinals below $2^{\aleph_0}$ to obtain a model of $ZFC + CH$. More precisely, there is a complete Boolean algebra $\mathbb{B}$ in any model $\mathcal{M}$ of $ZFC + \neg CH$ such that

$$(\Vdash_{\mathbb{B}} CH)^{\mathcal{M}}.$$

The rest of this case is then analogous to the first case: we use an ultrafilter to define a Boolean ultrapower and then return to an elementary equivalent model using Fact 25.

It's then easy to see that the argument above can be adapted to obtain the following result. Taking our lead from the example, let us say that a theory $T$ interprets $S$ *via forcing* if for all $\varphi \in S$, $T$ proves $\Vdash_{\mathbb{P}} \varphi$, where $\mathbb{P}$ is uniformly defined over $T$ by some formula.[47]

**Corollary 27.** *Any pair of theories that can interpret each other via forcing are parametrically, pointwise sententially equivalent.*[48]

This gives our best solution to the problem posed at the very beginning of this paper: is there a way of deflating a debate between users of $ZFC$ who disagree about whether some sentence $\varphi$ is true or not? We wondered whether we could offer translations that would reveal that they were arguing past each. Could they be understood as using different languages to say the same thing? We then saw that: mutual interpretation was too easy to obtain; bi-interpretability collapsed into identity; and sentential equivalence only worked in special cases. But with the generalization to pointwise comparison with parameters, the generalized notion of sentential equivalence gives us an equivalence relation that allows us to deflate one of the fundamental problems of set theory: the continuum hypothesis. If we think that $CH$ holds, then we can travel to and return from an equivalent world where it is false. Similarly if $\neg CH$ is true.

2.3.2. *The generic multiverse.* We finally have an equivalence relation that is serviceable for the comparison of theories extending $ZFC$. However, we've also noted some difficulties in arguing for the philosophical picture that motivates it. In an effort to alleviate this worry and to situate this relation in the context of contemporary set theory, we now highlight an intriguing relationship between

[47] In other words, $\mathbb{P}$ is defined over $T$ in the same way that constant symbols of one theory can be interpreted in another. As we noted above, this isn't, strictly speaking, an interpretation. Nonetheless, it can often be converted into one using the techniques described above.

[48] Recall that we are restricting our attention to set forcing here.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

pointwise parametric sentential equivalence and the generic multiverse. The latter notion has been investigated by Woodin and Steel in response to the incompleteness of $ZFC$ wrought by forcing [Woodin, 2012, Steel, 2014]. Very crudely, the generic multiverse responds to incompleteness by embracing it rather than attempting to remove it. Rather than attempting to solve problems shown to be be independent of $ZFC$ by forcing, we simply allow that there are set-concepts or worlds where such problems have affirmative solutions and worlds where they do not. The underlying idea is not dissimilar to the motivating problems of this paper. Indeed, I think its relationship with pointwise parametric sentential equivalence gives a clear perspective of a role that the generic multiverse can play in set theory.

First we define the generic multiverse. Our definition is a little different to that offered in [Woodin, 2012] since we only demand that our models are countable but not transitive. Countability is required to ensure that generic filters exist, but we avoid transitivity since our approach to forcing uses Boolean valued ultrapowers which are generally ill-founded. This helps us to get the relationship with generalized sentential equivalence, however, a similar relation was also proposed by Woodin in [Maddy and Meadows, 2020]. It will be helpful to say that a *generic refinement* is obtained when we employ Fact 25 to define a ground model of the universe. Roughly speaking it is the inverse of the operation that gives us generic extensions.

**Definition 28.** (Woodin) Let $\mathcal{M}$ be a countable model of $ZFC$. Let the *generic multiverse generated from* $\mathcal{M}$, denoted $\mathbb{V}_{\mathcal{M}}$, be the set of models $\mathcal{N}$ such that there is a sequence $\langle \mathcal{M}_0, ..., \mathcal{M}_n \rangle$ where $\mathcal{M} = \mathcal{M}_0$ and $\mathcal{N} = \mathcal{M}_n$ such that for all $i < n$, $\mathcal{M}_{i+1}$ is either a generic extension or a generic refinement of $\mathcal{M}_i$.

Thus, the generic multiverse of some $\mathcal{M}$ consists of all those models that can be obtained by a sequence of generic extensions and refinements starting from $\mathcal{M}$. We have a set of models closed under generic extension and its inverse. Using a result of [Usuba, 2017], it turns out that every such model can be obtained by a single generic refinement following by a generic extension.[49] Our next problem is to define a suitable equivalence relation using the generic multiverse.

**Definition 29.** Let us say that theories $T$ and $S$ extending $ZFC$ are *generic multiverse equivalent* if

$$\{\mathbb{V}_{\mathcal{M}} \mid \mathcal{M} \models T\} = \{\mathbb{V}_{\mathcal{N}} \mid \mathcal{M} \models S\}$$

where we restrict our attention to models that are hereditarily countable.

The essential idea here is to blur the models associated with a theory $T$ by considering their generic multiverses. Note that whenever we have

$$\{\mathcal{M} \mid \mathcal{M} \models T\} = \{\mathcal{N} \mid \mathcal{N} \models S\}$$

then $S$ and $T$ are the same theory. By using a collection of generically related models rather than a single model, we wipe away the generically superficial differences between those models. Thus, comparing these generic multiverses associated allows us to ignore the kinds of difference that are caused by forcing. If we don't think these differences are significant, then generic multiverse equivalence is a good way to track equivalences that ignore these differences. As with pointwise parametric sentential

[49] See [Maddy and Meadows, 2020] for more details. [Koellner, 2010] contains a similar result.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

equivalence, we see that $ZFC + CH$ and $ZFC + \neg CH$ are generic multiverse equivalent. Given a model $\mathcal{M}$ of $ZFC + CH$, there is a model $\mathcal{M}[G]$ of $ZFC + \neg CH$ in $\mathbb{V}_{\mathcal{M}}$; and given a model $\mathcal{N}$ of $ZFC + \neg CH$ there is a model $\mathcal{N}[H]$ of $ZFC + CH$ in $\mathbb{V}_{\mathcal{M}}$.

Given the pointwise parametric sentential equivalence also obtains these kinds of equivalences, this raises questions about its relationship with generic multiverse equivalence. Our goal now is to show that – under certain restrictions – generic multiverse equivalence implies pointwise parametric sentential equivalence. The restrictions are necessary since generic multiverse equivalence uses genuine generic extension for forcing arguments while pointwise parametric sentential equivalence uses Boolean valued ultrapowers. It can be seen that there are complete theories given by generic filters that cannot be matched by ultrapower constructions. A relatively simple restriction is to restrict our attention to theories extending $ZFC$ by $\Sigma_2$ statements. Essentially, a $\Sigma_2$ statements says that something holds in an initial segment of the cumulative hierarchy.[50] A typical example is the statement that a measurable cardinal exists. The following fact then underpins our claim.

**Fact 30.** *(Woodin) Suppose there are unboundedly many Woodin cardinal and that $\varphi$ is a $\Sigma_2$ sentence. Then whenever $\mathbb{P} \in V$ is such that $\Vdash_{\mathbb{P}} \varphi$ and $\mathbb{Q}$ is a poset, there is some $\dot{\mathbb{S}} \in V^{\mathbb{Q}}$ such that:*

$$\Vdash_{\mathbb{Q} * \dot{\mathbb{S}}} \varphi.$$

Informally speaking, this tells that whenever a $\Sigma_2$ sentence $\varphi$ can be forced to be true but is then forced to be false, it can still be forced to be true again. For this reason, it is often known as the $\Sigma_2$-resurrection theorem. We can then state our claim and sketch its proof.

**Theorem 31.** *[Meadows, 2023c] Suppose $T$ and $S$ are extensions of $ZFC$ plus there are unboundedly many Woodin cardinals by $\Sigma_2$ statements. Then if $T$ and $S$ are generic multiverse equivalent, then they are also parametrically, pointwise sententially equivalent.*

*Proof.* (Sketch) Suppose $T$ and $S$ are as above and that they are generically multiverse equivalent. Let $ZFC^*$ denote $ZFC$ augmented with a proper class of Woodin cardinals. Then let $T$ be $ZFC^* + \Phi$ where $\Phi$ is $\Sigma_2$; and $S$ be $ZFC^* + \Psi$ where $\Psi$ is also $\Sigma_2$. Now let $\mathcal{M}$ be a model of $T$. By symmetry of argument and results from above, it will suffice to show that there is a generic extension of $\mathcal{M}$ in which $S$ holds. Since $T$ and $S$ are generic multiverse equivalent, there must be a generic refinement $\mathcal{M}^*$ of $\mathcal{M}$ with a generic extension $\mathcal{N}$ satisfying $S$. But then using Fact 30, we see that $\mathcal{M}$ must also have a generic extension satisfying $S$ as required. □

Thus, we see that parametric, pointwise sentential equivalence provides us with a non-trivial lower bound below generic multiverse equivalence. This speaks to both to the naturalness of generic multiverse equivalence and the value of the framework we've been developing by being able to capture it. But is there some reason to prefer one of the other? In defense of parametric, pointwise sentential equivalence, we have in it an equivalence relation that emerges through a natural sequence of generalizations based on a linguistic translation: it's just definability with parameters! By contrast generic multiverse equivalence essentially relies on forcing and generic extension. These techniques are peculiar to set theory and as such, one might worry about that the resulting equivalence relation is too parochial

[50] We'll discuss $\Sigma_2$-statements in more detail in Section 4.1 where we also observe their relationship with satisfiability in second order logic.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

to support much philosophical significance. Moreover, the paucity of generic sets means that we must restrict our attention to toy models that are countable. No such restriction is required with parametric, pointwise sentential equivalence. On the other hand, John Steel offers a defense of generic multiverse equivalence based on the quality of models it delivers in comparison to those offered by sentential equivalence relations [Steel, 202?].[51] As we observed above, in order to simulate the effect of forcing arguments, we must use Boolean valued ultrapowers in parametric, pointwise sentential equivalence. In general, the effect of this is that the models produced by such ultrapowers will be ill-founded. Steel says that such interpretations fail to *preserve meaning*. I am unsure how to offer a complete theory of meaning here, but we can illustrate the effect by considering how the meaning of the term "ordinal" changes as we move from some ground model to an ill-founded Boolean ultrapower. While "ordinals" are well-founded in the ground model, in general, they are not in the ultrapower. We might say that the meaning of the term "ordinal" has not been preserved in this translation. By contrast, when we use generic extensions, we obtain a model with exactly the same ordinals as those of the ground model. Thus, the term "ordinal" and, indeed, much more is preserved in the move to a generic extension. For this reason, we may come to prefer generic multiverse equivalence. We leave any final adjudication on this matter to the reader, but note that the close relationship illustrated above is remarkable and arguably casts them them both in a better light together.

This concludes our discussion of using "talking past each other" as a means of resolving debates about extensions of $ZFC$. Our goal was to find a plausible way of modeling the idea that some mutually inconsistent set theories could be considered as providing merely different ways of saying essentially the same thing. The hope then was that the disagreement could be dissolved in translation. Having discarded a number of equivalence relations that were too weak or strong, we then slowly worked our way into a Goldilocks zone where equivalence wasn't too easy to obtain and yet was applicable to contemporary set theory. In pointwise, parametric sentential agreement we found a plausible alignment with contemporary practice while – at the same time – making the philosophical story about the naturalness of this relation more challenging.

## 3. The Common Ground

Having explored the idea of merely verbal dispute in set theory, I now want to investigate a distinct but closely related reason to not take all debates in set theory seriously. Perhaps there is *enough common* ground between competing set theories that the disagreements lying outside that region can be largely ignored. This idea has obvious parallels in physics where people often talk about empirically equivalent theories. Such theories have empirical information as their common ground and as such, for the empirical purposes of physics any theory will be as good as any empirically equivalent alternative.[52] Similarly with regard to the foundations of mathematics, we might imagine only caring about, say,

[51] Note that Steel's comments in that paper actually relate to the following result:

**Theorem 32.** *[Meadows, 2021] Let $MV^*$ be Steel's multiverse theory $MV$ augmented by the statement that the multiverse has a core that satisfies $V = HOD$. Then $ZFC + V = HOD +$ the ground axiom is sententially equivalent to $MV^*$.*

The proof of this also makes use of Boolean ultrapowers, so I think Steel's responses also apply in the current context. For a precise description of $MV$ and the rationale behind its axioms see [Maddy and Meadows, 2020] and [Steel, 2014]. The ground axiom is discussed in detail in [Reitz, 2007].

[52] A classic example of this kind is the relationship between Lagrangian and Hamiltonian mechanics. See [Barrett, 2019] for a detailed discussion.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

how some set theory deals with analysis. We might see the rest of the theory as providing a kind of ancillary scaffolding that facilitates a more complete theory of analysis, but not care so much about that scaffolding itself. Perhaps we shouldn't care about what's happening in the heavens so much as what is happening here and now closer to the ground.

Our project in the current section is to generalize the tools of relative interpretation to a theory of common ground between theories. We call this *partial interpretation*. We outline some of the basic theory here and illustrate some its limitations. Then in the following section, we shall apply these tools to develop a better understanding how much common ground we can expect to find between competing set theories. In contrast to our work on arguing past each other, we shall be working our way up to stronger equivalences rather than trying to weaken relations that are too strong for practical application.

3.1. **Partial equivalence.** Let's start with a plausible example of something like *empirical* equivalence in set theory. We'll then use this as our guide to providing a general theory of partial equivalence using interpretability. After that, we'll apply our framework to some other examples and then show that it has some disappointing limitations. Our prototypical example is well-known, however, it will be useful to sketch a quick proof of it since it provides the basis for our investigation. Let us say that $T$ interprets $S$ *using an inner model* if the mod-functor $t : mod(T) \to mod(S)$ is such that for all models $\mathcal{M}$ of $T$, $t(\mathcal{M})$ is an *inner model* of $\mathcal{M}$; i.e., $t(\mathcal{M})$ is a model of $S$ with the same ordinals as $\mathcal{M}$.

**Proposition 33.** *Let $T$ and $S$ be theories extending $ZFC$ and suppose that $T$ interprets $S$ using an inner model while $S$ interprets $T$ via forcing. Then $T$ and $S$ agree on $\Pi^1_2$ sentences; i.e., for all $\Pi^1_2$ sentences $\varphi$,*

$$T \vdash \varphi \;\Leftrightarrow\; S \vdash \varphi.$$

*Proof.* Let $\varphi$ be a $\Pi^1_2$ sentence. ($\Leftarrow$) Suppose that $T \nvdash \varphi$. Then fix a model $\mathcal{M}$ of $T$ where $\mathcal{M} \models \neg\varphi$. Since $T$ interprets $S$ via an inner model, $\mathcal{M}$ defines an inner model $N$ of $\mathcal{M}$ such that $N \models S$. Now $N$ and $\mathcal{M}$ share the same ordinals, so the Levy-Schoenfield theorem[53] tells us that $N \models \neg\varphi$. Thus, $S \nvdash \varphi$. ($\Rightarrow$) Similarly, if $S \nvdash \varphi$, we may fix a model $\mathcal{N}$ satisfying $S \cup \{\neg\varphi\}$. Without loss of generality, we suppose $\mathcal{N}$ is countable and so by our assumption we may fix an $\mathcal{N}$-generic $g$ such that $\mathcal{N}[g] \models T$. Again, $\mathcal{N}$ and $\mathcal{N}[g]$ share the same ordinals, so we have $\mathcal{N}[g] \models \neg\varphi$ and thus, $T \nvdash \varphi$ as required. □

The proof above is driven by the fact that we are using good interpretations in the context of theories that are strong enough to ensure that $\Pi^1_2$ sentences are preserved. The classic example here is $ZFC + CH$ and $ZFC + \neg CH$. Traditionally, a model of $ZFC + CH$ can be obtained by using the interpretation that restricts us to $L$; and a model of $ZFC + \neg CH$ can be obtained by forcing to add $\aleph_2$ many reals. Informally speaking, this tells us that if we are happy – as the mathematical community generally is – to use $ZFC$ in our mathematical practice, then no matter what *natural strengthening* of this theory we take up, we cannot disagree about $\Pi^1_2$-statements.[54] Following the idea above, we might take the $\Pi^1_2$-statements as the concrete ones that we really care about and the possible exotic extensions of $ZFC$ as mere instruments for learning more about what is concrete. This informal gloss only talks

[53] See Theorem 13.15 in [Kanamori, 2003].

[54] By this, we mean that there will never be a $\Pi^1_2$ sentence $\varphi$ of the stronger theory such that $ZFC$ proves its negation. See [Steel, 2014] and [Maddy and Meadows, 2020] for more discussion of this.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

about natural strengthenings, so the scope of this claim is arguably limited. The idea that motivates this is that the ways of extending $ZFC$ that have turned out to have been mathematically fruitful are linked to the large cardinal hierarchy by inner model and forcing arguments.[55]

This example gives us a reasonable prototype for "empirical equivalence" in the foundations of mathematics. We now put our interpretative tools to work to abstract out what is arguably the underlying mechanism of the example. This will allow us to draw some links with the work of the previous sections and highlight the fact that "empirical equivalence" is fruitfully understood through the lens of relative interpretability. We start by describing the idea semi-formally and then give the proper definition. Suppose we have two theories $T$ and $S$ that purport to be foundations for mathematics. Then we might consider a particular topic of mathematics that is discussed in some language $\mathcal{L}^*$ and wonder whether they agree about it. For example, we might consider the ways in which $T$ and $S$ are applied to analysis. To do this we consider the ways in which $T$ and $S$ interpret the language of analysis. Note, however, in the example above, we only obtain agreement over the $\Pi^1_2$ fragment of the language of analysis. In order to model this situation better we shall overload our notation and let $\mathcal{L}^*$ represent both a language and some fragment of the sentences associated with that language.[56] We shall then speak of $\mathcal{L}^*$-sentences as those sentences of the language associated with $\mathcal{L}^*$ that are members of the fragment of associated with $\mathcal{L}^*$.

**Definition 34.** Let $T$ and $S$ be theories in languages $\mathcal{L}_S$ and $\mathcal{L}_T$ respectively and let $\mathcal{L}^*$ be some language associated with a fragment of its sentences. Then suppose we have translations $p$ and $q$ determining mod-functors such that:[57]

$$p : mod(T) \to mod(\mathcal{L}^*) \leftarrow mod(S) : q.$$

We say that $T$ and $S$ are *agree on* $\mathcal{L}^*$ *via* $p$ *and* $q$ if for all $\mathcal{L}^*$-sentences $\varphi$[58]

$$T \vdash p(\varphi) \;\Leftrightarrow\; S \vdash q(\varphi).$$

The idea here that we interpret $\mathcal{L}^*$ in $T$ and $S$ using $p$ and $q$; and then we say that they agree about $\mathcal{L}^*$ if the respective translations give the same $\mathcal{L}^*$-theory; i.e., the same theorems in the fragment associated with $\mathcal{L}^*$. We might think of $\mathcal{L}^*$ as the "empirical" domain that we really care about. In contrast to most of the relationships discussed in Section 2, it appears to be very weak. For instance, there is no demand for any kind of back-and-forth agreement between the theories. Indeed, we haven't even required that they be mutually interpretable. Rather, it seems most similar in structure to mutual faithful interpretability, as in Definition 10, and it is perhaps best understood as a partial version of that. Nonetheless, it does have a simple model-theoretic gloss.

**Proposition 35.** *The following are equivalent:*

[55] See [Koellner, 2010] and [Steel, 2014] for more detailed discussion of this.

[56] Recall that a fragment is just a subset of the set of sentences in some language. Strictly speaking, it would be more precise to denote what we have called $\mathcal{L}^*$ above as a pair $\langle \mathcal{L}^*, \Gamma \rangle$ where $\mathcal{L}^*$ is a language and $\Gamma$ is a subset of the first order sentences associated with $\mathcal{L}^*$. I've opted to avoid this to keep things a little cleaner on the page, however, I shall take care to use footnotes to head off potential confusions.

[57] Note that I'm abusing notation by writing $mod(\mathcal{L}^*)$ to denote all of the models of language associated with $\mathcal{L}^*$. We might call these $\mathcal{L}^*$-models.

[58] Recall that below Theorem 4, I noted that we'd abuse notation and use $t$ for both the translation of $\mathcal{L}^*$ into $\mathcal{L}_T$ and the mod-functor $t^* : mod(T) \to mod(\mathcal{L}^*)$. It seems apropos to mention it here since this is the first time since then that we've abused our notation in this way.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

(1) *$T$ and $S$ agree on $\mathcal{L}^*$ via $p$ and $q$;*

(2) *The pointwise images of the models of $T$ and $S$ through $p$ and $q$ respectively give the same models of $\mathcal{L}^*$ up to elementary equivalence in $\mathcal{L}^*$, or more formally*[59]

$$(p``mod(T)/\equiv_{\mathcal{L}^*}) = (q``mod(S)/\equiv_{\mathcal{L}^*}).$$

*Proof.* ($1 \to 2$) Suppose $\mathcal{M}$ is a model of $T$. We show that there is a model $\mathcal{N}$ of $S$ such that $p(\mathcal{M}) \equiv_{\mathcal{L}^*} q(\mathcal{N})$. This will suffice for one direction and the other is similar. First, we observe that any $\varphi \in \mathcal{L}^*$ that is true in $p(\mathcal{M})$ is such that $q(\varphi)$ is consistent with $S$. To see this note that since $p(\mathcal{M})$ is a model of $T$, $p(\varphi)$ is consistent with $T$ and so by (1) we see that $q(\varphi)$ is consistent with $S$. From here, it is easy to see that for any finite subset $\Delta$ of $Th_{\mathcal{L}^*}(p(\mathcal{M}))$, $q``\Delta$ is consistent with $S$. And so by compactness, we may fix a model $\mathcal{N}$ of $S$ where $q``Th_{\mathcal{L}^*}(p(\mathcal{M}))$ is satisfied. Thus, $q(\mathcal{N})$ satisifies $Th_{\mathcal{L}^*}(p(\mathcal{M}))$ as required.

($2 \to 1$) Suppose $T \cup \{p(\varphi)\}$ is consistent. We show that $S \cup \{q(\varphi)\}$ is consistent and leave the other (very similar) direction to the reader. First, we fix a model $\mathcal{M}$ of $T \cup \{p(\varphi)\}$ and then fix a model $\mathcal{N}$ of $S$ such that $p(\mathcal{M}) \equiv_{\mathcal{L}^*} q(\mathcal{N})$. Then we see that since $\varphi$ is satisfied by $p(\mathcal{M})$, $\varphi$ is also satisfied by $q(\mathcal{N})$ and so $\mathcal{N}$ is a model of $S \cup \{q(\varphi)\}$ which suffices by soundness. □

Observe now that Proposition 33 can be understood within this framework. To see this, first recall that there are many natural translations of the language of analysis into the language of set theory. Let us fix one of them can call it the *standard translation*.

**Proposition 36.** *Let $T$ and $S$ be theories extending $ZFC$ that are equiconsistent as witnessed by inner model interpretations or forcing. Then $T$ and $S$ agree on the $\Pi^1_2$ fragment of analysis via the standard translation of analysis in set theory.*

In Section 2.3, we noted above that forcing arguments using generic extension don't strictly give us interpretations since they augment the domain. For this reason, we made use of Boolean valued ultrapowers to get around this. However, to make the passage of our current inquiry a little smoother, we'll now proceed as if generic extensions do give us interpretations.[60] It's also worth noting that the standard translation works on every sentence in analysis, not just the $\Pi^1_2$ sentences. However, this has no effect on the results.

We now have a general notion of partial interpretation based on interpretability that fits our canonical example. It turns out that with the addition of large cardinal axioms, further examples of agreement can be gained that encompass the entire language of analysis and beyond. This phenomenon has been discussed extensively by John Steel in his work on eventual monotonicity and what he calls a *theory of the concrete* [Steel, 2014, Feferman et al., 2000, Steel, 2010]. For one example, we have the following.

**Theorem 37.** *[Steel, 2014] $ZFC + AD^{L(\mathbb{R})}$ and $ZFC$ plus there are infinitely many Woodin cardinals agree on the theory of the analysis ($\Pi^1_\omega$).*

[59] A model of $\mathcal{L}^*$ is a model of the language associated with $\mathcal{L}^*$. We write $\mathcal{M} \equiv_{\mathcal{L}^*} \mathcal{N}$ to indicate that $\mathcal{M}$ and $\mathcal{N}$ satisfy exactly the same $\mathcal{L}^*$-sentences. Similarly, we write $Th_{\mathcal{L}^*}(\mathcal{M})$ for the set of $\mathcal{L}^*$-sentences that are satisfied in $\mathcal{M}$.

[60] For one way of making this precise see [Meadows, 2022] in the context of countable transitive models. To expand this approach to deal with models that might now be well-founded, see the discussion in the appendix of [Maddy and Meadows, 2020].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

Very roughly, we start by establishing the equiconsistency of these theories: a forcing argument takes us from $ZFC$ and the Woodin cardinals to a model of $ZFC + AD^{L(\mathbb{R})}$; and an inner model argument takes us from a model of $ZFC + AD^{L(\mathbb{R})}$ to a model of $ZFC$ with infinitely many Woodin cardinals.[61] The arguments then proceed by showing that sufficient structure is preserved in these interpretations to ensure that the theory of analysis is preserved.[62]

3.2. **The limitations of empirical agreement.** As we observed above, agreement seems very similar to mutual faithful interpretability. Moreover, we saw reasons in Section 2.1.2 to be pessimistic about the philosophical significance of faithful interpretability; in particular, it was too easy to obtain and provided an implausible analysis of our target idea of arguing past each other. Perhaps worryingly, with our definition of agreement above, we are only concerned with the translations of consequences of the theories. We are not offering a more natural, structural connection between the theories themselves. As such, the fact that $T$ and $S$ agree on $\Pi^1_2$ sentences doesn't tell us much about $T$ and $S$ or their connection. Perhaps this is enough for an argument based on common ground, but we'll see in this section that it leaves some unsatisfying equivalences on the table that are difficult to remove by formal means. To see this recall that in the example from the previous section, we considered theories were mutually related by forcing or inner model arguments. But this is not required to obtain agreement as we have defined it. Here is an arguably pathological illustration of this.

**Proposition 38.** *Suppose there is a countable transitive model of $ZFC$ plus an inaccessible cardinal.*[63] *Then there are mutually interpretable theories $T$ and $S$ extending $ZFC$ that agree on $\Pi^1_2$ sentences but are not mutually related by forcing or inner model interpretations.*

*Proof.* Let $T$ be $ZFC$ plus the statement that there is an inaccessible cardinal. Let $S$ be $ZFC$ plus every statement $\varphi \in \Pi^1_2$ that $T$ proves. Noting that $T$ and $S$ have the same $\Pi^0_1$-consequences, we see using Theorem 8 that they are mutually interpretable. Then since a model of $T$ is a model of $S$, we see $T$ interprets $S$ by the trivial inner model, itself. So it suffices to show that $S$ cannot interpret $T$ by a forcing or inner model construction. To see this, use our assumption to fix a countable transitive model $M$ of $T$ and let $M^\dagger = L^M$. Then let $M^* = V_\kappa^{M^\dagger}$ where $\kappa$ is the least inaccessible cardinal according to $M^\dagger$. Then $M^\dagger$ is a model of $S$ that thinks there are no inaccessible cardinals. To complete the proof we observe that $M^*$ has no inner model or generic extension that thinks there is an inaccesible cardinal. To see this, first note that since $M^*$ satisifies $V = L$ is has no proper inner models and that $M^*$ doesn't satisfy $T$. Then observe that if $G$ is $M^*$-generic, $M^*[G]$ cannot think there is an inaccessible since this is a $\Pi_1$-statement and so by downward absoluteness, this would imply that $M^*$ also had an inaccessible. □

The connection between $T$ and $S$ from the proof above is clearly very weak as is evidenced by the fact that, in general, a model of $S$ will only be able to define a model of $T$ that is ill-founded. There is also something sufficiently odd about $S$ to warrant consideration about whether such theories should

[61] See Theorem 32.16 and the sketch below it from [Kanamori, 2003] for more details. And for a thorough versions of these arguments see Neeman's chapter in [Foreman and Kanamori, 2009] and Koellner and Woodin's chapter that follows it.

[62] See [Steel, 2014, Maddy and Meadows, 2020] and [Meadows, 2021] for further discussion of the philosophical significance of this phenomenon.

[63] This assumption is overkill, but makes for an easy proof.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

be taken seriously. In favor of $S$, we might observe that it appears to provide a way for us to obtain more results in analysis without explicitly committing to the existence of large cardinals. This could be seen as an advantage by those who are shy about the apparent ontological excesses of contemporary set theory. However, the way in which this advantage is obtained is in some sense parasitic: we are enjoying the benefits of theft over honest toil by simply lifting the$\Pi^1_2$ theorems of $T$ in order to even formulate $S$. John Steel is critical of this move and calls it an *instrumental dodge* [Feferman et al., 2000]. While it is true that $S$ doesn't commit itself to the existence of an inaccessible cardinal and this might seem like a benefit, we are left to wonder how one might extend or even just gain a proper understanding of $S$, without investigating and understanding $T$ better. It seems that the proper way to investigate $S$ must go through $T$ and so the apparent alleviation of large cardinal burdens is more of a dodge than genuine ontological parsimony. More practically, the use of $S$ invokes a kind of bad faith in that the only way to obtain the extra $\Pi^1_2$ theorems is to prove them using $T$ and then sweep them up into $S$ by fiat. One wonders whether the inaccessible cardinal of $T$ has really gone away after all.

This kind of pathology motivates a brief (and somewhat disappointing) effort to obtain a deeper connection between theories and the languages they interpret. Our idea is to not merely demand that the $\mathcal{L}^*$-consequence of two theories are the same, but that the theories themselves are mutually interpretable in a way that preserves their interpretation into $\mathcal{L}^*$. The following definition is intended to capture this idea.

**Definition 39.** Let $T$ and $S$ be theories in $\mathcal{L}_S$ and $\mathcal{L}_T$ respectively and let $\mathcal{L}^*$ be some language associated with a fragment of its sentences. Suppose we have $t : mod(T) \leftrightarrow mod(S)$ and

$$p : mod(T) \to mod(\mathcal{L}^*) \leftarrow mod(S) : q.$$

We say that $t$ *(sententially) preserves* $\mathcal{L}^*$ *from* $T$ *over* $p$ *to* $S$ *over* $q$ if for all $\mathcal{L}^*$-sentences $\varphi$

$$T \vdash p(\varphi) \leftrightarrow t \circ q(\varphi).$$

If in addition, there is some $s : mod(S) \to mod(T)$ such that $s$ preserves $\mathcal{L}^*$ from $S$ over $q$ to $T$ over $p$, we say that $s$ and $t$ *mutually preserve* $\mathcal{L}^*$ *between* $S$ *and* $T$ *over* $q$ *and* $p$.

The key difference between preservation and agreement is that we also demand some connection between $T$ and $S$ in that they must be mutually interpretable, which then gives us a conduit through which we can track the information about $\mathcal{L}^*$ that is passed between these theories. Before we give an example, we note that a more convenient model-theoretic characterization is also available.

**Proposition 40.** *Given $T, S, t, s, \mathcal{L}^*, p$ and $q$ as in Definition 39, the following are equivalent:*

(1) *$t$ (sententially) preserves $\mathcal{L}^*$ from $T$ over $p$ to $S$ over $q$; and*
(2) *$p(\mathcal{M}) \equiv_{\mathcal{L}^*} q \circ t(\mathcal{M})$ for all $\mathcal{M} \in mod(T)$.*

*Proof.* To see this observe that the following are equivalent:

- $p(\mathcal{M}) \equiv_{\mathcal{L}^*} q \circ t(\mathcal{M})$ for all $\mathcal{M} \in mod(T)$;
- $p(\mathcal{M}) \models \varphi$ iff $q \circ t(\mathcal{M})$ for all $\mathcal{M} \in mod(T)$ and $\varphi \in \mathcal{L}^*$;
- $\mathcal{M} \models p(\varphi)$ iff $\mathcal{M} \models t \circ q(\varphi)$ for all $\mathcal{M} \in mod(T)$ and $\varphi \in \mathcal{L}^*$; and

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

- $T \vdash p(\varphi) \leftrightarrow t \circ q(\varphi)$ for all $\varphi \in \mathcal{L}^*$.

□

Suppose $T$ and $S$ are set theories and that $\mathcal{L}^*$ is the language of analysis. Then if the relation defined above holds and we are given some model $\mathcal{M}$ of $T$, $\mathcal{M}$ defines a model $p(\mathcal{M})$ of arithmetic that is elementary equivalent to the model of arithmetic $q \circ t(\mathcal{M})$ defined in the model $t(\mathcal{M})$ of $S$ that is defined in $\mathcal{M}$. It is then easy to see that mutual preservation between theories implies agreement.[64] Moreover, we see that the scenario of Proposition 33 also gives us mutual preservation.

**Proposition 41.** *Let $T$ and $S$ be theories extending $ZFC$ that are equiconsistent as witnessed by inner model interpretations or forcing. Then those interpretations mutually preserve the $\Pi^1_2$ fragment of analysis $T$ and $S$ via the standard translation of analysis in set theory.*

*Proof.* Let $\mathcal{M}$ be a countable model of $T$ and suppose that $t : mod(T) \leftrightarrow mod(S) : s$ are interpretations given by forcing or inner models. Let $p$ be the translation of analysis into the language of set theory. We claim $p(\mathcal{M}) \equiv_{\Pi^1_2} p \circ t(\mathcal{M})$. This then holds since as we've remarked, forcing and inner model arguments preserve $\Pi^1_2$ statements. Similarly, $p(\mathcal{N}) \equiv_{\Pi^1_2} p \circ s(\mathcal{N})$ for models $\mathcal{N}$ of $S$. Thus, we have mutual preservation. □

This suggests that our initial abstraction of Proposition 33 to obtain the definition of agreement missed an important feature of the relationship between such theories. So are there any theories that satisfy Definition 34 but not Definition 39? Here is an easy example where agreement holds but preservation fails.

**Proposition 42.** *There exist theories $T$ and $S$ that are $\mathcal{L}$-faithful over $p$ and $q$ but $\mathcal{L}$ cannot be mutually preserved between $T$ and $S$ over $p$ and $q$.*

*Proof.* Since $GBN$ is a model theoretic conservative extension of $ZFC$, we see that $ZFC$ and $GBN$ agree on the language of analysis over the standard translations. However, $ZFC$ and $GBN$ are not mutually interpretable since – by a reflection argument – $ZFC$ cannot interpret $GBN$. □

This does give us a strict hierarchy between agreement and mutual preservation, but there is something cheap about the failure since we blocked it by selecting a pair of theories that are very close but are, somewhat annoyingly, not mutually interpretable. So can we can have mutually interpretable theories that agree on some language but do not mutually preserve it? An affirmative answer to this question might put us in a position to formally identify theories that exploit an instrumental dodge, like those from Proposition 38. Unfortunately and perhaps surprisingly, a generalization of Lindström's argument for Theorem 11 shows that it many cases, agreement is sufficient for mutual preservation. To keep things short and simple, we provide a relatively specific example and then merely remark about how it can be generalized.

**Theorem 43.** *Suppose $T$ and $S$ are essentially reflective, recursively axiomatizable theories extending $ZFC$ that agree on the theory of arithmetic. Then $T$ and $S$ mutually preserve arithmetic.*

[64] For example, given the setup of Definition 39 and model $\mathcal{M}$ of $T$, we see that $p(\mathcal{M}) \equiv q \circ t(\mathcal{M})$ where $t(\mathcal{M})$ satisfies $S$. A similar argument applies to models of $S$, and so we see that $(p``mod(T)/\equiv) = (q``mod(S)/\equiv)$.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*Proof.* Suppose $\mathcal{M}$ is a model of $S$. We aim to define a model $\mathcal{N}$ of $T$ in $\mathcal{M}$ with the same theory of arithmetic. More precisely, we want to show that $\mathcal{M}$ can define a model of

$$T^* = T \cup \{\varphi \in \mathcal{L}_{Ar} \mid \mathcal{M} \models \varphi\}.$$

Fix a formula defining $T^*$ and let $T^{*\mathcal{M}}$ be the denotation of that formula in $\mathcal{M}$. Note that it can be arranged that the standard part of $T^{*\mathcal{M}}$ is $T^*$. Now observe that for any finite subset $\Gamma_0$ of $T$ and finite set $\Gamma_1$ of arithmetic sentences,

$$T \vdash \bigwedge \Gamma_1 \to Con(\Gamma_0 \cup \Gamma_1).$$

This holds since $T$ extends $ZFC$ and so it can prove that $\Gamma_0$ holds in some initial segment of the universe which, of course, contains its version of $\mathbb{N}$. Now since this sentence is itself arithmetic and $T$ and $S$ agree on arithmetic, we see that $S$ also proves this sentence.

We now claim that $\mathcal{M} \models Con(\Delta)$ for all finite subsets of $T^*$. To see this, let $\Delta$ be such a set and let $\Delta = \Delta_0 \cup \Delta_1$ where $\Delta_0 \subseteq T$ and $\Delta_1$ is a set of arithmetic sentences that are true in $\mathcal{M}$. Then since $\Delta_1$ is satisfied in $\mathcal{M}$ and $S \vdash \bigwedge \Delta_1 \to Con(\Delta_0 \cup \Delta_1)$ we see that $\mathcal{M} \models Con(\Delta_0 \cup \Delta_1)$ as required.

Now we can define the model of $T^*$. We do this in two cases noting that either $\mathcal{M} \models Con(T^{*\mathcal{M}})$ or $\mathcal{M} \not\models Con(T^{*\mathcal{M}})$. In the first case, we let $\mathcal{N}$ be what $\mathcal{M}$ thinks is the $L$-least model satisfying $T^{*\mathcal{M}}$. In the second case, we let $c$ be what $\mathcal{M}$ thinks is the largest natural number such that the set $X$ of codes of formulae from $T^{*\mathcal{M}}$ up to and including $c$ are consistent. Such a $c$ must exist in $\mathcal{M}$ since it thinks that $T^{*\mathcal{M}}$ is inconsistent. Moreover, by a previous claim, we see that $c$ must be in the nonstandard part of $\mathcal{M}$'s natural numbers. Thus, $X$ includes $T^*$ and is consistent according to $\mathcal{M}$. Thus, we let $\mathcal{N}$ be the $L$-least model of $X$. □

Thus, we see that under a few conditions that are easily satisfied, mutual agreement between theories over some language implies that they also preserve it. Thus, the effort to demand more connection between theories puts us in no better position with regard to the problem of the instrumental dodge. Moreover, this result can easily be generalized as the following theorem demonstrates.

**Theorem 44.** *Suppose $T$ and $S$ are recursively axiomatizable theories extending $ZFC$ and the statement that $0^{\#}$ exists. Moreover suppose that $T$ and $S$ agree on their $L$ consequences. Then $T$ and $S$ can mutually preserve their theories of $L$.*

We omit the proof, but the key point is that since $T$ and $S$ both think that $0^{\#}$ exists, they can both define their theories of $L$.[65] For a particularly pathological example, we could let $T$ be $ZFC$ plus the existence of a measurable cardinal and $S$ be $ZFC$ plus the statement that $0^{\#}$ exists and all of the $L$-consequences of $T$. Then $T$ and $S$ mutually preserve their theories of $L$. This is another obvious example of an instrumental dodge. So our efforts to avoid this pathology appear to have been in vain. For this reason, we'll content ourselves with mere agreement for the rest of this paper. Despite this, these results shed some interesting light on the relationship between some natural generalizations of mutual faithful interpretability and sentential equivalence to partial contexts and further work should be done here.

[65] See Chapter 18 of [Jech, 2003] for more details.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

## 4. Maximizing the zone of agreement

We now have a serviceable theory of partial interpretation in place, but we still seem to lack a satisfying explanation of what is driving the agreement we saw in Proposition 33. What is it about their connection that leads them to agree on so much? Our goal in this final section is twofold. First, we aim to refine our analysis of Proposition 33 by developing a better understanding of the structure that is preserved by the forcing and inner model interpretations used there. This will give us an account that is much more amenable to generalization and leads naturally to our second goal: to better understand the boundaries of the common ground that can be forged between strong set theories. Following Woodin an Koellner, we shall calibrate the our common ground using what they call strong logics. Indeed, the ensuing discussion is heavily influenced by the work of Koellner, Steel, Väänänen and Woodin [Koellner, 2010, Steel, 2014, Feferman et al., 2000, Ikegami and Väänänen, 2015, Woodin, 2001, 2012]. Steel's discussion of a theory of the concrete in [Steel, 2014] and Koellner's intriguing story of logics between $\beta$-logic and $\Omega$-logic in [Koellner, 2010] provided the initial inspiration for this section. My somewhat modest goal has been to pull these threads together and slow the passage down so that we might take a closer look at what lies along the way. Nonetheless, we should note that strong logics are just one instrument, among many, that we might use to calibrate agreement. In the context of this paper, they deliver a good way of slowing things down and a surprising connection between second order logic and forcing, as is nicely illustrated in [Ikegami and Väänänen, 2015]. In a similar fashion to Section 2, we will conduct another Goldilocks search. In that section, we quickly found some unsuitable lower bounds and then worked our way down from the impractical relations like bi-interpretability to sentential equivalence and various generalizations thereof. By contrast, here we will quickly find an impractical upper bound and then we'll commence a long journey upward from the logic underlying Proposition 33 with the help of large cardinal assumptions.

Let us start by discussing what Woodin and Koellner call strong logics. Rather than considering fragments of languages, like the $\Pi^1_2$ fragment of analysis, we are going to instead consider agreement on consequence relations strengthening first order logic by restricting the kinds of model that we admit. As always, we shall restrict our attention to the language of set theory, $\mathcal{L}_\in$, unless otherwise stated. What we shall call an $X$-logic will then be determined by a (generally proper) class $X$ of models of $\mathcal{L}_\in$. When given sentences $\Gamma \cup \{\varphi\}$ from $\mathcal{L}_\in$, we shall say that $\varphi$ is an $X$*-consequence* of $\Gamma$, abbreviated $\Gamma \models_X \varphi$, if every $\mathcal{M} \in X$ that satisfies $\Gamma$ also satisfies $\varphi$.[66] In the case of Proposition 33, there is a particularly natural $X$-logic that gives a very natural counterpart to the $\Pi^1_2$ sentences of analysis. Let $\beta$ be the class of well-founded models and consider $\beta$-logic. We then see that:

**Proposition 45.** *The following are Turing equivalent:*

(1) $\{\varphi \in \mathcal{L}_\in \mid \ \models_\beta \varphi\}$;

[66] This is essentially the approach to strong logics used by Woodin in [2012]. We note also that there is good reason to not think of $X$-logics as genuine logics. In the model-theoretic tradition, a logic is not restricted to a particular language, like $\mathcal{L}_\in$, and there is an expectation that it will satisfy certain invariance conditions. A typical example of an orthodox strong logic is that obtained by extending first order logic with a new quantifier $W$ which is such that $Wxy\varphi(x,y)$ holds in some model iff $\varphi(x,y)$ is a well-founded relation. A general definition of logics like these can be found in Definition 1.1.1 of [Ebbinghaus, 1985]. Even if we allow languages extending $\mathcal{L}_\in$, then the logics considered below all fail to satisfy this definition since they do not have the renaming properly. To emphasize this distinction, we thus refer to what Woodin and Koellner call strong logics as $X$-logics, where the $X$ is some implicit class of $\mathcal{L}_\in$-models. We thank an anonymous referee for this helpful suggestion.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

(2) *The set of true $\Pi^1_2$ sentences.*

*Proof.* ($1 \leq_T 2$) Note that $\models_\beta \varphi$ iff

$$\forall M \; (M \text{ is well-founded} \rightarrow M \models \varphi) \tag{4.1}$$

and this statement is clearly $\Pi^1_2$. So the Turing machine sending $\varphi$ to the corresponding version of 4.1 witnesses that $1 \leq_{1\text{-}1} 2$. ($2 \leq_T 1$) Let $t : \Pi^1_2 \rightarrow \mathcal{L}_\in$ take $\varphi$ and return a sentence of $\mathcal{L}_\in$ saying that the universe is $V_{\omega+1}$ and (the standard translation) of $\varphi$ holds. Then $\varphi \Leftrightarrow \models_\beta t(\varphi)$ as required. □

Thus, we might say that $\beta$-logic is the natural logical counterpart to the $\Pi^1_2$ sentences of analysis. The proof above also establishes that the consequence relation for $\beta$-logic is $\Pi^1_2$-complete. With this in hand, we can describe a suitable notion of agreement between theories regarding a logic.

**Definition 46.** Let us say that $T$ and $S$ *completely agree* on an $X$-logic, if

(1) $T \models$ "$\Gamma \models_X \varphi$" iff $S \models$ "$\Gamma \models_X \varphi$"; and
(2) $T \models$ "$\Gamma \not\models_X \varphi$" iff $S \models$ "$\Gamma \not\models_X \varphi$"

where $\Gamma \cup \{\varphi\}$ is a recursive set of $\mathcal{L}_\in$ sentences.

This is a strengthening of our previous conception agreement since we are demanding that $T$ and $S$ not only agree on the $X$-consequences, they must also agree on $X$-satisfiability. This seems like the natural notion of agreement between $X$-logics and it helps avoid the effects of idiosyncratic upward absoluteness facts. Now we are ready to introduce the new ingredient into our analysis of Proposition 33: structural preservation. The idea here is that beyond demanding that our theories are mutually interpretable, we also want those interpretations to preserve the structure that guarantees the agreement we desire. We start with a form of agreement tailored to Proposition 33.

**Definition 47.** We say that a mod-functor $t : mod(T) \rightarrow mod(S)$ *preserves* $L$ if $L^{\mathcal{M}} = L^{t(\mathcal{M})}$ whenever $\mathcal{M} \models T$.

It is easy to see that both inner model and forcing interpretations preserve $L$ since they essentially just make a model wider or thinner and $L$ is the thinnest inner model of $ZFC$. However, there are other interpretations that preserve $L$ that are neither forcing nor inner model interpretations. For one species of example, consider the symmetric extensions used to obtain models where the axiom of choice fails.[67] We now put these pieces together to offer a new analysis of Proposition 33.

**Theorem 48.** *If $T$ and $S$ are extensions of $ZFC$ that are mutually interpretable via interpretations that preserve $L$, then $T$ and $S$ completely agree about $\beta$-logic.*

*Proof.* Suppose $t : mod(T) \leftrightarrow mod(S) : s$ are such interpretations. Then suppose $T \nvdash$ "$\Gamma \models_\beta \varphi$". Then we may fix $\mathcal{M} \models$ "$\Gamma \not\models_\beta \varphi$". Since $t(\mathcal{M})$ preserves $L$ and $\not\models_\beta$ is a $\Sigma^1_2$ property, the Lévy-Shoenfield theorem tells us that $t(\mathcal{M}) \models$ "$\Gamma \not\models_\beta \varphi$" and so $S \nvdash$ "$\Gamma \models_\beta \varphi$". Essentially the same argument suffice shows $T \nvdash$ "$\Gamma \not\models_\beta \varphi$" implies $S \nvdash$ "$\Gamma \not\models_\beta \varphi$" and analogous arguments then work in the other direction. □

[67] See [Jech, 2008] for the canonical source on this technique.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

It is worth noting that the interpretations used above can involve both forcing and inner models. For example, let $T$ be the theory where we augment $ZFC$ by the statement $V = L[c]$ where $c$ is a Cohen real; and let $S$ be the theory where we add to $ZFC$ the statement $V = L[r]$ where $r$ is a random real. Neither $T$ nor $S$ can interpret the other by an inner model interpretation or generic extension alone. However, $T$ may interpret $S$ by going to its version of $L$ *and then* forcing to add a random real. This two step interpretation clearly preserves $L$ and thus fits the template above. Similar remarks apply to $S$.

This modified perspective gives a promising approach to carving out larger tracts of agreement between set theories while also offering a kind of explanation for why the agreement occurs. Our eventual goal will be to generalize the kind of structural preservation employed above to obtain even more common ground via agreement over stronger $X$-logics. But before we move on, it's also worth remarking that agreement on $\beta$-logic is not at all trivial. It is well known that much of ordinary mathematics can be construed as going on below the $\Pi^1_2$-level (in the codes).[68] Perhaps it is no coincidence that $ZFC$ is so good at handling countable transitive models: the natural calculators of the $\Pi^1_2$ realm. As such, it would not be absurd for someone to think that agreement over $\beta$-logic is all the common ground we need between our set theories and that debates about theories that agree on $\beta$-logic are just quibbles over our preferred mathematical scaffolding. Nonetheless, to further our investigation we shall take it that a good foundation should have room for future expansion in mathematics and as such, we would like to understand where the limits of agreement might lie.[69]

4.1. **Too much to ask for.** Now we aim to explore the limits of agreement using the framework proposed above. An obvious and natural first question leaps to mind: what is the strongest $X$-logic we could hope to incorporate into the common ground?[70] The obvious and naive answer is: we'd like to preserve the logic that has just one model, the entire universe. But since the universe is not itself a set, there are tedious Tarskian issues around defining this as an $X$-logic. Moreover, in including the whole universe we would have departed from our project to obtain merely partial agreement. So let's look for what seems to be the next natural step down. Let $V.$ be the class consisting of all rank initial segments of the universe; i.e., $V_\alpha$ for all ordinals $\alpha$. Then for $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_\in$, we say that $\varphi$ is a $V.$ consequence of $\Gamma$, abbreviated, $\Gamma \models_{V.} \varphi$, if for all $\alpha$, $V_\alpha \models \varphi$ whenever $V_\alpha \models \Gamma$. Like $\beta$-logic, $V.$-logic has some very natural cousins.

**Theorem 49.** *(Essentially Väänänen) The following are Turing equivalent:*

(1) $\{\varphi \in \mathcal{L}_\in \mid \ \models_{V.} \varphi\}$*;*
(2) *the set of true* $\Pi_2$ *sentences; and*
(3) $\{\varphi \in \mathcal{L}_\in \mid \ \models_{SOL} \varphi\}$.

*where* $\models_{SOL}$ *is the consequence relation for full second order logic.*[71]

[68] See for example [Simpson, 1999].

[69] For further discussion of this, see Maddy's notion of the Generous Arena [Maddy, 2016].

[70] I'm being a little vague with what it means for one $X$-logic to be stronger than another. Informally and roughly, we might think of one $X$-logic as being stronger than another if it allows us to access more of the true theory of sets. For a more formal relation, we might say $X$-logic is stronger than $Y$-logic if $X \subseteq Y$.

[71] See [Shapiro, 1991] for more details.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

*Proof.* ($1 \leq_T 2$) Note that $\models_{V.} \varphi$ iff

$$\forall \alpha \forall M (M = V_\alpha \to M \models \varphi) \tag{4.2}$$

where the right hand is $\Pi_2$ since $M = V_\alpha$ is $\Pi_1$.[72] So the Turing machine taking $\varphi$ to the corresponding version of 4.2, we get a 1-1 reduction. ($2 \leq_T 1$) Let $\varphi$ be $\Pi_2$ and $\psi$ be a sentence of $\mathcal{L}_\in$ saying that the universe is a $\beth$-fixed point. Then it can be seen that $\varphi$ holds iff

$$\models_{V.} \psi \to \varphi.$$

To see this note that if $\alpha$ is a $\beth$-fixed point, then $V_\alpha$ is a $\Sigma_1$ elementary submodel of $V$.[73] Now suppose $\varphi$ is of the form $\forall x \chi(x)$ and suppose it is not true. Then we may fix some $x$ such that $\neg\chi(x)$. Let $\alpha$ be a $\beth$-fixed point where $x \in V_\alpha$. Then $V_\alpha \models \psi \wedge \neg\chi(x)$ and so $\not\models_{V.} \psi \to \varphi$. The proofs of ($2 \leq_T 3$) and ($3 \leq_T 2$) are very similar to ($2 \leq_T 1$) and ($1 \leq_T 2$) respectively. For ($2 \leq_T 3$), we use of sentence of $SOL$ that says its universe is a $\beth$-fixed point. For ($3 \leq_T 2$), we exploit the fact that being a full model of $SOL$ is $\Pi_1$ to see that $\models_{SOL}$ is $\Pi_2$. □

I think it's fair to say that these equivalences reveal that $V$.-logic is a very natural $X$-logic. We might even say that $V$.-logic gives us the strongest, natural notion of satisfiability we could hope to make sense of. When $\varphi$ is $V$.-satisfiable it is true in an initial segment of the universe that is maximal in the sense that nothing can be added to it without increasing its rank. But is it a plausible place to find agreement between strong set theories? Can it fit in the common ground? Given this relationship with second order logic, we might wonder if a triviality result like Zermelo's Theorem 1 or Enayat's Corollary 19 is on the cards. Perhaps the only way for theories to agree about $V$.-logic is for those theories to be identical. This is not quite the case.

**Theorem 50.** *Let $ubGCH$ be the statement that the GCH holds unboundedly; more accurately, that $\forall \alpha \exists \beta > \alpha\ 2^{\aleph_\beta} = \aleph_{\beta+1}$. Then $U = ZFC + ubGCH$ and $B = ZFC + \neg ubGCH$ agree*[74] *on $V$.-logic, but they are not identical.*

*Proof.* The theories are obviously not identical, so it suffices to show they agree on $V$.-logic. Suppose $U \not\vdash$ "$\Gamma \models_{V.} \varphi$" and fix countable $\mathcal{M}$ of $U$ such that $\mathcal{M}$ thinks $\Gamma \not\models_{V.} \varphi$. Thus there is some $\alpha \in Ord^{\mathcal{M}}$ such that $V_\alpha^{\mathcal{M}} \models \Gamma \cup \{\neg\varphi\}$. Now we may class force above $\alpha$ in $\mathcal{M}$ to obtain $\mathcal{M}[G]$ such that $\mathcal{M}[G] \models B$ and $V_\alpha^{\mathcal{M}} = V_\alpha^{\mathcal{M}[G]}$.[75] Thus, we see that $B \not\vdash$ '$\Gamma \models_{V.} \varphi$" as required. A similar argument works in the other direction. □

Thus, we see that agreement between theories over $V$.-logic doesn't collapse to identity. But unlike the example from Proposition 33 and its refinement in Theorem 48, it's not clear that some kind of structural feature is being preserved in the interpretations used above. The forcing arguments that link these theories are, so to speak, tinkering with the heavens rather than preserving something feature

[72] For more detail see Lemma 0.2 in [Kanamori, 2003].

[73] For more detail see the discussion above Proposition 22.3 of [Kanamori, 2003].

[74] For ease, we only consider agreement here rather than complete agreement. Note also that Solovay showed that if there is a strongly compact cardinal, then *ubGCH* holds [Solovay, 1974]. This is a rare example where the existence of a large cardinal has a profound effect on global properties of the universe. Nonetheless, we may obtain an analogous theorem by considering unbounded instances of the generalized continuum hypothesis at *regular* cardinals.

[75] See Section VIII.4 on Easton forcing in [Kunen, 2006] for more details.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

that occurs a little closer to the ground. Worse still, agreement over $V$.-logic also distinguishes theories that seem to be very close to each other. Here is a simple example.

**Proposition 51.** *$T_0 = ZFC + 2^{\aleph_{16}} = \aleph_{17}$ does not agree with $T_1 = ZFC + 2^{\aleph_{16}} \neq \aleph_{17}$ about $V$.-logic.*

To see this, let $\varphi$ be a sentence of $\mathcal{L}_{\in}$ that says there is a largest cardinal and it is $\aleph_{17}$. Then $T_0$ proves that $\models_{V.} \varphi \rightarrow 2^{\aleph_{16}} = \aleph_{17}$ while $T_1$ does not. From a practical perspective of applicable set theory, there is not much difference between these theories. But $V$.-logic is designed to take such differences seriously. It is not difficult to see that this phenomena can be generalized to larger definable cardinals. For these reasons, we now leave $V$.-logic behind, although we'll return to a generalization of this $X$-logic when we come to the end of this investigation.

4.2. **The long road.** Our search for the common ground now has two examples of $X$-logics at opposite ends of the common ground spectrum: $\beta$-logic and $V$.-logic. While $\beta$-logic delivers a modest but well-understood amount of common ground, we argued that $V$.-logic was beyond the pale. In this section, we shall explore some of the vast region between them, with a view to opening up an even wider body of common ground. Our plan is to lift the ideas underlying $\beta$-logic agreement to deliver a sequence of stronger and stronger $X$-logics: our long road. While we cannot hope to provide a comprehensive survey here, we will aim to generalize our analysis in such a way that we end up with a template for logical agreement. Our main tools for this journey will be from inner model theory and large cardinals. This section is somewhat technical in comparison to those preceding it and there are fewer philosophical remarks to adorn it. However in essence, our goal is to simply enumerate a number of natural strengthenings of $\beta$-logic and illustrate the tools that can be used to achieve this effect. I believe a clear and patient illustration of these techniques should provide the first steps toward a more edifying foundational understanding of them in this context. The reader less willing to get bogged in details may simply focus on the defining conditions of these $X$-logics and the results delimiting the common ground that they obtain.

4.2.1. *$\gamma$-logic.* We start by recalling that $\beta$-consequences form a $\Pi^1_2$ set and set our initial sights on defining a natural $X$-logic whose consequence relation is $\Pi^1_3$. To do this, we must place a stronger restriction on our target models than mere well-foundedness and for this purpose, we appeal to the theory of sharps.[76] Given a set $x$ we say that $x^{\#}$ exists if there is a non-trivial elementary embedding $j : L(x) \rightarrow L(x)$ where $L(x)$ thesmallest inner model of $ZF$ which contains $x$ as an element.[77] This fact can be used to obtain a transitive model $M_0^{\#}(x)$ of the form $\langle L_\lambda(x), \in, U\rangle$ satisfying certain conditions,[78] the most salient of which are that:

- $M_0^{\#}(x)$ thinks $U$ is a $\kappa$-complete, normal ultrafilter on some $\kappa < \lambda$; and
- $M_0^{\#}(x)$ is *linearly iterable* by $U$ and its images; i.e., for all $\alpha \in Ord$, $Ult_\alpha(M_0^{\#}(x), U)$ is well-founded.[79]

[76] See [Schimmerling, 2001] for an excellent overview of this theory.

[77] There are many equivalent definitions of $x^{\#}$, however, this is most convenient for our purposes. For more details see: [Schimmerling, 2001]; Chapters 18 and 19 of [Jech, 2003]; and Chapters 9, 14 and 21 of [Kanamori, 2003].

[78] See Definition 2.6 in [Schimmerling, 2001].

[79] A detailed description of this process can be found in Schimmerling [2001].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

Such models are known as *mice*. The final condition above has a useful implication: by iterating $M_0^\#(x)$ more than $\omega_1$-times we obtain an elementary embedding $j : M_0^\#(x) \to M^*$ where $M^*$ is a transitive model with $\omega_1 \subseteq M^*$. The Lévy-Shoenfield theorem then tells us that $M^*$ is correct about $\Pi^1_2$ statements and thus, so is $M_0^\#(x)$. So mice of the form $M_0^\#(x)$ are, in effect, perfect calculators of $\Pi^1_2$ truth, much in the same way that countable $\beta$-models are perfect calculators of $\Pi^1_1$ truth. We put them to work in our project by using them to define a new $X$-logic.

**Definition 52.** Let $\gamma$ be the $X$-logic consisting of the set of countable transitive models closed under the operation $x \mapsto M_0^\#(x)$ taking a set to the associated mouse containing it.

We now record a few useful properties of $\gamma$-logic.[80] Recall that a model $N$ is $\underset{\sim}{\Pi}^1_n$-correct if for all $\underset{\sim}{\Pi}^1_n$ statements $\varphi(x)$ about the reals and any real $x$, we have $\varphi(x)^N \leftrightarrow \varphi(x)^V$.

**Proposition 53.** *(1) Being an element of $\gamma$ is a $\Pi^1_2$ property.*

*(2) $\models_\gamma$ is $\Pi^1_3$.*

*(3) If $N$ is in $\gamma$ and $N \models ZFC^-$, then $N[g]$ is $\underset{\sim}{\Pi}^1_2$ correct whenever $g$ is (set) generic over $N$.*

*Proof.* (1) Noting that iterability through just the countable ordinals is sufficient for iterability through all the ordinals, it can be seen that the property of being $M_0^\#(x)$ is a $\Pi^1_2$-property.[81] Then we observe that $N \in \gamma$ iff

$$\underbrace{\forall x \in N \exists y \in N\ \underbrace{(y = M_0^\#(x))}_{\Pi^1_2}}_{\Pi^1_2}.$$

Note that the first two quantifiers don't increase the complexity since they quantify over elements of $N$ and are thus, in effect, arithmetic. (2) Observe that $\models_\gamma \varphi$ iff

$$\underbrace{\forall \mathcal{M} \in \mathbb{R}(\underbrace{\mathcal{M} \in \gamma}_{\Pi^1_2} \to \underbrace{\mathcal{M} \models \varphi}_{\Delta^1_1})}_{\Pi^1_3}.$$

(3) Let $N \in \gamma$ and let $g$ be $N$-generic. Then it can be seen that $N[g]$ is still closed under sharps. Now suppose $\varphi(y)$ is $\Pi^1_2$ where $y \in \mathbb{R}^{N[g]}$. We claim that

$$\varphi(y)^V \Leftrightarrow \varphi(y)^{N[g]}.$$

The ($\Rightarrow$) direction follows by downward absoluteness from $V$ to $N[g]$. For the ($\Leftarrow$) direction we see that $M_0^\#(y)$ exists and is a member of $N[g]$. Suppose that $\varphi(y)^V$. Since $M_0^\#(y)$ is linearly iterable, there is an elementary embedding $j : M_0^\#(y) \to M^*$ where $\omega_1^V \subseteq M^*$. It then follows by the Lévy-Shoenfield theorem that $M^* \models \varphi(y)$ and so by the elementarity of $j$, $M_0^\#(y) \models \varphi(y)$.[82] By downward absoluteness again, we see that $N[g]$ thinks that $M_0^\#(y)$ possesses the $\Pi^1_2$ property of being linearly iterable, and so we may work in $N[g]$ to obtain $i \in N[g]$ such that $i : M_0^\#(y) \to M^\dagger$ is an elementary

[80] The reader might worry that unlike $\beta$-logic, $\gamma$ logic is defined by a closure property on models. But in fact, $\beta$-logic can also be described as a closure property over $\omega$-models. See Theorem 2.23 of [Bagaria et al., 2006].

[81] See Theorem 14.11 in [Kanamori, 2003].

[82] Note that the critical point of the embedding is $> \omega$, so reals are not moved by $j$.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

embedding where $\omega_1^{N[g]} \subseteq M^\dagger$. Then we see that $M^\dagger \models \varphi(y)$ by the elementarity of $i$, and $N[g] \models \varphi(y)$ by the Lévy-Shoenfield theorem as required. □

We thus, have some understanding of the complexity of this $X$-logic and a demonstration that it is a natural generalization of $\beta$-logic. Note that in the argument for (3) as in the argument for Theorem 48 we make heavy use of the Lévy-Shoenfield theorem. In the case of Theorem 48, we did this with the inner model $L$, whereas here we use the merely countable $M_0^\#$ in (3) to get the same effect. We might say that $M_0^\#$ *localizes* the inner model $L$. Indeed we can obtain $L$ from $M_0^\#$ by iterating $M_0^\#$ along the entirety of the ordinals.[83] These observations are useful for our current problem: generalizing Theorem 48 to obtain agreement on $\gamma$-logic. While we demanded the $L$ was preserved in Theorem 48, we are now going to need to preserve a richer structure. For this purpose, we introduce the inner model $M_1$. In essence, $M_1$ is the canonical inner model containing one Woodin cardinal. A little more specifically, $M_1$ can be obtained from its sharp $M_1^\#$ by iterating its active measure out of the universe. $M_1^\#$ is then a stronger version of a mouse than $M_0^\#$ that contains a Woodin cardinal and satisfies a minimality condition.[84] Moreover, the iterability required of $M_1^\#$ makes use of a non-linear iteration described through the device of an iteration tree. We shall be particularly interested in what is known as $\omega_1$*-iterability*.[85] The following remarkable fact is the driving force behind our current interest in $M_1$ and other mice with Woodin cardinals.

**Fact 54.** *[Neeman, 1995]*[86] *Let $x \in \mathbb{R}$ and let $M$ an $\omega_1$-iterable premouse satisfying $ZF^-$ plus the existence of a Woodin cardinal $\delta$ where $\delta$ is countable in $V$. Then there is an elementary embedding $i : M \to M^*$ fixing $\delta$ such that there is a $Col(\omega, \delta)$-generic $g$ over $M^*$ where $x \in M^*[g]$.*

Informally speaking, this tells us that every real number can be *captured* in a generic extension of an iteration of such a mouse. Or as a slogan: every real is a generic somewhere! This process is known as *genericity iteration*. This fact will allow us to gain a kind of access to witnesses of existential statements involving the reals. With this in hand, we may obtain our agreement result.

**Theorem 55.** *Suppose $T$ and $S$ are theories that extending $ZFC$ that imply that $M_1$-exists, it is $\omega_1$-iterable, and that its Woodin cardinal is countable. Suppose also that $S$ and $T$ are mutually interpretable by interpretations that preserve $M_1$ and its iterability. Then $T$ and $S$ completely agree on $\gamma$-logic.*

*Proof.* Since $\models_\gamma$ is $\Pi^1_3$, it will suffice to show that $T$ and $S$ agree on $\Pi^1_3$ statements. Let $\forall x \psi(x)$ be $\Pi^1_3$ where $\psi(x)$ is $\Sigma^1_2$. Suppose that $T \nvdash \forall x \psi(x)$ and fix a model $\mathcal{M}$ of $T$ such that $\mathcal{M} \models \exists x \neg\psi(x)$ and fix $x$ from $\mathcal{M}$ witnessing this. We aim to show that there is a model of $S$ where this statement is also true. To this end, let $t(\mathcal{M})$ be the model of $S$ given by the $M_1$-preserving interpretation, so we have $M_1^{\mathcal{M}} = M_1^{t(\mathcal{M})}$. Working in $\mathcal{M}$, we then use Fact 54 to obtain and elementary embedding

[83] This gives us a structure that is longer than the ordinals, but if we truncate it at the critical point of its ultrafilter, we obtain $L$. This process is often known as *iterating a measure out of the universe.*

[84] A proper definition can be found in [Steel, 1995], [Müller et al., 2020] and in Steel's chapter in [Foreman and Kanamori, 2009].

[85] We'll aim to avoid specific discussion of iteration trees below, but see Definition 5.10 in [Martin and Steel, 1994] for more detail.

[86] A proof can be found in Section 7 of Neeman's chapter in [Foreman and Kanamori, 2009]. There is also another version of this result from Woodin, but this requires $(\omega_1 + 1)$-iterability: see Section 7.2 of Steel's chapter in [Foreman and Kanamori, 2009] or [Farah, 2020].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

$i : M_1 \to M^*$ such that there is an $M^*$-generic $g$ for $Col(\omega, \delta)$ where $x \in M^*[g]$. Since $M^*$ is an inner model and $\neg\psi(x)$ is $\Pi^1_2$, we see that $\psi(x)^{M^*[g]}$ and so we have

$$\begin{aligned} &(\Vdash_{Col(\omega,\delta)} \exists x \neg\psi(x))^{M^*} \\ \Leftrightarrow &(\Vdash_{Col(\omega,\delta)} \exists x \neg\psi(x))^{M_1} \end{aligned}$$

by the elementarity of $i$. Now working in $t(\mathcal{M})$ we may let $h$ be $Col(\omega, \delta)$-generic over $M_1$ and so we see $(\exists x \neg\psi(x))^{M[h]}$. Then since $M_1[h]$ is $\utilde{\Pi}^1_2$-correct in $t(\mathcal{M})$ and $\exists x \neg\psi(x)$ is $\Sigma^1_3$, we see that $t(\mathcal{M}) \models \exists x \neg\psi(x)$ and so, $S \nvDash \forall x \psi(x)$ as required. The rest of the cases are similar. □

Informally speaking, we start with a model $\mathcal{M}$ of $T$ satisfying $\exists x \neg\psi(x)$. We use a genericity iteration to capture a witness of this. Elementarity then ensures that this existential statement is also forced in $M_1$. Then since $M_1$ is shared by $\mathcal{M}$ and $t(\mathcal{M})$ we may work in $t(\mathcal{M})$ to obtain a generic extension of $M_1$ where the existential statement holds. Absoluteness between $M_1$ and $t(\mathcal{M})$ then finishes the proof. The key element, however, is the use of the genericity iteration to capture the required witness. This gives us a nice analogy and generalization of Theorem 48. Instead of merely preserving $L$ we preserve $M_1$ and instead of merely agreeing on $\beta$-logic, we agree on $\gamma$-logic. Here is a helpful illustration of its application.

**Example 56.** Consider the following theories:[87]

(1) $ZFC$ plus there is a Woodin cardinal;
(2) $ZFC$ plus $\Pi^1_2$-determinacy;
(3) $ZFC$ plus there is an $\omega_2$-saturated ideal on $\omega_1$.

Woodin showed that these theories are equiconsistent. Moreover, the interpretations witnessing this can be seen to be $M_1$-preserving and so by Theorem 55, we see that they all agree on $\gamma$-logic.

Thus, we have a number of natural extensions of $ZFC$ that all agree on $\gamma$-logic and indeed on any $\Pi^1_3$ statement. This cements our claim to having generalized $\beta$-logic.

4.2.2. *$\gamma_n$-logic.* Our next goal is to show how this technique can be taken further up through the projective hierarchy by preserving more structure in the form of inner models with the pleasing side effect of also obtaining agreement on stronger $X$-logics. First we describe our next $X$-logic. For $n \geq 1$ and $x \in \mathbb{R}$, $M^\#_n(x)$ is the canonical countable mouse with $n$ Woodin cardinals. Its canonicity is witnessed by a minimality condition.[88]

**Definition 57.** For $n \geq 1$, let $\gamma_n$ be the set of countable transitive models $N$ closed under $x \mapsto M^\#_n(x)$ for for $x \in \mathbb{R}$ where $M^\#_n(x)$ is $\omega_1$-iterable in $N$ and $V$.

For the inner models that our interpretations will preserve, we note that $M_n(x)$ is the inner model obtained from $M^\#_n(x)$ by iterating its topmost and active extender out of the universe.[89] It is the canonical inner model containing $n$ many Woodin cardinals and $x \in \mathbb{R}$.[90] The $M_n$s will play a similar

[87] See Theorem 32.17 from [Kanamori, 2003] and the ensuing discussion on that page.

[88] See the discussion around Theorem 1.2 in [Steel, 1995] for more detail.

[89] An extender can be understood as a directed system of ultrafilters that are able to represent strong embeddings that mere ultrafilters. See Chapter 26 of [Kanamori, 2003] for details.

[90] See [Steel, 1995] for a proper definition and discussion.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

role to $L$ and $M_1$ in our previous agreement results. Toward our agreement claim, we then observe that the models in $\gamma_n$ are increasingly correct; and that this correctness corresponds with preservation of inner models of the form $M_n$. We note that for some technical issues, these results are sensitive to whether we use odd or even $n \in \omega$.[91] The correctness result rests on the following fact:

**Fact 58.** *Suppose $n \geq 0$ is even and $M_n^{\#}(x)$ exists and is $\omega_1$-iterable for all $x \in \mathbb{R}$. Then $M_n^{\#}(x)$, is $\underset{\sim}{\Sigma}^1_{n+2}$-correct.*[92]

**Lemma 59.** *(1) If $n$ is odd and $N \in \gamma_n$, then $N$ is $\underset{\sim}{\Pi}^1_{n+2}$-correct.*[93]

*(2) If $V$ and $W$ are models of ZFC where $M_n^V = M_n^W$ for odd $n$, then for all $\varphi \in \Pi^1_{n+2}$*

$$\varphi^V \Leftrightarrow \varphi^W.$$

*Proof.* (1) We proceed by induction supposing that we have established the claim for $m < n$. Fix $N$ as described above and let $\exists x \psi(x, y)$ be $\Sigma^1_{n+2}$ with $y \in \mathbb{R}^N$ and $\psi(x, y) \in \Pi^1_{n+1}$.[94] Now fix $x \in V$ witnessing this. We claim that

$$\exists x \psi(x, y)^V \leftrightarrow N \models \exists x \psi(x, y).$$

($\Rightarrow$) Since $N \in \gamma_n$, we know that $M_n^{\#}(y) \in N$ and is $\omega_1$-iterable in $N$ and $V$. Let $\delta_0$ be the least Woodin cardinal according to $M^\dagger = M_n^{\#}(y)$. Using Fact 54, we may fix $i : M^\dagger \to M^*$ and $g$ that is $Col(\omega, \delta_0)$-generic over $M^*$ such that $x \in M^*[g]$. Now it can be seen that $M^*[g]$ is $M_{n-1}^{\#}(x \oplus y)$ and so by Fact 58, it is $\underset{\sim}{\Sigma}^1_{n+1}$-correct.[95] This means that $M^*[g] \models \psi(x, y)$ and so we have by elementarity that[96]

$$\begin{aligned} &(\Vdash_{Col(\omega,\delta_0)} \exists x \psi(x, y))^{M^*} \\ \Leftrightarrow &(\Vdash_{Col(\omega,\delta_0)} \exists x \psi(x, y))^{M^\dagger}. \end{aligned}$$

Now working in $N$ we may fix $h \in N$ that is $Col(\omega, \delta)$-generic over $M^\dagger$. So we have $M^\dagger[h] \models \exists x \psi(x, y)$. Fixing some $x \in \mathbb{R}^{M^\dagger[h]}$ witnessing this, it can be seen that $M^\dagger[h]$ is $M_{n-1}^{\#}(x \oplus y)$ and thus $\underset{\sim}{\Sigma}^1_{n+1}$-correct by Fact 58.[97] Finally, by upward absoluteness, we see that $\exists x \psi(x, y)$ holds in $V$.

($\Leftarrow$) Suppose $\psi(x, y)$ is of the form $\forall z \chi(x, y, z)$ where $\chi(x, y, z)$ is $\Sigma^1_n$. Now fix $x \in N$ such that $N \models \forall z \chi(x, y, z)$ and suppose toward a contradiction that

$$\exists z \neg \chi(x, y, z)^V.$$

Noting that $\exists z \neg \chi(x, y, z)$ is $\Pi^1_{n+1}$ and thus $\Sigma^1_{n+1}$ we may reuse the ($\Rightarrow$) part of proof to obtain the result. The proof of (2) is very similar. □

[91] The restriction to even and odd $n \in \omega$ is related to periodicity phenomena in the project hierarchy. See [Müller et al., 2020] for a discussion of how it relates to our current problem. See [Moschovakis, 1980] for a discussion of periodicity theorems.

[92] See Lemma 1.17 of [Müller et al., 2020] for a proof.

[93] The base case here was observed on page 1660 of Steel's chapter in [Foreman and Kanamori, 2009].

[94] Note that $\underset{\sim}{\Sigma}^1_{n+2}$-correctness is the same as $\underset{\sim}{\Pi}^1_{n+2}$-correctness and it works a little more smoothly in this proof.

[95] See the proof of Lemma 1.1.7 in [Müller et al., 2020] for more details on this point. A perhaps easier argument can be made to show that $M^*[g]$ is still $(n-1)$-iterable in essentially the sense of [Neeman, 1995] since $g$ is obtained by a small forcing in relation to the remaining Woodin cardinals of $M^*$. The argument of Lemma 1.1.7 is then easily adapted to show that sensibly defined $(n-1)$-iterable models are $\underset{\sim}{\Sigma}^1_{n+1}$-correct.

[96] Note $\delta_0$ is not moved by $i$ since the embedding is determined by a branch of length $\omega$.

[97] Again, see the proof of Lemma 1.1.7 in [Müller et al., 2020] for more details.

Using this lemma, we can now obtain agreement over $\gamma_n$-logics using strong theories and interpretations that preserve the right inner models.

**Corollary 60.** *Let $n \in \omega$ be even. Suppose $T$ and $S$ are theories extending $ZFC$ that imply that $M_{n+1}$-exists, it is $\omega_1$-iterable, and that its Woodin cardinal is countable. Suppose also that $S$ and $T$ are mutually interpretable by interpretations that preserve $M_{n+1}$ and its iterability. Then $T$ and $S$ completely agree on $\gamma_n$-logic.*

*Proof.* To see this first note that $M_n^{\#}(x)$ is a $\Pi^1_{n+2}$-singleton.[98] Thus, by a similar calculation to that in Proposition 53 (1) and (2), we see that the consequence relation of $\gamma_n$ is $\Pi^1_{n+3}$. Then Lemma 59 can be used to close out the proof. □

Through the use of Woodin cardinals and preservation of their canonical models, we now have a way of obtaining agreement for $X$-logics representing any odd level of the projective hierarchy. Thus, we've generalized the ideas of $\gamma$-logic to through every rung of the theory of analysis.

4.2.3. *$\delta$-logic.* We now continue our journey by considering the limit of the projective logics above. In contrast to the previous sections, instead of defining a new class of target models, we are going to gather up what we've already obtained. As such, we shall define the consequence relation as kind of supremum of the weaker logics below it.[99]

**Definition 61.** Given $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_\in$, we say that $\varphi$ is a $\delta$-consequence of $\Gamma$, abbreviated $\Gamma \models_\delta \varphi$, if there is some $n \in \omega$ such that for all $M \in \gamma_n$, $M \models \varphi$ whenever $M \models \Gamma$.

Informally speaking, $\varphi$ is a thus $\delta$-consequence of $\Gamma$ if $\varphi$ is a $\gamma_n$-consequence of $\Gamma$ for some $n \in \omega$. Noting that the $\gamma_n$-logics get stronger as we go to greater $n \in \omega$, we see that $\varphi$'s being a $\delta$-consequence of $\Gamma$ is effectively saying that $\varphi$ is eventually a consequence of $\Gamma$. Perhaps unsurprisingly, we can then see that $\delta$-logic is closely related to the true theory of analysis.

**Proposition 62.** *For all $x \in \mathbb{R}$ and $n \in \omega$, suppose $M_n^{\#}(x)$ exists and is $\omega_1$-iterable. Then the following sets are arithmetical in each other:*

*(1) The set of true $\Pi^1_\omega$ sentences; and*

*(2) $\{\varphi \in \mathcal{L}_\in \mid ZFC^- \models_\delta \varphi\}$.*

*Proof.* ($1 \leq_{\Pi^0_\omega} 2$) Suppose $\varphi \in \Pi^1_{n+2} \subseteq \Pi^1_\omega$ for some odd $n \in \omega$. Then fix a countable transitive model $N \in \gamma_n$. But then we see by Lemma 59 (1) that $N$ is $\underset{\sim}{\Pi}^1_{n+2}$-correct, so we see that if $\varphi$ is true iff $N \models \varphi$. It is then easy to see that $\varphi$ is true iff $ZFC^- \models_{\gamma_n} \varphi$. And it will suffice to show that:

$$ZFC^- \models_{\gamma_n} \varphi \;\Leftrightarrow\; ZFC^- \models_\delta \varphi$$

since this establishes that the truth of $\Pi^1_\omega$ sentences is reducible to whether they are $\delta$ consequences of $ZFC^-$. To see this first note that the ($\Rightarrow$) direction follows by definition. In the ($\Leftarrow$) direction suppose that $ZFC^- \not\models_{\gamma_n} \varphi$. First note every $N \in \gamma_m$ for $m \geq n$ is $\underset{\sim}{\Pi}^1_{n+2}$-correct and so we have $ZFC^- \not\models_{\gamma_m} \varphi$ for all such $m$. On the other hand, suppose $i < n$. Then fix $N \in \gamma_n$ such that $N \models ZFC \cup \{\neg\varphi\}$.

[98] See the observation after Corollary 4.11 in [Steel, 1995].

[99] This can be understood as a precursor to our eventual destination: $\Omega$-logic.

Then it can be seen that $N$ is also closed under $x \mapsto M_i^{\#}(x)$ and so $N \in \gamma_i$. Thus $ZFC^- \not\models_{\gamma_i} \varphi$ and so $ZFC^- \not\models_\delta \varphi$ as required. ($2 \leq_{\Pi^0_\omega} 1$) Let $\varphi \in \mathcal{L}_\in$ and let $Tr_{\Pi^1_\omega}$ denote the true $\Pi^1_\omega$ sentences. Then it suffices to describe a Turing machine using $Tr_{\Pi^1_\omega}$ as an oracle that takes $\varphi \in \mathcal{L}_\in$ and halts iff $ZFC^- \models_\delta \varphi$. To do this we just check through each $n \in \omega$ whether the statement

$$\forall N(N \in \gamma_n \to N \models \varphi)$$

is in $Tr_{\Pi^1_\omega}$. If it is we halt; it not we repeat the process at $n+1$. □

As above, we can then obtain agreement on $\delta$-logic as follows.

**Theorem 63.** *If $T$ and $S$ are theories that extending $ZFC$ that imply that for all $n \in \omega$, $M_n$-exists and is $\omega_1$-iterable, and which are mutually interpretable by interpretations that preserve each $M_n$ and its iterability. Then $T$ and $S$ completely agree on $\delta$-logic.*

*Proof.* Suppose $T \nvdash$ "$\Gamma \models_\delta \varphi$" and fix a model $\mathcal{M}$ of $T$ witnessing this. Let $t(\mathcal{M})$ be the model of $S$ obtained by the $M_n$-preserving interpretation. Then working in $\mathcal{M}$ we see that for all $n \in \omega$ there is some $N \in \gamma_n$ such that $N \models \Gamma \cup \{\neg\varphi\}$. We aim to show this also holds in $t(\mathcal{M})$. Working in $\mathcal{M}$, let $n \in \omega$. Then the statement

$$\exists N(N \in \gamma_n \ \wedge \ N \models \Gamma \cup \{\neg\varphi\})$$

is $\Sigma^1_k$ for some $k > n$.[100] Then since $t(\mathcal{M})$ preserves $M_i$ for all $i \in \omega$, it can be seen by Lemma 59 (2) that the statement above is also preserved from $\mathcal{M}$ to $t(\mathcal{M})$ and so $S \nvdash$ "$\Gamma \models_\delta \varphi$". The rest of the cases are similar. □

This gives us a natural generalization of Theorem 55 and Corollary 60 to the limit of the $\gamma_n$-logics, which in effect, give us agreement on the entire theory of analysis. Here then is a pair of theories that mutually interpret each other with sufficient structural preservation that they agree on $\delta$-logic.

**Example 64.** Consider the following theories:

- $ZFC$ plus determinacy of all sets of reals in $L(\mathbb{R})$; and
- $ZFC$ plus there are infinitely many Woodin cardinals.

Woodin showed that these theories are equiconsistent. Moreover, the interpretations witnessing this preserve the $M_n$s and their iterability. Thus, they completely agree on $\delta$-logic and the theory of analysis.

At this point, we have now developed a strategy for obtaining a large amount of common ground. Moreover for some readers, agreement on the theory of the real numbers may all the common ground they need. Nonetheless, there are many more steps that could be described with the help of inner model theory and, in particular, the core model induction. Rather than explore this terrain in great depth, we now aim to offer a more general perspective on what lies ahead.

[100] This can be made more precise but it is not necessary for the claim.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

4.2.4. *Typical $X$-logics.* We would like to take a more general attitude to the kinds of closure we saw above. Recall with $\gamma$-logic we sought models that were closed under the operation of generating mice for sets. Given that such mice can be easily represented as reals, we might re-imagine these closure properties as closure properties with respect to particularly useful sets of reals. We can then use models closed in this way to give us new $X$-logics. This sets us up with a couple of problems. First, we must find a suitable conception of what constitutes a *useful set of reals*; and second, we must say what it means for a model to be *closed* with regard to such a set. In response to the first problem, we turn to the universally Baire sets. We give a quick overview of these sets and then describe a way of building $X$-logics with them. Recall that given a $T$ tree on $\omega \times \gamma$, the *projection* of $T$, denoted $p[T]$ is the set of $x \in \omega^\omega$ such that there is some $f : \omega \to \gamma$ where for all $n \in \omega$, $\langle x \restriction n, f \restriction n \rangle \in T$. For an example, every $\Sigma^1_1$ set is the projection of a recursive tree $T$ on $\omega \times \omega$.[101] More generally, these projections give us convenient and robust representations for sets of reals whose extension may change depending on the context of the background universe.

**Definition 65.** [Feng et al., 1992] Let us say that a $A \subseteq \mathbb{R}$ is *$\kappa$-universally Baire* if there exists trees $T$ and $S$ on $\omega \times \gamma$ for some $\gamma$ which are such that:

(1) $A = p[T] = \mathbb{R} \backslash p[S]$; and
(2) $\Vdash_{\mathbb{P}} \mathbb{R} = p[T] \cup p[S]$, whenever $\mathbb{P}$ has cardinality $< \kappa$ .

We say that $A$ is *universally Baire* if it is $\kappa$-universally Baire for all $\kappa$.

Informally speaking, being universally Baire generalizes the property of being $\underset{\sim}{\Pi}^1_1$. Recall that a $\underset{\sim}{\Pi}^1_1$ set $A \subseteq \mathbb{R}$ can be represented as the projection of a tree on $\omega \times \omega_1$ and its complement can be represented by the projection of a tree on $\omega \times \omega$. The absoluteness of well-foundedness then gives us a way of identifying natural counterparts to such sets in other models like generic extensions. The property of being universally Baire is intended to abstract out the essentials of this kind of representation. Thus, we may use the trees $T$ and $S$ to identify versions of $A$ in generic extensions of the universe. With sufficient large cardinals, more point classes of reals become universally Baire. For example, if every set has a sharp, the $\underset{\sim}{\Pi}^1_2$-sets become universally Baire.

Given a universally Baire set $A$ as witnessed by trees $T$ and $S$, we now describe a way of representing $A$ in generic extensions $V[G]$ of the universe $V$. Of course, the extension of $p[T]$ will generally not still be $A$ in $V[G]$ as additional reals could fall into the projection of $T$. Thus, it is convenient to adopt a name $A_G$ that is intended to denote $p[T]^{V[G]}$. For this purpose we let $A_G$ be the set of those $x \in \mathbb{R}^{V[G]}$ such that there is some $T \in V$ for which $A = p[T]^V$ and $x \in p[T]^{V[G]}$.[102] So we use the trees from the ground model that represent $A$ to represent its counterpart in the generic extension. In particular, if $T$ and $S$ witness that $A$ is universally Baire and $G$ is $V$-generic, then $A_G = p[T]^{V[G]}$. Next we offer an appropriate notion of *closure* that a model can have with respect to a universally Baire set. There are a number of similar notions on the market, but this one is the most common.[103]

[101] See [Moschovakis, 1980] for more detail.

[102] See section 2.1 of [Bagaria et al., 2006] for more detail.

[103] For some examples: strong closure is discussed in [Bagaria et al., 2006]; related notions of correctness and closure are discussed in [Farah, 2020]; and related notions of faithfulness and invariance are discussed in [Koellner, 2010].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Definition 66.** (Woodin) Say that a countable transitive model $M$ is $A$-*closed* if for all $\mathbb{P} \in M$ and $V$-generic $G$ for $\mathbb{P}$

$$V[G] \models M[G] \cap A_G \in M[G].$$

The essential idea here is that an $A$-closed model $M$ is able to correctly identify its overlap with $A$ even as $A$ is extended generically. For an example, if every real has a sharp and $N \in \gamma$, i.e., a $\gamma$-logic model, then $N$ is $A$-closed where $A = \{\langle i, x\rangle \mid i \in x^{\#}\}$ is a set designed to represent the function $x \mapsto x^{\#}$.[104] With this in hand, we may then give a very general definition of an $X$-logic that is faithful to some universally Baire set and which also generalize $\gamma$-logic.

**Definition 67.** Given ub $A \subseteq \mathbb{R}$, let us say that $\varphi$ is an $A$-consequence of $\Gamma$, $\Gamma \models_A \varphi$, if every $A$-closed model $M$, if $M \models \Gamma$, then $M \models \varphi$.

Thus, the target structures for our $X$-logics are the $A$-closed models for a universally Baire $A \subseteq \mathbb{R}$. In the presence of a sharp for every real, we see that $\gamma$-logic is an example of what we might call an $A$-*closed logic*. Our goal now is to extend the template of Theorems 48, 55 and 63 by showing that if we preserve something like inner model structure, we end up agreeing on $A$-closed logics. Our problem then is to identify a suitable kind of structure to preserve. To this end, let us first observe that $A$-closed logics are robust under forcing.

**Lemma 68.** *Suppose $A \subseteq \mathbb{R}$ is universally Baire. Then for all $\mathbb{P}$ and $\mathbb{P}$-generic $G$ over $V$*

$$\Gamma \models_A \varphi \;\Leftrightarrow\; V[G] \models \text{“}\Gamma \models_{A_G} \varphi.\text{”}$$

*Proof.* ($\Rightarrow$) See the proof of Theorem 2.35 in [Bagaria et al., 2006]. ($\Leftarrow$) Suppose $\Gamma \not\models_A \varphi$ and fix an $A$-closed model $M$ such that $M \models \Gamma \cup \{\neg\varphi\}$ witnessing this. We then claim that $M$ is $A_G$-closed in $V[G]$ whenever $G$ is $\mathbb{P}$-generic over $V$. To see this fix trees $T$ and $S$ witnessing that $A$ is universally Baire. Then suppose that $\mathbb{Q} \in M$ and $H$ is $\mathbb{Q}$-generic over $V[G]$. It suffices to show that

$$V[G \times H] \models M[H] \cap A_{G\times H} \in M[H].$$

To do this, observe that for $x \in \mathbb{R}^{M[H]}$ we have

$$\begin{aligned} x \in A_{G\times H} &\Leftrightarrow x \in p[T]^{V[G\times H]} \\ &\Leftrightarrow (T_x \text{ is ill-founded})^{V[G\times H]} \\ &\Leftrightarrow (T_x \text{ is ill-founded})^{V[H]} \Leftrightarrow x \in p[T]^{V[H]} \Leftrightarrow x \in A_H. \end{aligned}$$

Thus, $M[H] \cap A_{G\times H} = M[H] \cap A_H$ and so the claim follows since $M$ is $A$-closed. Thus, $V[G] \not\models$ “$\Gamma \models_{A_G} \varphi$” as required. □

This gives us a clue to the kind of structure our interpretations should preserve. In the examples above, we preserved an inner model or a collection of such models. In this case, we are going to do something

[104] Here we use the representation $x^{\#}$ instead of $M_0^{\#}$. The former is the theory of $Ult_\omega(M_0^{\#}, U)$ truncated to its critical point in the language $\mathcal{L}_\in(c_n)_{n\in\omega}$ where $c_n$ denotes the critical point of the $n^{th}$ iteration of $M_0^{\#}$. See Fact 5.6 in [Larson, 2011] for a proof of this.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

similar, but less fine-grained: we are going to demand that the interpretations preserve their generic multiverses. Thus, given theories $T$ and $S$, and interpretations

$$t : mod(T) \leftrightarrow mod(S) : s$$

witnessing mutual interpretability, we want $\mathbb{V}_{\mathcal{M}} = \mathbb{V}_{t(\mathcal{M})}$ for models $\mathcal{M}$ of $T$ and $\mathbb{V}_{\mathcal{N}} = \mathbb{V}_{s(\mathcal{N})}$ for models $\mathcal{N}$ of $S$.[105] It is not difficult to see that such interpretations witness that $T$ and $S$ are generic multiverse equivalent.[106]

We should note that this is a much more difficult equivalence relation to satisfy than those we've considered so far along our road. First, as we discussed in Section 2.3.2, generic multiverse equivalence is plausible criterion for saying that entire theories convey the same information, not merely parts of those theories as we've been considering in Sections 3 and 4. To see the effect of this, we recall a pair of theories that are not generic multiverse equivalent but still preserve $\beta$-logic. Let $T$ be $ZFC$ plus the existence of a measurable cardinal. And let $S$ be $ZFC$ plus the existence of a precipitous ideal on $\omega_1$.[107] Then we can generically extend any model of $T$ to obtain a model of $S$; and any model of $S$ has a proper inner model satisfying $T$.[108] These interpretations preserve $L$ and so they completely agree on $\beta$-logic. However, they are not generic multiverse equivalent. To see this note that there are models of $S$ that satisfy $V = L(V_{\omega_1+2})$; and we can (class) generically extended such a model to obtain a model $\mathcal{M}$ of $S$ plus the ground axiom.[109] $\mathcal{M}$ is then the bedrock or core of its generic multiverse, but no model of $T$ can be obtained from $\mathcal{M}$ by generic extension. Thus, $T$ and $S$ are not generic multiverse equivalent. Despite this, there is some reason to doubt its importance. For one thing, the process used to obtain a model of the ground axiom involves endless fiddling with instances of the generalized continuum hypothesis, similar to that used in the argument for Theorem 50. There we worried that our focus had shifted to far away from the concrete. Perhaps we should have similar worries about such models and theories here. For another thing, there is an attitude among some set theorists that the models we really care about are those obtained by generic extension from canonical models. For example, Martin and Steel [1994] optimistically remark with regard to the theory of core models, "We believe that one day the theory will reach models for all the large cardinal hypotheses used by set theorists. This will mean that all of the many models of $ZFC$ they have produced can be built by forcing from core models." Roughly speaking, in the example above, we are invited to focus on the models that can be obtained from models of $T$ by (set) generic extension.[110] With these reservations and replies regarding our *preservation* constraint out of the way, we are almost ready to state our general agreement theorem.

A final tweak is, however, required. Since our discussion in this paper is about comparison between theories rather than structures, we need some way of dealing with universally Baire sets as syntactic items. The following somewhat technical definition is intended to address this.

[105] Recall Definition 28.

[106] Recall Definition 29. I think it's unlikely that the converse holds, although I do not have a proof of this.

[107] Similar remarks should apply to the theories from Example 56, but the interpretations used there are much more complicated to describe.

[108] See Theorem 22.33 in [Jech, 2003].

[109] See Theorem 3.5 in [Reitz, 2007].

[110] A more recent example of this general idea can be found in the introduction to Steel [2022]. With regard to a theory $T$ extending $ZFC$, Steel remarks, "The consistency strength of $T$ is determined by the minimal mouse $M$ have a generic extension satisfying $T$, and thus the consistency strength hierarchy is mirrored in the mouse order."

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Definition 69.** Let us say that a formula $\Phi(x) \in \mathcal{L}_\in$ is *definably universally Baire* if whenever:

- $W$ is a world in the generic multiverse;
- $U$ is a generic refinement of $W$ such that $W = U[G]$; and
- $U[H]$ is a set generic extension of $U$,

then $U$ thinks there is a universally Baire set $A$ such that

$$\Phi^W = A_G \ \& \ \Phi^{U[H]} = A_H$$

where $\Phi^W = \{x \in \mathbb{R} \mid \Phi(x)^W\}$ and similarly for $U[H]$.

Informally speaking, the idea here is provide a definition $\Phi(x)$ that identifies a universally Baire set in a uniform manner across the generic multiverse. We also note that although we have articulated this definition in the language of the generic multiverse, it can also be defined more fussily in the language of set theory. We shall then write $\Gamma \models_\Phi \varphi$ to mean that every real coding a model $\mathcal{M}$ that is $\Phi^{\mathcal{M}}$-closed is such that $\mathcal{M} \models \varphi$ whenever $\mathcal{M} \models \Gamma$. We call this $\Phi$*-logic.* Finally with this in hand, we are ready for the agreement theorem.

**Theorem 70.** *Suppose that $T$ and $S$ are theories extending $ZFC$ that imply that some $\Phi(x)$ is definably universally Baire. Suppose also that $T$ and $S$ are mutually interpretable by interpretations that preserve the generic multiverse. Then $T$ and $S$ completely agree on $\Phi$-logic.*

*Proof.* Suppose $T \nvDash$ "$\Gamma \models_\Phi \varphi$." Fix a countable model $\mathcal{M}$ of $T$ witnessing this and let $\mathbb{V}_\mathcal{M}$ be the generic multiverse of $\mathcal{M}$. Since $T$ and $S$ are generic multiverse equivalent, we may fix some $\mathcal{N}$ in $\mathbb{V}_\mathcal{M}$ such that $\mathcal{N} \models T$. Moreover, it can then be seen that there is some $\mathcal{Q}$ in $\mathbb{V}_\mathcal{N}$ such that $\mathcal{Q}$ is a generic refinement of both $\mathcal{M}$ and $\mathcal{N}$.[111] Then using the definition of definably universally Baire and Lemma 68 twice, we see that both $\mathcal{P}$ and $\mathcal{N}$ satisfy that $\Gamma \models_\Phi \varphi$. Thus $S \nvDash$ "$\Gamma \models_\Phi \varphi$". The other cases are similar. □

Taking a little stock, we've now learned how to obtain a large amount of common ground and laid out a general strategy for obtaining more of it with a general logical prototype for obtaining agreement between strong theories. It would pleasing to set out a few more natural milestones beyond the realm of infinitely many Woodin cardinals, but we leave this for future work. It is also worth observing at this point that the logical framework we have exploited here gives us a nice perspective on the length of the road through the common ground. By contrast and as is well-known, the next obvious jump beyond analysis and up the complexity hierarchy leads us into third order arithmetic and the continuum hypothesis, which is beyond the realm of possible agreement, at least as we currently understand it.[112] This could make it appear as though there is not much common ground beyond second order arithmetic, but the hierarchy of universally Baire sets shows that there is a rich vein of structure to explore and understand.

[111] This follows by a simple induction on the paths linking models in the generic multiverse using Usuba's downward directed grounds theorem [Usuba, 2017].

[112] Nonetheless, it is worth noting Woodin's theorem that if there is a proper class of measurable Woodin cardinals, then any pair of set-generic extensions of the universe that satisfy the continuum hypothesis will agree on all $\Sigma^2_1$ statements. This doesn't fit the logical template offered here, but is certainly remarkable. A proof this can be found in Section 3.2 of [Larson, 2004].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

4.3. **The limits of agreement.** Having offered a way of understanding a typical $X$-logic in our common ground, we come to our final question: where if anywhere does the road end? In this short section, we shall argue that Woodin's $\Omega$-logic stakes out a plausible culmination and delimitation of our common ground project. Ironically, this brings us back to one of the recurring themes of this paper: second order logic. At the beginning of this section, we argued that second order logic in the form of $V$.-logic was too strong to fit into a reasonable conception of common ground. $\Omega$-logic can be understood as responding to this problem by providing a natural weakening of $V$.-logic using the second of our recurring themes: forcing. More precisely, we weaken $V$.-logic by allowing more target structures into the scope of its consequence relation. We do this by considering initial segments of *generic extensions* of the universe and not just of the universe itself. It is defined as follows.

**Definition 71.** We say that $\varphi$ is an $\Omega$-*consequence* of $\Gamma$, $\Gamma \models_\Omega \varphi$, if for all $\mathbb{P}$, $\alpha \in Ord$ and $G$ that is $\mathbb{P}$-generic over $V$, $V_\alpha^{V[G]} \models \varphi$ whenever $V_\alpha^{V[G]} \models \Gamma$.

Outside set theory, such an $X$-logic might seem artificial. However, in close analogy to the case of $V$.-logic, $\Omega$-logic bears a very close relationship to a natural weakening of second order logic defined below.

**Definition 72.** [Ikegami and Väänänen, 2015] Say that a second order sentence $\varphi$ is a *Boolean second order logic consequence* (BSOL) of $\Gamma$, $\Gamma \models_{BSOL} \varphi$, if for all $\mathbb{P} \in V$ and $\mathbb{P}$-generic $G$ over $V$ if $M \in V[G]$ is a full model of second order logic satisfying $\Gamma$, then $M \models \varphi$.

Thus, if we were looking for a counterexample to the claim that $\models_{BSOL} \varphi$, we may seek that model not just in our universe but in any generic extension thereof. Analogously to Theorem 49, we then see that:

**Fact 73.** *(Väänänen & Ikegami, Woodin) The following are Turing equivalent:*[113]

(1) $\{\varphi \mid\ \models_\Omega \varphi\}$;
(2) $\{\varphi \in \Pi_2 \mid \forall \mathbb{P}\ \Vdash_{\mathbb{P}} \varphi\}$; *and*
(3) $\{\varphi \mid\ \models_{BSOL} \varphi\}$.

Thus, we see that $\Omega$-logic can be understood as a very sensible generalization of second order logic for the problems we've seen above. It's also not difficult to see that $\Omega$-logic is weaker than $V$.-logic, but is it plausible that natural theories can agree on it? Can it be part of the common ground? The following theorem gives us an affirmative answer.

**Theorem 74.** *(Essentially Woodin) Suppose $T$ and $S$ are generic multiverse equivalent theories extending $ZFC$ and implying that there is a proper class of Woodin cardinals. The $T$ and $S$ completely agree on $\Omega$-logic.*

So by generically weakening $V$.-logic, we end up with something amenable to the common ground. The following fact is the key to the proof.

[113] The proof is very similar to that of Theorem 49. More details can be found in [Ikegami and Väänänen, 2015].

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

**Fact 75.** *(Woodin)*[114] *Suppose there is a proper class of Woodin cardinals. Then for all $\mathbb{P}$ and $\mathbb{P}$-generic $G$ over $V$*

$$\Gamma \models_\Omega \varphi \Leftrightarrow V[G] \models \text{“}\Gamma \models_\Omega \varphi.\text{”}$$

*Proof.* (of Theorem 74) Suppose $T \nvdash$ “$\Gamma \models_\Omega \varphi$” and fix a model $\mathcal{M}$ of $T$ witnessing this. Then exploiting generic multiverse equivalence, we may fix a model $\mathcal{N} \in \mathbb{V}_{\mathcal{M}}$ satisfying $S$. Now fix $\mathcal{P} \in \mathbb{V}_{\mathcal{M}}$ such that $\mathcal{P}$ is a generic refinement of both $\mathcal{M}$ and $\mathcal{N}$. Then by two applications of Fact 75, we see that $\mathcal{N} \models$ “$\Gamma \not\models_\Omega \varphi$” and so $S \nvdash$ “$\Gamma \models_\Omega \varphi$” as required. The other cases are similar. □

Looking over the loose threads before us, we see that by weakening $V$.-logic to its generic counterpart the possibility of a natural form of agreement returns. This delivers a pleasing rapprochement to a recurring tension in this paper. Time and time again, we have seen second order logic and variations on that theme enforce a kind of rigidity on the possibility of set theoretic alternatives, while forcing and generic extension have offered a much greater degree of flexibility. In Ω-logic, there is a sense in which this tension is resolved through their entanglement. We have an $X$-logic that aims to give as much of the sublime power of second order logic as possible while still remaining within the common ground. Nonetheless, it remains unclear if or how this is related to our discussion of the long road in the previous section. We appear to have a way of agreeing on a very strong $X$-logic, but no story about how this is related to the $X$-logics that were based on models with natural closure properties. This is where Woodin's Ω-conjecture enters our story.

**Conjecture 76.** *(Woodin)*[115] *Suppose there is a proper class of Woodin cardinals. The following are equivalent:*

(1) $\Gamma \models_\Omega \varphi$;
(2) *There is universally Baire $A \subseteq \mathbb{R}$ such that*

$$\Gamma \models_A \varphi.$$

The answer proposed here is clear: Ω-logic is the limit of the $A$-closed logics where $A$ is universally Baire. Like $\delta$-logic it is a limit logic that collects up all the logics below it. Note that only (1→2) is open.[116] However, it is known to be true in all currently known canonical inner models for large cardinals.[117] If the Ω-conjecture is true, we would learn, among other things, that generified second order logic sits at the end of the long road of Section 4.2. We thus have a plausible limit to the kind of agreement we can expect between strong, natural extensions of $ZFC$: Ω-logic. Moreover, we have a conjecture that aims to explain this phenomenon as a generalization of our prototypical example of agreement with $\beta$-logic. We started out with the idea of developing a greater region of partial agreement between set theories. Having observed that second order logic provided a reasonable bound at which point agreement seems to fail, we then took our first steps up through the large cardinal

[114] See Theorem 1.8 in [Bagaria et al., 2006] for a proof.

[115] See [Woodin, 2001, 2004, 2011b, 2012] for more discussion of this conjecture. Note that Woodin thinks of (2) as a kind of proof theory where the universally Baire set $A$ is the proof. Thus, he writes $\Gamma \vdash_\Omega \varphi$. We do not follow him in this practice here.

[116] A proof (2→1) can be found in the proof of Theorem 3.3 in [Bagaria et al., 2006].

[117] See Section 9.4 of [Larson, 2011]. In fact, it suffices for there to be a Woodin cardinal $\delta$ for which the collapse of a countable Skolem hull of $H_{\delta^+}$ has a universally Baire iteration strategy.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

hierarchy and greater levels of agreement. We isolated a reasonably general conception of a strong logic and what is required to agree upon it. And finally we have seen that the limit of that project could bring us back to a synthesis of the two recurring themes of this paper: second order logic and forcing. While there may be other reasonable places to mark the end of agreement, it's difficult to ignore the elegance and uniformity of the account offered above.

**Conclusions.** I think I should thank the readers who have made it this far. I hope it has been a rewarding journey. In the first half of this paper, we considered whether different set theories are really so different after all. Perhaps the proponents of various alternatives would be better understood as talking past each other and really just saying the same thing in different languages. We used interpretability to compare theories and worked to isolate natural equivalence relations between those theories that plausibly explained our dispositions to think of them in some contexts as equivalent. This led us into the realm of the generic multiverse. In the second half of this paper, we restricted our scope and considered some conceptions of partial agreement motivated by the idea maximizing the common ground. We marked out the limits of plausible agreement and then embarked on the long road of increasing agreement and inner model theory culminating in our final destination: Ω-logic.

## References

Peter Aczel. *Non-Well-Founded Sets*. CSLI Lecture Notes, 1988.

Carolin Antos. Class forcing in class theory. In Carolin Antos, Sy-David Friedman, Radek Honzik, and Claudio Ternullo, editors, *The Hyperuniverse Project and Maximality*, pages 1–16. Birkhäuser, 2018.

Joan Bagaria, Neus Castells, and Paul Larson. *An Ω-logic Primer*, pages 1–28. Birkhäuser Basel, Basel, 2006.

Mark Balaguer. A platonist epistemology. *Synthese*, 103(3):303–325, 1995.

Thomas William Barrett. Equivalent and inequivalent formulations of classical mechanics. *British Journal for the Philosophy of Science*, 70(4):1167–1199, 2019.

Thomas William Barrett and Hans Halvorson. Morita equivalencce. *The Review of Symbolic Logic*, 9 (3):556–582, 2016.

J. Barwise and L. Moss. *Vicious Circles: On the Mathematics of Non-Wellfounded Phenomena*. Center for the Study of Language and Information Publication Lecture Notes. Cambridge University Press, 1996.

Tim Button and Sean P. Walsh. *Philosophy and Model Theory*. Oxford, UK: Oxford University Press, 2018.

Keith J. Devlin. *Constructibility*. Springer-Verlag, Berlin, 1984.

H.-D Ebbinghaus. Extended logics: The general framework. In J. Barwise and S. Feferman, editors, *Model-Theoretic Logics*, pages 25–76. Springer, 1985.

Ali Enayat. Variations on a Visserian theme. In Jan van Eijck, Rosalie Iemhoff, and Joost J. Joosten, editors, *Liber Amicorum Alberti: a tribute to Albert Visser*, pages 99–110. College Publications, 2016.

Ali Enayat and Mateusz Łełyk. Categoricity-like properties in the first order realm. *Journal for the Philosophy of Mathematics*, 1:63–98, Sep. 2024.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

Ilijas Farah. *The extender algebra and $\Sigma^2_1$-absoluteness*, volume 4 of *Lecture Notes in Logic*, pages 141–176. Cambridge University Press, 2020.

S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49(1):35–92, 1960.

Solomon Feferman, Harvey M. Friedman, Penelope Maddy, and John R. Steel. Does mathematics need new axioms? *The Bulletin of Symbolic Logic*, 6(4):401–446, 2000.

Qi Feng, Menachem Magidor, and Hugh Woodin. Universally baire sets of reals. In Haim Judah, Winfried Just, and Hugh Woodin, editors, *Set Theory of the Continuum*, pages 203–242, New York, NY, 1992. Springer US.

Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96(10):533–540, 1999.

M. Foreman and A. Kanamori. *Handbook of Set Theory*. Springer Netherlands, 2009.

Marco Forti and Furio Honsell. Set theory with free construction principles. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, Ser. 4, 10(3):493–522, 1983.

Alfredo Roque Freire and Joel David Hamkins. Bi-interpretation in weak set theories, 2020.

D. Guaspari. Partially conservative extensions of arithmetic. *Transactions of the American Mathematical Society*, 254:47–68, 1979.

J. D. Hamkins and D. E. Seabold. Well-founded Boolean ultrapowers as large cardinal embeddings. *ArXiv e-prints*, June 2012.

Joel David Hamkins. The set-theoretic multiverse. *The Review of Symbolic Logic*, 5:416–449, 2012.

Joel David Hamkins. *Lectures on the Philosophy of Mathematics*. Cambridge, Massachusetts: The MIT Press, 2020.

Wilfrid Hodges. *A Shorter Model Theory*. CUP, Cambridge, 1997.

Daisuke Ikegami and Jouko Väänänen. Boolean-valued second-order logic. *Notre Dame Journal of Formal Logic*, 56(1):167–190, 2015.

Thomas Jech. *Set Theory*. Springer, Heidelberg, 2003.

T.J. Jech. *The Axiom of Choice*. Dover Books on Mathematics Series. Dover Publications, 2008.

A. Kanamori. *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*. Springer, 2003.

Peter Koellner. Strong logics of first and second order. *Bulletin of Symbolic Logic*, 16(1):1–36, 2010.

Kenneth Kunen. *Set Theory: an introduction to independence proofs*. Elsevier, Sydney, 2006.

Paul B. Larson. Three days of $\omega$-logic. *Annals of the Japan Association for Philosophy of Science*, 19: 57–86, 2011.

P.B. Larson. *The Stationary Tower: Notes on a Course by W. Hugh Woodin*. University lecture series. American Mathematical Soc., 2004.

Richard Laver. Certain very large cardinals are not created in small forcing extensions. *Annals of Pure and Applied Logic*, 149(1):1–6, 2007.

P. Lindström. *Aspects of Incompleteness: Lecture Notes in Logic 10*. Lecture notes in logic. Taylor & Francis, 2003.

Per Lindström. Some results on interpretability. In B. H. Mayoh F. V. Jensen and K. K. Møller, editors, *Proceedings of the 5th Scandinavian Logic Symposium*, pages 329–361. Aalborg University Press, 1979.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

Per Lindström. On faithful interpretability. In *Computation and Proof Theory*, volume 1104 of *Springer Lecture Notes in Mathematics*, pages 279–288. Springer, 1984.

P. Maddy. *Naturalism in Mathematics*. Oxford Scholarship Online. Philosophy module. Clarendon Press, 1997.

Penelope Maddy. Set-theoretic foundations. American Mathematical Society, 2016.

Penelope Maddy and Toby Meadows. A reconstruction of Steel's multiverse project. *Bulletin of Symbolic Logic*, 26(2):118–169, 2020.

Penelope Maddy and Jouko Väänänen. *Philosophical Uses of Categoricity Arguments*. Elements in the Philosophy of Mathematics. Cambridge University Press, 2023.

D. A. Martin and J. R. Steel. Iteration trees. *Journal of the American Mathematical Society*, 7(1): 1–73, 1994.

Toby Meadows. What can a categoricity theorem tell us? *The Review of Symbolic Logic*, 6(3):524–544, 2013.

Toby Meadows. Two arguments against the generic multiverse. *Review of Symbolic Logic*, 14(2): 347–379, 2021.

Toby Meadows. What is a restrictive theory? *The Review of Symbolic Logic*, pages 1–39, 2022. doi: 10.1017/S1755020322000181.

Toby Meadows. Forcing revisited. *Mathematical Logic Quarterly*, 69(3):287–340, 2023a.

Toby Meadows. Beyond linguistic interpretation. *Review of Symbolic Logic*, 2023b. doi: 10.1017/S1755020323000321.

Toby Meadows. Risk and theoretical equivalence in mathematical foundations. *Synthese*, 202, 11 2023c. doi: 10.1007/s11229-023-04374-1.

Y.N. Moschovakis. *Descriptive Set Theory*. North Holland, 1980.

Andrzej Mostowski. Recent results in set theory. In Imre Lakatos, editor, *Problems in the Philosophy of Mathematics*. North Holland Publishing Company, Amsterdam, 1967.

Sandra Müller, Ralf Schindler, and W. Hugh Woodin. Mice with finitely many woodin cardinals from optimal determinacy hypotheses. *Journal of Mathematical Logic*, 20(Supp01):1950013, 2020.

Itay Neeman. Optimal proofs of determinacy. *Bulletin of Symbolic Logic*, 1(3):327–339, 1995.

Charles Parsons. The uniqueness of the natural numbers. *Iyyun: The Jerusalem Philosophical Quarterly*, 39:13–44, 1990.

Jonas Reitz. The ground axiom. *Journal of Symbolic Logic*, 72(4):1299–1317, 2007.

Ernest Schimmerling. The abc's of mice. *Bulletin of Symbolic Logic*, 7(4):485–503, 2001.

Stewart Shapiro. *Foundations without Foundationalism: a case for second order logic*. OUP, Oxford, 1991.

Saharon Shelah. The future of set theory. *Israel Mathematical Conference Proceedings*, 6, 1991.

J. C. Shepherdson. Inner models for set theory - part I. *Journal of Symbolic Logic*, 16(3):161–190, 1951.

Stephen G. Simpson. *Subsystems of Second Order Arithmetic*. Springer, Berlin, 1999.

Robert M. Solovay. Strongly compact cardinals and the GCH. In *Proceedings of the Tarski Symposium, University of California*, pages 365–372. American Matehematical Society, 1974.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press

J. R. Steel. Projectively well-ordered inner models. *Annals of Pure and Applied Logic*, 74(1):77–104, 1995.

John Steel. Generically invariant set theory. *unpublished manuscript*, 202?

John R. Steel. Generic absoluteness and the continuum problem. *unpublished handout*, 2004.

John R. Steel. The triple helix. *unpublished presentation slides*, 2010.

John R. Steel. Gödel's program. In Juliette Kennedy, editor, *Interpreting Gödel: Critical Essays*. Cambridge University Press, 2014.

J.R. Steel. *A Comparison Process for Mouse Pairs*. Lecture Notes in Logic. Cambridge University Press, 2022.

Alfred Tarski. I: A general method in proofs of undecidability. In Alfred Tarski, editor, *Undecidable Theories*, volume 13 of *Studies in Logic and the Foundations of Mathematics*, pages 1–34. Elsevier, 1953.

T. Usuba. The downward directed grounds hypothesis and very large cardinals. *ArXiv e-prints*, July 2017.

Toshimichi Usuba. The downward directed grounds hypothesis and very large cardinals. *Journal of Mathematical Logic*, 17(02):1750009, 2017.

Jouko Väänänen. An extension of a theorem of Zermelo. *The Bulletin of Symbolic Logic*, 25(2):208–212, 2019.

Albert Visser. Categories of theories and interpretations. In Ali Enayat, Iraj Kalantari, and Mojtaba Moniri, editors, *Logic in Tehran. Proceedings of the workshop and conference on Logic, Algebra and Arithmetic, held October 18-22, 2003*, volume 26, pages 284–341. ASL, Wellesley, Mass., 2006.

Albert Visser and Harvey M. Friedman. When bi-interpretability implies synonymy. *Logic Group preprint series*, 320, 2014.

W. Woodin. The continuum hypothesis, part I. *Notices of the AMS*, 48(6):567–576, 2001.

W. Hugh Woodin. Set theory after russell; the journey back to eden. In G. Link, editor, *100 Years of Russell's Paradox*. De Gruyter, 2004.

W. Hugh Woodin. The realm of the infinite. In Michael Heller and W. Hugh Woodin, editors, *Infinity: New Research Frontiers*. Cambridge University Press, Cambridge, 2011a.

W. Hugh Woodin. *The Realm of the Infinite*. Cambridge University Press, 2011b.

W. Hugh Woodin. *The Continuum Hypothesis, the Generic Multiverse of Sets, and the $\Omega$ Conjecture*. Cambridge University Press, 2012.

Ernst Zermelo. On boundary numbers and domains of sets: new investigations in the foundations of set theory. In *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*. Oxford University Press, 1976.

https://doi.org/10.1017/bsl.2025.10093 Published online by Cambridge University Press