Fast-collapsing theories

Samuel A. Alexander[*]
Department of Mathematics, the Ohio State University

November 12, 2013

Abstract

Reinhardt's conjecture, a formalization of the statement that a truthful knowing machine can know its own truthfulness and mechanicalness, was proved by Carlson using sophisticated structural results about the ordinals and transfinite induction just beyond the first epsilon number. We prove a weaker version of the conjecture, by elementary methods and transfinite induction up to a smaller ordinal.

1 Introduction

This is a paper about idealized truthful mechanical knowing agents who know facts in a quantified arithmetic-based language that also includes a connective for their own knowledge ($K(1+1=2)$ is read "I (the agent) know $1+1=2$"). It is well known ([4], [6], [9], [10], [11], [12]) that such an agent cannot simultaneously know its own truthfulness and its own code. Reinhardt conjectured that, while knowing its own truthfulness, such a machine can know it has *some* code, without knowing which. This conjecture was proved by Carlson [6]. The proof uses sophisticated structural results from [5] about the ordinals, and involves transfinite induction up to $\epsilon_0 \cdot \omega$.

We will give a proof of a weaker result, but will do so in an elementary way, inducting only as far as $\omega \cdot \omega$. Along the way, we will develop some machinery that is interesting in its own right. Carlson's proof of Reinhardt's conjecture is based on stratifying knowledge (see [8] for a gentle summary). This can be viewed as adding operators K^α for knowledge after time α where α takes ordinal values. Under certain assumptions, theories in such stratified language *collapse* at positive integer multiples of ϵ_0, in the sense that if ϕ only contains superscripts $< \epsilon_0 \cdot n$ (n a positive integer) then $K^{\epsilon_0 \cdot n}\phi$ holds if and only if $K^{\epsilon_0 \cdot (n+1)}\phi$ does. In this paper, collapse occurs at positive integer multiples of ω, hence the name: *Fast-collapsing theories*.

Our result is weakened in the sense that the background theory of knowledge is weakened. The schema $K(\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi))$ (ucl denotes universal closure) is weakened by adding the requirement that K not be nested deeper in ϕ than in ψ (the unrestricted schema $\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi)$ is preserved, but the knower is not required to *know* it); the schema $\mathrm{ucl}(K\phi \rightarrow KK\phi)$ is forfeited entirely; and a technical axiom called Assigned Validity (made up of valid formulas with numerals plugged in to their free variables) is added to the background theory of knowledge.

On the bright side, our result is stated in a more general way (we mention in passing how the full unweakened result could also be so generalized, but leave those details for later work). Casually, our main theorem has the following form:

> A truthful knowing agent whose knowledge is sufficiently "generic" can be taught its own truthfulness and still remain truthful.

Here "generic" is a specific technical term, but it is inclusive enough to include knowledge that one has some code, thus the statement addresses Reinhardt's conjecture.

In Section 2 we present some preliminaries.

In Section 3 we develop *stratifiers*, maps from unstratified language to stratified language. These are the key to fast collapse. They debuted in [1] and [3].

[*] Email: alexander@math.ohio-state.edu

In Section 4 we discuss *uniform* stratified theories. A key advantage of stratifiers is that they turn unstratified theories into uniform stratified theories.

In Section 5 we define some notions of genericity of an axiom schema, and establish the genericity of some building blocks of background theories of knowledge.

In Section 6 we state our main theorem and make closing remarks.

2 Preliminaries

Definition 1. (Standard Definitions) Let $\mathscr{L}_{\mathrm{PA}}$ be the language $(0, S, +, \cdot)$ of Peano arithmetic and let $\mathscr{L}$ be an arbitrary language.

1. For any $e \in \mathbb{N}$, W_e is the range of the eth partial computable function. The binary predicate $\bullet \in W_\bullet$ is $\mathscr{L}_{\mathrm{PA}}$-definable so we will freely act as if $\mathscr{L}_{\mathrm{PA}}$ actually contains this predicate symbol.

2. If an $\mathscr{L}$-structure $\mathscr{M}$ is clear from context, an *assignment* is a function taking variables into the universe of $\mathscr{M}$.

3. If s is an assignment, x is a variable, and $a \in \mathscr{M}$, $s(x|a)$ is the assignment that agrees with s except that $s(x|a)(x) = a$.

4. We define $\mathscr{L}_{\mathrm{PA}}$-terms $\overline{n}$ ($n \in \mathbb{N}$), called *numerals*, so that $\overline{0} = 0$ and $\overline{n+1} = S(\overline{n})$.

5. If ϕ is an $\mathscr{L}$-formula, $\mathrm{FV}(\phi)$ is the set of free variables of ϕ. If $\mathrm{FV}(\phi) = \emptyset$ then ϕ is a *sentence*.

6. If ϕ is an $\mathscr{L}$-formula, x is variable, and u is an $\mathscr{L}$-term, $\phi(x|u)$ is the result of substituting u for all free occurrences of x in ϕ.

7. A *universal closure* of an $\mathscr{L}$-formula ϕ is a sentence $\forall x_1 \cdots \forall x_n \phi$. We write $\mathrm{ucl}(\phi)$ to denote a universal closure of ϕ.

8. We use the word *theory* as synonym for *set of sentences*.

9. If T is an $\mathscr{L}$-theory and $\mathscr{M}$ is an $\mathscr{L}$-structure, $\mathscr{M} \models T$ means that $\mathscr{M} \models \phi$ for all $\phi \in T$.

10. If T is an $\mathscr{L}$-theory, we say $T \models \phi$ if $\mathscr{M} \models \phi$ whenever $\mathscr{M} \models T$.

11. A *valid* $\mathscr{L}$-formula is one that holds in every $\mathscr{L}$-structure.

12. For any formulas ϕ_1, ϕ_2, ϕ_3, we write $\phi_1 \rightarrow \phi_2 \rightarrow \phi_3$ to abbreviate $\phi_1 \rightarrow (\phi_2 \rightarrow \phi_3)$.

We will repeatedly use the following standard fact without explicit mention: if ψ is a universal closure of ϕ, then in order to prove $\mathscr{M} \models \psi$, it suffices to let s be an arbitrary assignment and show that $\mathscr{M} \models \phi[s]$.

For quantified semantics we work in Carlson's *base logic*, defined as follows.

Definition 2. (The Base Logic) A *language $\mathscr{L}$ in the base logic* is a first-order language $\mathscr{L}_0$ together with a set of symbols called *operators*. Formulas of $\mathscr{L}$ are defined in the usual way, with the clause that whenever ϕ is an $\mathscr{L}$-formula and K is an $\mathscr{L}$-operator, $K\phi$ is also an $\mathscr{L}$-formula (and $\mathrm{FV}(K\phi) = \mathrm{FV}(\phi)$). Syntactic parts of Definition 1 extend to the base logic in obvious ways. Given such an $\mathscr{L}$, an *$\mathscr{L}$-structure $\mathscr{M}$* is a first-order $\mathscr{L}_0$-structure $\mathscr{M}_0$ together with a function that takes one $\mathscr{L}$-formula ϕ, one $\mathscr{L}$-operator K, and one assignment s, and outputs True or False—in which case we write $\mathscr{M} \models K\phi[s]$ or $\mathscr{M} \not\models K\phi[s]$, respectively—satisfying the following three conditions (where ϕ ranges over $\mathscr{L}$-formulas and K ranges over operators):

1. Whether or not $\mathscr{M} \models K\phi[s]$ is independent of $s(x)$ if $x \notin \mathrm{FV}(\phi)$.

2. (Alphabetic Invariance) If ψ is an *alphabetic variant* of ϕ, meaning that it is obtained from ϕ by renaming bound variables while respecting binding of the quantifiers, then $\mathscr{M} \models K(\phi)[s]$ if and only if $\mathscr{M} \models K(\psi)[s]$.

3. (Weak Substitution)[1] If the variable y is substitutable for the variable x in ϕ, then $\mathscr{M} \models K\phi(x|y)[s]$ if and only if $\mathscr{M} \models K\phi[s(x|s(y))]$.

Theorem 3. (Completeness and compactness) Let $\mathscr{L}$ be an r.e. language in the base logic.

1. The set of valid $\mathscr{L}$-formulas is r.e.

2. For any r.e. $\mathscr{L}$-theory T, $\{\phi : T \models \phi\}$ is r.e.

3. There is an effective algorithm, given (a Gödel number for) an r.e. $\mathscr{L}$-theory T, to find (a Gödel number for) $\{\phi : T \models \phi\}$.

4. If T is an $\mathscr{L}$-theory and $T \models \phi$ (ϕ any $\mathscr{L}$-formula), there are $\tau_1, \ldots, \tau_n \in T$ such that $(\bigwedge_i \tau_i) \rightarrow \phi$ is valid.

Proof. By interpreting the base logic in first-order logic. For details, see [1]. □

Definition 4. Let $\mathscr{L}_{\mathrm{EA}}$ be the language of Epistemic Arithmetic from [13], so $\mathscr{L}_{\mathrm{EA}}$ extends $\mathscr{L}_{\mathrm{PA}}$ by a unary operator K. An $\mathscr{L}_{\mathrm{EA}}$-structure (more generally an $\mathscr{L}$-structure where $\mathscr{L}$ extends $\mathscr{L}_{\mathrm{PA}}$) has *standard first-order part* if its first-order part has universe $\mathbb{N}$ and interprets $0, S, +, \cdot$ in the intended ways.

Definition 5. Suppose $\mathscr{L}$ extends $\mathscr{L}_{\mathrm{PA}}$ and ϕ is an $\mathscr{L}$-formula with $\mathrm{FV}(\phi) \subseteq \{x_1, \ldots, x_n\}$. For any assignment s into $\mathbb{N}$, we define

$$\phi^s \equiv \phi(x_1|\overline{s(x_1)}) \cdots (x_n|\overline{s(x_n)}),$$

the sentence obtained by replacing all free variables in ϕ by numerals according to s.

Definition 6. For any $\mathscr{L}_{\mathrm{EA}}$-theory T, the intended structure for T is the $\mathscr{L}_{\mathrm{EA}}$-structure $\mathscr{N}_T$ that has standard first-order part and interprets K so that for any $\mathscr{L}_{\mathrm{EA}}$-formula ϕ and assignment s,

$$\mathscr{N}_T \models K\phi[s] \text{ if and only if } T \models \phi^s.$$

We say T is *true* if $\mathscr{N}_T \models T$.

It is easy to check that the structures $\mathscr{N}_T$ of Definition 6 really are $\mathscr{L}_{\mathrm{EA}}$-structures (they satisfy Conditions 1–3 of Definition 2). The following lemma shows that they accurately interpret quantified formulas in the way one would expect.

Lemma 7. For any $\mathscr{L}_{\mathrm{EA}}$-theory T, $\mathscr{L}_{\mathrm{EA}}$-formula ϕ and assignment s,

$$\mathscr{N}_T \models \phi[s] \text{ if and only if } \mathscr{N}_T \models \phi^s.$$

Proof. Straightforward induction. □

Armed with these definitions, we can make more precise some things we suggested in the introduction. Let T_{SMT} be the following $\mathscr{L}_{\mathrm{EA}}$-theory ($\phi$ and ψ range over $\mathscr{L}_{\mathrm{EA}}$-formulas):

1. (E_1) $\mathrm{ucl}(K\phi)$ whenever ϕ is valid.

2. (E_2) $\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi)$.

3. (E_3) $\mathrm{ucl}(K\phi \rightarrow \phi)$.

4. (E_4) $\mathrm{ucl}(K\phi \rightarrow KK\phi)$.

5. The *axioms of Epistemic Arithmetic*, by which we mean the axioms of Peano Arithmetic with the induction schema extended to $\mathscr{L}_{\mathrm{EA}}$.

6. (Mechanicalness) $\mathrm{ucl}(\exists e \forall x (K\phi \leftrightarrow x \in W_e))$ provided $e \notin \mathrm{FV}(\phi)$.

7. $K\phi$ whenever ϕ is an instance of lines 1–6 or (recursively) 7.

[1] Note that the general substitution law, where y is replaced by an arbitrary term, is not valid in modal logic.

Combining lines 6 and 7 yields the *Strong Mechanistic Thesis*, $K(\mathrm{ucl}(\exists e\forall x(K\phi \leftrightarrow x \in W_e)))$. One of the main results of [6] is that T_{SMT} is true, that is, $\mathscr{N}_{T_{\mathrm{SMT}}} \models T_{\mathrm{SMT}}$. To establish $\mathscr{N}_{T_{\mathrm{SMT}}} \models E_3$, Carlson uses transfinite recursion up to $\epsilon_0 \cdot \omega$, as well as deep structural properties (from [5]) about the ordinals. That $\mathscr{N}_{T_{\mathrm{SMT}}}$ satisfies lines 2, 5, 6, and 7, is trivial; that it satisfies line 4 follows from the fact that it satisfies lines 1–2. Line 1 would be trivial if we added the following line to T_{SMT}:

1b. (Assigned Validity) ϕ^s, whenever ϕ is valid and s is any assignment.

Theorems from [6] imply Assigned Validity is already a consequence of T_{SMT}, so this addition is not necessary, however it becomes necessary if (say) line 2 is weakened.

The main result in this paper is that by weakening E_2, removing E_4, and adding Assigned Validity, we remove the need to induct up to $\epsilon_0 \cdot \omega$. Induction up to $\omega \cdot \omega$ suffices, and the computations from [5] can also be avoided. This is surprising because we do not weaken E_3, the lone schema for which such sophisticated methods were used before.

Definition 8. For any $\mathscr{L}_{\mathrm{EA}}$-formula ϕ, let $\mathrm{depth}(\phi)$ denote the depth to which K operators are nested in ϕ, more formally:

- If ϕ is an $\mathscr{L}_{\mathrm{PA}}$-formula then $\mathrm{depth}(\phi) = 0$.
- If $\phi \equiv K(\phi_0)$ then $\mathrm{depth}(\phi) = \mathrm{depth}(\phi_0) + 1$.
- If $\phi \equiv (\rho \rightarrow \sigma)$ then $\mathrm{depth}(\phi) = \max\{\mathrm{depth}(\rho), \mathrm{depth}(\sigma)\}$.
- If $\phi \in \{(\neg\phi_0), (\forall x\phi_0)\}$ then $\mathrm{depth}(\phi) = \mathrm{depth}(\phi_0)$.

Now let T^w_{SMT} be the $\mathscr{L}_{\mathrm{EA}}$-theory containing the following schemas:

1. E_1 and E_3.
2. Assigned Validity: ϕ^s whenever ϕ is valid and s is any assignment.
3. (E_2') $\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi)$ provided $\mathrm{depth}(\phi) \le \mathrm{depth}(\psi)$.
4. The axioms of Epistemic Arithmetic.
5. Mechanicalness.
6. $K\phi$ whenever ϕ is an instance of lines 1–5 or (recursively) 6.

Our main result (obtained by inducting only up to $\omega \cdot \omega$) will imply T^w_{SMT} is true.

3 Stratifiers

Definition 9. Let $\mathscr{L}_{\omega\cdot\omega}$ be the language obtained from $\mathscr{L}_{\mathrm{PA}}$ by adding operators K^α for all $\alpha \in \omega \cdot \omega$. For any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ, let

$$\mathrm{On}(\phi) = \{\alpha \in \omega \cdot \omega \,:\, K^\alpha \text{ occurs in } \phi\}.$$

An example of an $\mathscr{L}_{\omega\cdot\omega}$-formula: $\forall x(K^\omega K^{\omega\cdot 7+2} K^{53} K^0 (x = 0) \rightarrow K^{\omega\cdot 7+3}(x = 0))$.

Definition 10. (Stratifiers) For any infinite subset $X \subseteq \omega \cdot \omega$, the *stratifier given by* X is the function $\bullet^+$ that takes $\mathscr{L}_{\mathrm{EA}}$-formulas to $\mathscr{L}_{\omega\cdot\omega}$-formulas in the following way.

1. If ϕ is atomic, $\phi^+ \equiv \phi$.
2. If ϕ is $\phi_1 \rightarrow \phi_2$, $\neg\phi_1$, or $\forall x\phi_1$, then ϕ^+ is $\phi_1^+ \rightarrow \phi_2^+$, $\neg\phi_1^+$, or $\forall x\phi_1^+$, respectively.
3. If ϕ is $K\phi_0$, then $\phi^+ \equiv K^\alpha \phi_0^+$ where α is the smallest ordinal in $X \backslash \mathrm{On}(\phi_0^+)$.

By a *stratifier*, we mean a stratifier given by some X. By the *veristratifier*, we mean the stratifier given by $X = \{\omega \cdot 1, \omega \cdot 2, \ldots\}$. If $\bullet^+$ is a stratifier and T is an $\mathscr{L}_{\mathrm{EA}}$-theory, T^+ denotes $\{\phi^+ \,:\, \phi \in T\}$.

For example, if $\bullet^+$ is the veristratifier, then

$$(K(1=0) \rightarrow KK(1=0))^+ \equiv K^{\omega}(1=0) \rightarrow K^{\omega\cdot 2}K^{\omega}(1=0).$$

Lemma 11. Suppose ϕ is an $\mathscr{L}_{\text{EA}}$-formula, s is an assignment into $\mathbb{N}$, and $\bullet^+$ is a stratifier. If $\alpha, \beta \in \omega \cdot \omega$ are such that $(K\phi)^+ \equiv K^{\alpha}\phi^+$ and $(K\phi^s)^+ \equiv K^{\beta}(\phi^s)^+$, then $\alpha = \beta$.

Proof. By induction. □

Lemma 12. Suppose ϕ and ψ are $\mathscr{L}_{\text{EA}}$-formulas and $\bullet^+$ is a stratifier. Let $\alpha, \beta \in \omega \cdot \omega$ be such that $(K\phi)^+ \equiv K^{\alpha}\phi^+$ and $(K\psi)^+ \equiv K^{\beta}\psi^+$. Then $\text{depth}(\phi) < \text{depth}(\psi)$ if and only if $\alpha < \beta$.

Proof. By induction. □

Definition 13. For any $\mathscr{L}_{\omega\cdot\omega}$-structure $\mathscr{M}$ and stratifier $\bullet^+$, let $\mathscr{M}^+$ be the $\mathscr{L}_{\text{EA}}$-structure that has the same universe and interpretation of $\mathscr{L}_{\text{PA}}$ as $\mathscr{M}$, and that interprets K so that for any $\mathscr{L}_{\text{EA}}$-formula ϕ and assignment s,

$$\mathscr{M}^+ \models K\phi[s] \text{ if and only if } \mathscr{M} \models (K\phi)^+[s].$$

It is easy to check that if $\mathscr{M}$ is an $\mathscr{L}_{\omega\cdot\omega}$-structure then $\mathscr{M}^+$ really is an $\mathscr{L}_{\text{EA}}$-structure (it satisfies Conditions 1–3 of Definition 2). From now on we will suppress this remark when defining new structures.

Lemma 14. Let $\mathscr{M}$ be an $\mathscr{L}_{\omega\cdot\omega}$-structure, $\bullet^+$ a stratifier. For any $\mathscr{L}_{\text{EA}}$-formula ϕ and assignment s,

$$\mathscr{M}^+ \models \phi[s] \text{ if and only if } \mathscr{M} \models \phi^+[s].$$

Proof. A straightforward induction. □

Definition 15. For any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ, ϕ^- is the $\mathscr{L}_{\text{EA}}$-formula obtained by changing every operator of the form K^{α} in ϕ into K. If T is an $\mathscr{L}_{\omega\cdot\omega}$-theory, $T^- = \{\phi^- : \phi \in T\}$.

Example 16. $\left(K^{\omega\cdot 8+3}\forall x K^{17}(x=y)\right)^- \equiv K\forall x K(x=y)$.

Lemma 17. Let $\bullet^+$ be a stratifier. For any $\mathscr{L}_{\text{EA}}$-formula ϕ, $(\phi^+)^- \equiv \phi$.

Proof. Straightforward. □

Definition 18. If $\mathscr{M}$ is an $\mathscr{L}_{\text{EA}}$-structure, let $\mathscr{M}^-$ be the $\mathscr{L}_{\omega\cdot\omega}$-structure that has the same universe as $\mathscr{M}$, agrees with $\mathscr{M}$ on $\mathscr{L}_{\text{PA}}$, and interprets each K^{α} so that for any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ and assignment s,

$$\mathscr{M}^- \models K^{\alpha}\phi[s] \text{ if and only if } \mathscr{M} \models K\phi^-[s].$$

In [6] (Definition 5.4), $\mathscr{M}^-$ is the *stratification of* $\mathscr{M}$ *over* $\omega \cdot \omega$.

Lemma 19. For any $\mathscr{L}_{\text{EA}}$-structure $\mathscr{M}$, $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ and assignment s,

$$\mathscr{M}^- \models \phi[s] \text{ if and only if } \mathscr{M} \models \phi^-[s].$$

Proof. A straightforward induction. □

Theorem 20.

1. For any valid $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ, ϕ^- is valid.

2. For any $\mathscr{L}_{\text{EA}}$-formula ϕ and stratifier $\bullet^+$, ϕ is valid if and only if ϕ^+ is valid.

Proof.

(1) Let ϕ be a valid $\mathscr{L}_{\omega\cdot\omega}$-formula. For any $\mathscr{L}_{\text{EA}}$-structure $\mathscr{M}$ and assignment s, since ϕ is valid, $\mathscr{M}^- \models \phi[s]$ and so by Lemma 19, $\mathscr{M} \models \phi^-[s]$. By arbitrariness of $\mathscr{M}$ and s, ϕ^- is valid.

(2, $\Rightarrow$) Assume ϕ is a valid $\mathscr{L}_{\text{EA}}$-formula. For any $\mathscr{L}_{\omega\cdot\omega}$-structure $\mathscr{M}$ and assignment s, since ϕ is valid, $\mathscr{M}^+ \models \phi[s]$, and $\mathscr{M} \models \phi^+[s]$ by Lemma 14. By arbitrariness of $\mathscr{M}$ and s, this shows ϕ^+ is valid.

(2, $\Leftarrow$) Assume ϕ is an $\mathscr{L}_{\mathrm{EA}}$-formula and ϕ^+ is valid. For any $\mathscr{L}_{\mathrm{EA}}$-structure $\mathscr{M}$ and assignment s, since ϕ^+ is valid, $\mathscr{M}^- \models \phi^+[s]$, and $\mathscr{M} \models (\phi^+)^-[s]$ by Lemma 19. By Lemma 17, $\mathscr{M} \models \phi[s]$. By arbitrariness of $\mathscr{M}$ and s, ϕ is valid. □

Definition 21. For any $\mathscr{L}_{\mathrm{EA}}$-theory T, let

$$T^{\oplus} = \{\phi^+ \,:\, \phi \in T \text{ and } \bullet^+ \text{ is a stratifier}\}.$$

Example 22. Suppose T is the $\mathscr{L}_{\mathrm{EA}}$-theory consisting of $K\phi \rightarrow KK\phi$ for all $\mathscr{L}_{\mathrm{PA}}$-sentences ϕ. Then $T^{\oplus}$ is the $\mathscr{L}_{\omega\cdot\omega}$-theory consisting of $K^{\alpha}\phi \rightarrow K^{\beta}K^{\alpha}\phi$ for all $\mathscr{L}_{\mathrm{PA}}$-sentences ϕ and ordinals $\alpha < \beta < \omega \cdot \omega$.

Theorem 23. (Upward proof stratification) For any $\mathscr{L}_{\mathrm{EA}}$-theory T, $\mathscr{L}_{\mathrm{EA}}$-sentence ϕ, and stratifier $\bullet^+$, the following are equivalent.

1. $T \models \phi$.
2. $T^+ \models \phi^+$.
3. $T^{\oplus} \models \phi^+$.

This theorem is so-named because it is an upside-down version of a harder theorem that we called [1] *proof stratification.* In non-upward proof stratification, T and ϕ are taken in the *stratified* language and the theorem states that $T \models \phi$ if and only if $T^- \models \phi^-$. This uses complicated hypotheses on T and ϕ. Versions of these hypotheses could be stated in an elementary way, but *a priori* they might imply T is inconsistent (in which case Theorem 23 is trivial). The only way we know to exhibit consistent theories that satisfy such hypotheses is to exploit the machinery from [5] on the Σ_1-structure of the ordinals.

Proof of Theorem 23. Let T, ϕ and $\bullet^+$ be as in Theorem 23.

(1 $\Rightarrow$ 2) Assume $T \models \phi$. By Theorem 3, there are $\tau_1, \ldots, \tau_n \in T$ such that $(\bigwedge_i \tau_i) \rightarrow \phi$ is valid. By Theorem 20, $\left(\bigwedge_i \tau_i^+\right) \rightarrow \phi^+$ is valid, showing $T^+ \models \phi^+$.

(2 $\Rightarrow$ 3) Trivial: $T^+ \subseteq T^{\oplus}$.

(3 $\Rightarrow$ 1) Assume $T^{\oplus} \models \phi^+$. By Theorem 3 there are $\tau_1, \ldots, \tau_n \in T^{\oplus}$ such that $(\bigwedge_i \tau_i) \rightarrow \phi^+$ is valid. By definition of $T^{\oplus}$ there are $\sigma_1, \ldots, \sigma_n \in T$ and stratifiers $\bullet^1, \ldots, \bullet^n$ such that each $\tau_i \equiv \sigma_i^i$. By Lemma 17

$$\left(\left(\bigwedge_i \sigma_i^i\right) \rightarrow \phi^+\right)^- \equiv \left(\bigwedge_i \sigma_i\right) \rightarrow \phi,$$

so Theorem 20 guarantees $(\bigwedge_i \sigma_i) \rightarrow \phi$ is valid, and $T \models \phi$. □

4 Uniform Theories and Collapsing Knowledge

Definition 24. Suppose $X \subseteq \omega \cdot \omega$ and $h : X \rightarrow \omega \cdot \omega$. For any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ, we define $h(\phi)$ to be the $\mathscr{L}_{\omega\cdot\omega}$-formula obtained by replacing K^{α} by $K^{h(\alpha)}$ everywhere K^{α} occurs in ϕ ($\alpha \in X$). (If $\alpha \notin X$, we do not change occurrences of K^{α} in ϕ.)

Example 25. Suppose $\alpha_1 < \cdots < \alpha_4$ are distinct ordinals in $\omega \cdot \omega$. Let $X = \{\alpha_2, \alpha_3\}$, let $h(\alpha_2) = \alpha_3$, $h(\alpha_3) = \alpha_4$. Then

$$h\left(K^{\alpha_3}K^{\alpha_2}K^{\alpha_1}(1=1)\right) \equiv K^{\alpha_4}K^{\alpha_3}K^{\alpha_1}(1=1).$$

Definition 26. An $\mathscr{L}_{\omega\cdot\omega}$-theory T is *uniform* if the following statement holds. For all $X \subseteq \omega \cdot \omega$, for all order-preserving $h : X \rightarrow \omega \cdot \omega$, for all $\phi \in T$, if $\mathrm{On}(\phi) \subseteq X$ then $h(\phi) \in T$.

Example 27. If T contains $K^1K^0(1=0)$ and T is uniform, then T must contain $K^{\beta}K^{\alpha}(1=0)$ for all $\alpha < \beta \in \omega \cdot \omega$.

Lemma 28. Suppose $\bullet^+$ is a stratifier, $X \subseteq \omega\cdot\omega$, $h : X \rightarrow \omega\cdot\omega$ is order preserving, and ϕ is an $\mathscr{L}_{\mathrm{EA}}$-formula with $\mathrm{On}(\phi^+) \subseteq X$. There is a stratifier $\bullet^*$ such that $\phi^* \equiv h(\phi^+)$.

Proof. Let $Y_0 = \{h(\alpha) \,:\, \alpha \in \mathrm{On}(\phi^+)\}$, $Y = Y_0 \cup \{\beta \in \omega \cdot \omega \,:\, \beta > Y_0\}$, and let $\bullet^*$ be the stratifier given by Y. By induction, for every subformula ϕ_0 of ϕ, $\phi_0^* \equiv h(\phi_0^+)$. □

Lemma 29. (Uniformity lemma) For any $\mathscr{L}_{\mathrm{EA}}$-theory T, $T^{\oplus}$ is uniform.

Proof. Let $X \subseteq \omega \cdot \omega$, let $h : X \to \omega \cdot \omega$ be order preserving, let $\phi \in T^\oplus$, and assume $\mathrm{On}(\phi) \subseteq X$. By definition of $T^\oplus$, $\phi \equiv \phi_0^+$ for some $\phi_0 \in T$ and some stratifier $\bullet^+$. By Lemma 28 there is a stratifier $\bullet^*$ such that $h(\phi_0^+) \equiv \phi_0^*$. This shows $h(\phi) \in T^\oplus$. □

Unfortunately, the range of $\oplus$ does not include every uniform $\mathscr{L}_{\omega\cdot\omega}$-theory. For example, suppose T is the $\mathscr{L}_{\omega\cdot\omega}$-theory consisting of

$$K^\alpha(\phi^+ \to \psi^+) \to K^\alpha\phi^+ \to K^\alpha\psi^+$$

for all $\mathscr{L}_{\mathrm{EA}}$-sentences ϕ and ψ and stratifiers $\bullet^+$ with $\mathrm{On}(\phi^+), \mathrm{On}(\psi^+) < \alpha \in \omega \cdot \omega$. The reader may check that despite being uniform, T is not $T_0^\oplus$ for any $\mathscr{L}_{\mathrm{EA}}$-theory T_0.

Definition 30. If $\mathscr{M}$ is an $\mathscr{L}_{\omega\cdot\omega}$-structure, $X \subseteq \omega \cdot \omega$, and $h : X \to \omega \cdot \omega$, we define an $\mathscr{L}_{\omega\cdot\omega}$-structure $h(\mathscr{M})$ that has the same universe as $\mathscr{M}$, agrees with $\mathscr{M}$ on the interpretation of $\mathscr{L}_{\mathrm{PA}}$, and interprets K^α so that for any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ and assignment s,

$$h(\mathscr{M}) \models K^\alpha\phi[s] \text{ if and only if } \mathscr{M} \models h(K^\alpha\phi)[s].$$

Lemma 31. Suppose $\mathscr{M}$, X, and h are as in Definition 30. For any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ and assignment s,

$$h(\mathscr{M}) \models \phi[s] \text{ if and only if } \mathscr{M} \models h(\phi)[s].$$

Proof. By induction. □

We will only need part 1 of the next lemma, we state part 2 for completeness.

Lemma 32. Suppose $\mathscr{M}$, X, and h are as in Definition 30 and ϕ is an $\mathscr{L}_{\omega\cdot\omega}$-formula.

1. If ϕ is valid then $h(\phi)$ is valid.
2. Assume h is injective. If $\mathrm{On}(\phi) \subseteq X$ and $h(\phi)$ is valid, then ϕ is valid.

Proof.

(1) Similar to Theorem 20.

(2) If $h(\phi)$ is valid then $h^{-1}(h(\phi))$ is valid by part 1. Since $\mathrm{On}(\phi) \subseteq X$, $h^{-1}(h(\phi)) \equiv \phi$. □

Definition 33. For any $\mathscr{L}_{\omega\cdot\omega}$-theory T and $\alpha \in \omega \cdot \omega$, let $T \cap \alpha = \{\phi \in T : \mathrm{On}(\phi) \subseteq \alpha\}$ be the subset of T where all superscripts are strictly bounded by α.

Example 34.

- For any $\mathscr{L}_{\omega\cdot\omega}$-theory T, $T \cap 0 = \{\phi \in T : \phi \text{ is an } \mathscr{L}_{\mathrm{PA}}\text{-sentence}\}$.
- For any $\mathscr{L}_{\omega\cdot\omega}$-theory T, $T \cap 1 = \{\phi \in T : \phi \text{ is an } \mathscr{L}_{\mathrm{PA}} \cup \{K^0\}\text{-sentence}\}$.
- For any $\mathscr{L}_{\mathrm{EA}}$-theory T, $T^\oplus \cap \omega = \{\phi^+ : \phi \in T \text{ and } \bullet^+ \text{ is given by some } X \subseteq \omega\}$.

Theorem 35. (The collapse theorem) Let T be a uniform $\mathscr{L}_{\omega\cdot\omega}$-theory. For any $0 < n \in \mathbb{N}$ and $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ with $\mathrm{On}(\phi) \subseteq \omega \cdot n$, $T \models \phi$ if and only if $T \cap (\omega \cdot n) \models \phi$.

Proof. The $\Leftarrow$ direction is trivial: $T \cap (\omega \cdot n) \subseteq T$. For $\Rightarrow$, assume $T \models \phi$. By Theorem 3 there are $\tau_1, \ldots, \tau_n \in T$ such that

$$\Phi \equiv \left(\textstyle\bigwedge_i \tau_i\right) \to \phi$$

is valid. Let $X = \mathrm{On}(\Phi) \cap (\omega \cdot n)$, $Y = \mathrm{On}(\Phi) \cap [\omega \cdot n, \infty)$, see Fig. 1. Then $|X|, |Y| < \infty$ and $X \cup Y = \mathrm{On}(\Phi)$.

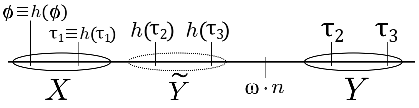

Figure 1: Collapse.

Since $|X| < \infty$ and $\omega \cdot n$ has no maximum element, there are infinitely many ordinals above X in $\omega \cdot n$. Thus since $|Y| < \infty$ we can find $\widetilde{Y} \subseteq \omega \cdot n$ such that $X < \widetilde{Y}$ and $|\widetilde{Y}| = |Y|$. It follows there is an order preserving function $h : X \cup Y \to X \cup \widetilde{Y}$ such that $h(x) = x$ for all $x \in X$.

By Lemma 32, $h(\Phi)$ is valid. Since $\mathrm{On}(\phi) \subseteq \omega \cdot n$, we have $\mathrm{On}(\phi) \subseteq X$ and $h(\phi) \equiv \phi$. Thus

$$h(\Phi) \equiv (\textstyle\bigwedge_i h(\tau_i)) \to h(\phi) \equiv (\textstyle\bigwedge_i h(\tau_i)) \to \phi.$$

Since T is uniform, each $h(\tau_i) \in T$. In fact, since $\mathrm{range}(h) \subseteq \omega \cdot n$, each $h(\tau_i) \in T \cap (\omega \cdot n)$, and the validity of $(\bigwedge_i h(\tau_i)) \to \phi$ witnesses $T \cap (\omega \cdot n) \models \phi$. □

Definition 36. If T is an $\mathscr{L}_{\omega\cdot\omega}$-theory, its intended structure is the $\mathscr{L}_{\omega\cdot\omega}$-structure $\mathscr{M}_T$ with standard first-order part that interprets the operators of $\mathscr{L}_{\omega\cdot\omega}$ so that for every $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ, assignment s, and $\alpha \in \omega \cdot \omega$,

$$\mathscr{M}_T \models K^\alpha \phi[s] \text{ if and only if } T \cap \alpha \models \phi^s.$$

Lemma 37. Suppose T is an $\mathscr{L}_{\omega\cdot\omega}$-theory. For any $\mathscr{L}_{\omega\cdot\omega}$-formula ϕ and assignment s, $\mathscr{M}_T \models \phi[s]$ if and only if $\mathscr{M}_T \models \phi^s$.

Proof. By induction. □

Recall from Definition 10 that the veristratifier is the stratifier given by $X = \{\omega \cdot 1, \omega \cdot 2, \ldots\}$.

Theorem 38. (The upward stratification theorem) Let $\bullet^+$ be the veristratifier. For any $\mathscr{L}_{\mathrm{EA}}$-theory T, $\mathscr{L}_{\mathrm{EA}}$-formula ϕ, and assignment s, $\mathscr{N}_T \models \phi[s]$ if and only if $\mathscr{M}_{T^\oplus} \models \phi^+[s]$.

Again, the theorem is so-named because it is an upside-down version of a harder theorem that equates $\mathscr{M}_T \models \phi[s]$ with $\mathscr{N}_{T^-} \models \phi^-[s]$ for stratified T and ϕ under more complicated hypotheses.

Proof of Theorem 38. By induction on ϕ. The only nontrivial case is when ϕ is $K\psi$. Then $\phi^+ \equiv K^\alpha \psi^+$ for some α. By definition of the veristratifier, $\alpha = \omega \cdot n$ for some $0 < n \in \mathbb{N}$, and $\mathrm{On}(\psi^+) \subseteq \omega \cdot n$. By Lemma 29, $T^\oplus$ is uniform, so we can use the collapse theorem (Theorem 35). The following are equivalent.

$$\begin{aligned}
\mathscr{N}_T &\models K\psi[s] && \\
T &\models \psi^s && \text{(Definition 6)}\\
T^\oplus &\models (\psi^s)^+ && \text{(Upward proof stratification—Theorem 23)}\\
T^\oplus \cap (\omega \cdot n) &\models (\psi^s)^+ && \text{(The collapse theorem—Theorem 35)}\\
T^\oplus \cap (\omega \cdot n) &\models (\psi^+)^s && \text{(Clearly } (\psi^s)^+ \equiv (\psi^+)^s\text{)}\\
\mathscr{M}_{T^\oplus} &\models K^{\omega\cdot n}\psi^+[s]. && \text{(Definition 36)}
\end{aligned}$$

□

Corollary 39. For any $\mathscr{L}_{\mathrm{EA}}$-theory T, in order to show $\mathscr{N}_T \models T$, it suffices to show $\mathscr{M}_{T^\oplus} \models T^\oplus$.

Corollary 39 provides a foothold for proving truth of self-referential theories by transfinite induction up to $\omega \cdot \omega$: in order to prove $\mathscr{N}_T \models T$, one can attempt to prove $\mathscr{M}_{T^\oplus} \models T^\oplus \cap \alpha$ for all $\alpha \in \omega \cdot \omega$ by induction on α.

5 Upward Generic Axioms

One way to state an epistemological consistency result, for example that a truthful machine can know itself to be true and recursively enumerable, is to show that the schemas in question are consistent with a particular background theory of knowledge. We take a more general approach: show that the doubted schemas are consistent with *any* background theory satisfying certain conditions.

We say that an $\mathscr{L}_{\mathrm{EA}}$-theory T is *K-closed* if $K\phi \in T$ whenever $\phi \in T$.

Definition 40. Suppose T_0 is an $\mathscr{L}_{\mathrm{EA}}$-theory.

1. T_0 is *generic* if $\mathscr{N}_T \models T_0$ for every $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
2. T_0 is *closed-generic* if T_0 is K-closed and $\mathscr{N}_T \models T_0$ for every K-closed $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
3. T_0 is *r.e.-generic* if T_0 is r.e. and $\mathscr{N}_T \models T_0$ for every r.e. $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
4. T_0 is *closed-r.e.-generic* if T_0 is K-closed, r.e., and $\mathscr{N}_T \models T_0$ for every K-closed r.e. $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.

Lemma 41.

1. Generic+r.e. implies r.e.-generic.
2. Generic+K-closed implies closed-generic.
3. Closed-generic+r.e. implies closed-r.e.-generic.
4. R.e.-generic+K-closed implies closed-r.e.-generic.

Proof. Straightforward. □

Lemma 42. Let $T = \cup_{i\in I} T_i$ where each T_i is an $\mathscr{L}_{\mathrm{EA}}$-theory.

1. If the T_i are generic, then T is generic.
2. If the T_i are closed-generic, then T is closed-generic.
3. If the T_i are r.e.-generic and T is r.e., then T is r.e.-generic.
4. If the T_i are closed-r.e.-generic and T is r.e., then T is closed-r.e.-generic.

Proof. Straightforward. □

Lemma 43. The $\mathscr{L}_{\mathrm{EA}}$-schema E_2, consisting of $\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi)$, is generic.

Proof. Suppose $T \supseteq E_2$ is arbitrary. For any $\mathscr{L}_{\mathrm{EA}}$-formulas ϕ and ψ and assignment s, if $\mathscr{N}_T \models K(\phi \rightarrow \psi)[s]$ and $\mathscr{N}_T \models K\phi[s]$, then

$$\begin{aligned}
T &\models \phi^s \rightarrow \psi^s && \text{(Definition 6)}\\
T &\models \phi^s && \text{(Definition 6)}\\
T &\models \psi^s && \text{(Modus Ponens)}\\
\mathscr{N}_T &\models K\psi[s], \text{ as desired.} && \text{(Definition 6)}
\end{aligned}$$

□

Definition 44. Suppose T_0 is an $\mathscr{L}_{\mathrm{EA}}$-theory.

1. T_0 is *upgeneric* if $\mathscr{M}_{T^\oplus} \models T_0^\oplus$ for every $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
2. T_0 is *closed-upgeneric* if T_0 is K-closed and $\mathscr{M}_{T^\oplus} \models T_0^\oplus$ for every K-closed $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
3. T_0 is *r.e.-upgeneric* if T_0 is r.e. and $\mathscr{M}_{T^\oplus} \models T_0^\oplus$ for every r.e. $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.
4. T_0 is *closed-r.e.-upgeneric* if T_0 is K-closed, r.e., and $\mathscr{M}_{T^\oplus} \models T_0^\oplus$ for every K-closed r.e. $\mathscr{L}_{\mathrm{EA}}$-theory $T \supseteq T_0$.

Lemma 45. (Compare Lemma 41)

1. Upgeneric+K-closed implies closed-generic.
2. Upgeneric+r.e. implies r.e.-upgeneric.
3. Closed-upgeneric+r.e. implies closed-r.e.-upgeneric.
4. R.e.-upgeneric+K-closed implies closed-r.e.-upgeneric.

Proof. Straightforward. □

Lemma 46. Suppose $T = \cup_{i\in I} T_i$ where the T_i are $\mathscr{L}_{\mathrm{EA}}$-theories.

1. If the T_i are upgeneric, then T is upgeneric.
2. If the T_i are closed-upgeneric, then T is closed-upgeneric.
3. If the T_i are r.e.-upgeneric and T is r.e., then T is r.e.-upgeneric.
4. If the T_i are closed-r.e.-upgeneric and T is r.e., then T is closed-r.e.-upgeneric.

Proof. Straightforward. □

Lemma 47.

1. Upgeneric implies generic.
2. Closed-upgeneric implies closed-generic.
3. R.e.-upgeneric implies r.e.-generic.
4. Closed-r.e.-upgeneric implies closed-r.e.-generic.

Proof. By the upward stratification theorem (Theorem 38). □

In light of Lemmas 43 and 47, the following shows that upgeneric is strictly stronger than generic.

Lemma 48. E_2 is not upgeneric. In fact E_2 is not even closed-r.e.-upgeneric.

Proof. Let T be the smallest K-closed $\mathscr{L}_{\mathrm{EA}}$-theory containing the following schemata.

1. E_2.
2. $K(1 = 0)$.
3. $K(1 = 0) \rightarrow (1 = 0)$.

Since $T \supseteq E_2$ is closed r.e., it suffices to exhibit some $\theta \in E_2$ and stratifier $\bullet^+$ such that $\mathscr{M}_{T^\oplus} \not\models \theta^+$. If $\bullet^+$ is the stratifier given by $X = \{0, 1, 2, \ldots\}$, the reader can check that

$$\theta \equiv \; K(K(1 = 0) \rightarrow (1 = 0)) \rightarrow KK(1 = 0) \rightarrow K(1 = 0)$$

works. □

Lemma 48 and the following demystify our reason for weakening E_2 to E_2'.

Lemma 49. The schema E_2', consisting of $\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi)$ whenever $\mathrm{depth}(\phi) \leq \mathrm{depth}(\psi)$ (Definition 8), is upgeneric.

Proof. Let $T \supseteq E_2'$ be arbitrary. Suppose ϕ and ψ are $\mathscr{L}_{\mathrm{EA}}$-formulas with $\mathrm{depth}(\phi) \leq \mathrm{depth}(\psi)$ and $\bullet^+$ is a stratifier, say with

$$\begin{aligned}(K\phi)^+ &\equiv K^\alpha \phi^+\\ (K\psi)^+ &\equiv K^\beta \psi^+\\ (K(\phi \rightarrow \psi))^+ &\equiv K^\gamma(\phi^+ \rightarrow \psi^+),\end{aligned}$$

we will show $\mathscr{M}_{T^\oplus}$ satisfies

$$(\mathrm{ucl}(K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi))^+ \equiv \mathrm{ucl}(K^\gamma(\phi^+ \rightarrow \psi^+) \rightarrow K^\alpha \phi^+ \rightarrow K^\beta \psi^+).$$

Note that by Lemma 12, $\alpha \leq \beta = \gamma$. Let s be an arbitrary assignment such that $\mathscr{M}_{T^\oplus} \models K^\gamma(\phi^+ \rightarrow \psi^+)[s]$ and $\mathscr{M}_{T^\oplus} \models K^\alpha \phi^+[s]$. Then

$$\begin{aligned}
T^\oplus \cap \gamma &\models (\phi^+)^s \rightarrow (\psi^+)^s && \text{(Definition 36)}\\
T^\oplus \cap \alpha &\models (\phi^+)^s && \text{(Definition 36)}\\
T^\oplus \cap \beta &\models ((\phi^+)^s \rightarrow (\psi^+)^s) \wedge (\phi^+)^s && \text{(Since } \alpha \leq \beta = \gamma\text{)}\\
T^\oplus \cap \beta &\models (\psi^+)^s && \text{(Modus Ponens)}\\
\mathscr{M}_{T^\oplus} &\models K^\beta \psi^+[s], \text{ as desired.} && \text{(Definition 36)}
\end{aligned}$$

□

Lemma 50. The Assigned Validity schema, consisting of ϕ^s whenever ϕ is valid and s is any assignment, is upgeneric.

Proof. Let $T \supseteq$ (Assigned Validity) be arbitrary. Suppose ϕ is valid, s is an assignment, and $\bullet^+$ is a stratifier. By Theorem 20, ϕ^+ is also valid. Thus $\mathscr{M}_{T^\oplus} \models \phi^+[s]$, and by Lemma 37, $\mathscr{M}_{T^\oplus} \models (\phi^+)^s$. By arbitrariness of ϕ, s, and $\bullet^+$, $\mathscr{M}_{T^\oplus} \models \text{(Assigned Validity)}^\oplus$. □

Lemma 51. Any set of true purely arithmetical sentences is upgeneric.

Proof. Trivial: $\mathscr{M}_T$ has standard first-order part. □

Lemma 52. The schema consisting of the axioms of Epistemic Arithmetic (Peano Arithmetic with induction extended to $\mathscr{L}_{\mathrm{EA}}$) is upgeneric.

Proof. Let $T \supseteq$ (Epistemic Arithmetic). Let σ be an axiom of Epistemic Arithmetic, $\bullet^+$ a stratifier. If σ is not an induction instance, then $\mathscr{M}_{T^\oplus} \models \sigma^+$ by Lemma 51. But suppose σ is an instance

$$\mathrm{ucl}(\phi(x|0) \rightarrow \forall x(\phi \rightarrow \phi(x|S(x))) \rightarrow \forall x \phi)$$

of induction, so that σ^+ is $\mathrm{ucl}(\phi^+(x|0) \rightarrow \forall x(\phi^+ \rightarrow \phi^+(x|S(x))) \rightarrow \forall x \phi^+)$. To show $\mathscr{M}_{T^\oplus} \models \sigma^+$, let s be an assignment and assume $\mathscr{M}_{T^\oplus} \models \phi^+(x|0)[s]$ and $\mathscr{M}_{T^\oplus} \models \forall x(\phi^+ \rightarrow \phi^+(x|S(x)))[s]$. Then

$$\begin{aligned}
\mathscr{M}_{T^\oplus} &\models \phi^+(x|0)^s && \text{(Lemma 37)}\\
\mathscr{M}_{T^\oplus} &\models (\phi^+)^{s(x|0)} && \text{(Clearly } \psi(x|0)^s \equiv \psi^{s(x|0)}\text{)}\\
\forall n \in \mathbb{N},\ \text{if } \mathscr{M}_{T^\oplus} &\models \phi^+[s(x|n)], \text{ then } \mathscr{M}_{T^\oplus} \models \phi^+(x|S(x))[s(x|n)] && \text{(First-order semantics of } \forall \text{ and } \rightarrow\text{)}\\
\forall n \in \mathbb{N},\ \text{if } \mathscr{M}_{T^\oplus} &\models (\phi^+)^{s(x|n)}, \text{ then } \mathscr{M}_{T^\oplus} \models (\phi^+(x|S(x)))^{s(x|n)} && \text{(Lemma 37)}\\
\forall n \in \mathbb{N},\ \text{if } \mathscr{M}_{T^\oplus} &\models (\phi^+)^{s(x|n)}, \text{ then } \mathscr{M}_{T^\oplus} \models (\phi^+)^{s(x|n+1)} && \text{(Clearly } \psi(x|S(x))^{s(x|n)} \equiv \psi^{s(x|n+1)}\text{)}\\
\forall n \in \mathbb{N},\ \mathscr{M}_{T^\oplus} &\models (\phi^+)^{s(x|n)} && \text{(Mathematical induction)}\\
\forall n \in \mathbb{N},\ \mathscr{M}_{T^\oplus} &\models (\phi^+)[s(x|n)] && \text{(Lemma 37)}\\
\mathscr{M}_{T^\oplus} &\models \forall x \phi^+[s], \text{ as desired.} && \text{(First-order semantics of } \forall\text{)}
\end{aligned}$$

□

Armed with Lemmas 42 and 46, computations such as Lemmas 43, 49, 50, 51 and 52 can be used as building blocks for background theories of knowledge. Often, schemas we would like as building blocks are not (up)generic in isolation, but become so when paired with other building blocks, as in the following three lemmas.

Lemma 53. $E_1 \cup$ (Assigned Validity) is upgeneric (E_1 consists of $\mathrm{ucl}(K\phi)$ whenever ϕ is valid).

Proof. Let $T \supseteq E_1 \cup$ (Assigned Validity). By Lemma 50, $\mathscr{M}_{T^\oplus} \models \text{(Assigned Validity)}^\oplus$, we need only show $\mathscr{M}_{T^\oplus} \models E_1^\oplus$. Let ϕ be valid, $\bullet^+$ any stratifier, and s any assignment. Since $T \supseteq$ (Assigned Validity), $T^\oplus$ contains the instance

$$(\phi^s)^+ \equiv (\phi^+)^s$$

of $\text{(Assigned Validity)}^\oplus$. In fact, $T^\oplus \cap \alpha$ contains $(\phi^+)^s$, where α is such that $(K\phi)^+ \equiv K^\alpha \phi^+$. Thus by Definition 36, $\mathscr{M}_{T^\oplus} \models K^\alpha \phi^+[s]$, that is, $\mathscr{M}_{T^\oplus} \models (K\phi)^+[s]$. This shows $\mathscr{M}_{T^\oplus} \models E_1^\oplus$. □

Lemma 54. For any upgeneric T_0, $T_0 \cup K(T_0)$ is upgeneric, where $K(T_0)$ consists of $K\phi$ whenever $\phi \in T_0$. Similarly with "upgeneric" replaced by "r.e.-upgeneric", "closed-upgeneric", "closed-r.e.-upgeneric", "generic", "r.e.-generic", "closed-generic", or "closed-r.e.-generic" throughout.

Proof. We prove the upgeneric statement. Suppose T_0 is upgeneric and $T \supseteq T_0 \cup K(T_0)$. Since T_0 is upgeneric and $T \supseteq T_0$, $\mathscr{M}_{T^\oplus} \models T_0^\oplus$. It remains to show $\mathscr{M}_{T^\oplus} \models (K\phi)^+$ for any sentence $\phi \in T_0$ and stratifier $\bullet^+$. Let α be such that $(K\phi)^+ \equiv K^\alpha \phi^+$. By Definition 10, $\mathrm{On}(\phi^+) \subseteq \alpha$ and thus $\phi^+ \in T_0^\oplus \cap \alpha \subseteq T^\oplus \cap \alpha$. Since $T^\oplus \cap \alpha \models \phi^+$, $\mathscr{M}_{T^\oplus} \models K^\alpha \phi^+$ as desired. □

We will not use the following lemma, but it illuminates differences between this paper's upward approach and Carlson's original downward approach.

Lemma 55. $E_1 \cup E_2 \cup E_4 \cup$ (Epistemic Arithmetic) is closed-generic.

Proof. Let T be a K-closed theory containing E_1, E_2, E_4 and (Epistemic Arithmetic).

By Lemma 43, $\mathscr{N}_T \models E_2$. By Lemmas 52 and 47, $\mathscr{N}_T \models$ (Epistemic Arithmetic). It remains to show $\mathscr{N}_T \models E_1 \cup E_4$. We will show $\mathscr{N}_T \models E_4$ and sketch $\mathscr{N}_T \models E_1$.

The typical sentence in E_4 is $\mathrm{ucl}(K\phi \rightarrow KK\phi)$. Let s be an assignment and assume $\mathscr{N}_T \models K\phi[s]$. Then

$$\begin{array}{rlr}
 & T \models \phi^s & \text{(Definition 6)} \\
\exists \tau_1, \ldots, \tau_n \in T \text{ s.t. } & (\wedge_{i=1}^n \tau_i) \rightarrow \phi^s \text{ is valid} & \text{(Theorem 3)} \\
 & T \models K\left((\wedge_{i=1}^n \tau_i) \rightarrow \phi^s\right) & (T \text{ contains } E_1) \\
 & T \models (\wedge_{i=1}^n K(\tau_i)) \rightarrow K\phi^s & \text{(Repeated applications of } E_2 \text{ in } T) \\
 & T \models \wedge_{i=1}^n K(\tau_i) & (T \text{ is } K\text{-closed}) \\
 & T \models K\phi^s & \text{(Modus Ponens)} \\
 & \mathscr{N}_T \models KK\phi[s]. & \text{(Definition 6)}
\end{array}$$

This shows $\mathscr{N}_T \models E_4$.

Because of the lack of Assigned Validity, showing $\mathscr{M}_T \models E_1$ is tricky. We indicate a rough sketch. Carlson's Lemmas 5.23 and 7.1 [6] (pp. 69 & 72) imply $T \models$ (Assigned Validity) (we invoke Lemma 7.1 with $\mathscr{Q}$ a singleton). As written, Lemma 5.23 demands T also contain E_3, but it can be shown this is unnecessary. Thus we may assume T contains Assigned Validity. By Lemmas 53 and 47, $\mathscr{N}_T \models E_1$. □

Lemma 55 explains why weakening E_2 to E_2' required two other seemingly-unrelated weakenings: adding Assigned Validity, and removing E_4 altogether.

Lemma 56. The Mechanicalness schema,

$$\mathrm{ucl}(\exists e \forall x (K\phi \leftrightarrow x \in W_e)) \quad (e \notin \mathrm{FV}(\phi)),$$

is r.e.-upgeneric.

Proof. Let T be any r.e. $\mathscr{L}_{\mathrm{EA}}$-theory containing the Mechanicalness schema. Let $\bullet^+$ be a stratifier and let α be such that $(K\phi)^+ \equiv K^\alpha \phi^+$. We must show

$$\mathscr{M}_{T^\oplus} \models \mathrm{ucl}(\exists e \forall x (K^\alpha \phi^+ \leftrightarrow x \in W_e)).$$

Let s be any assignment and note

$$\{q \in \mathbb{N} : \mathscr{M}_{T^\oplus} \models K^\alpha \phi^+[s(x|q)]\} = \{q \in \mathbb{N} : T^\oplus \cap \alpha \models (\phi^+)^{s(x|q)}\}. \qquad \text{(Definition 36)}$$

By the Church–Turing Thesis, the latter set is r.e., so there is some $p \in \mathbb{N}$ such that

$$W_p = \{q \in \mathbb{N} : \mathscr{M}_{T^\oplus} \models K^\alpha \phi^+[s(x|q)]\}.$$

For all $q \in \mathbb{N}$, the following biconditionals are equivalent:

$$\begin{aligned}
&\mathscr{M}_{T^\oplus} \models K^\alpha\phi^+ \leftrightarrow x \in W_e[s(e|p)(x|q)] \\
&\mathscr{M}_{T^\oplus} \models K^\alpha\phi^+[s(e|p)(x|q)] \text{ iff } \mathscr{M}_{T^\oplus} \models x \in W_e[s(e|p)(x|q)] && \text{(First-order semantics of } \leftrightarrow) \\
&\mathscr{M}_{T^\oplus} \models K^\alpha\phi^+[s(x|q)] \text{ iff } \mathscr{M}_{T^\oplus} \models x \in W_e[s(e|p)(x|q)] && \text{(Since } e \notin \mathrm{FV}(\phi)) \\
&\mathscr{M}_{T^\oplus} \models K^\alpha\phi^+[s(x|q)] \text{ iff } q \in W_p. && \text{(Since } \mathscr{M}_{T^\oplus} \text{ has standard first-order part)}
\end{aligned}$$

The latter is true by definition of p. By arbitrariness of q, $\mathscr{M}_{T^\oplus} \models \exists e \forall x(K^\alpha\phi^+ \leftrightarrow x \in W_e)[s]$. □

Corollary 57. (Recall the definition of T^w_{SMT} from the end of Section 2) Let $(T^w_{\mathrm{SMT}})\backslash E_3$ be the smallest K-closed theory containing E_1, Assigned Validity, E_2', Epistemic Arithmetic, and Mechanicalness. (Loosely speaking, T^w_{SMT} minus E_3.) Then $(T^w_{\mathrm{SMT}})\backslash E_3$ is r.e.-upgeneric.

6 The Main Result

With the machinery of Section 5, we are able to state our main result in a generalized form. Informally:

> An r.e.-upgeneric theory remains true upon augmentation by knowledge of its own truthfulness.

Reinhardt's conjecture (proved by Carlson) was that the Strong Mechanistic Thesis is consistent with a particular background theory of knowledge. We showed (Lemma 56) that Mechanicalness is r.e.-upgeneric. By Lemma 54, the pair consisting of Mechanicalness and the Strong Mechanistic Thesis, is r.e.-upgeneric. Thus as long as the background theory of knowledge is r.e. and built of r.e.-generic pieces along with truthfulness, the corresponding conjecture is a special case of this main result.

Recall (Definition 6) that an $\mathscr{L}_{\mathrm{EA}}$-theory T is *true* if $\mathscr{N}_T \models T$.

Theorem 58. Let T_0 be an r.e.-upgeneric $\mathscr{L}_{\mathrm{EA}}$-theory. Let T_1 be $T_0 \cup E_3$, that is, T_0 along with all axioms of the form $\mathrm{ucl}(K\phi \rightarrow \phi)$. Let T be the smallest K-closed theory containing T_1. Then T is true.

Proof. By Corollary 39 it is enough to show $\mathscr{M}_{T^\oplus} \models T^\oplus$. We will use transfinite induction up to $\omega \cdot \omega$ to show that for all $\alpha \in \omega \cdot \omega$, $\mathscr{M}_{T^\oplus} \models T^\oplus \cap \alpha$.

Let $\sigma \in T^\oplus \cap \alpha$. Then $\sigma \equiv \theta^+$ for some $\theta \in T$ and some stratifier $\bullet^+$. We will show $\mathscr{M}_{T^\oplus} \models \theta^+$.

Case 1: $\theta \in T_0$. Then $\mathscr{M}_{T^\oplus} \models \theta^+$ because $T \supseteq T_0$ is r.e. and T_0 is r.e.-upgeneric.

Case 2: θ is $K\phi$ for some sentence $\phi \in T$. Let α_0 be such that $(K\phi)^+ \equiv K^{\alpha_0}\phi^+$. By Definition 10, $\mathrm{On}(\phi^+) \subseteq \alpha_0$ and thus $\phi^+ \in T^\oplus \cap \alpha_0$, so $T^\oplus \cap \alpha_0 \models \phi^+$, so $\mathscr{M}_{T^\oplus} \models K^{\alpha_0}\phi^+$.

Case 3: θ is $\mathrm{ucl}(K\phi \rightarrow \phi)$ for some ϕ. Let α_0 be such that $(K\phi)^+ \equiv K^{\alpha_0}\phi^+$, so θ^+ is $\mathrm{ucl}(K^{\alpha_0}\phi^+ \rightarrow \phi^+)$. Since $\theta^+ \in T^\oplus \cap \alpha$, this forces $\alpha_0 < \alpha$. Let s be any assignment and assume $\mathscr{M}_{T^\oplus} \models K^{\alpha_0}\phi^+[s]$. Then:

$$\begin{aligned}
&\mathscr{M}_{T^\oplus} \models K^{\alpha_0}\phi^+[s] && \text{(Assumption)} \\
&T^\oplus \cap \alpha_0 \models (\phi^+)^s && \text{(Definition 36)} \\
&\mathscr{M}_{T^\oplus} \models (\phi^+)^s && \text{(By } \omega\cdot\omega\text{-induction, } \mathscr{M}_{T^\oplus} \models T^\oplus \cap \alpha_0) \\
&\mathscr{M}_{T^\oplus} \models \phi^+[s], \text{ as desired.} && \text{(Lemma 37)}
\end{aligned}$$

□

Corollary 59. T^w_{SMT} is true.

Proof. By Theorem 58 and Corollary 57. □

If one is willing to induct up to $\epsilon_0 \cdot \omega$ and use machinery from [5], it is possible (without the grievous sacrifices we have made in this paper) to generalize Reinhardt's conjecture to a statement of the form:

> Any r.e. theory that is generic in a very specific sense (one that allows E_2 as building block) remains true upon augmentation by knowledge of its own truthfulness. $(*)$

The specific notion of "generic" in order for this to work is somewhat complicated and hinges on [5], putting it out of the present paper's scope. It does admit Mechanicalness as building block, so that $(*)$ really is a generalization of Reinhardt's conjecture, and the notion also admits full E_2, which in turn allows building blocks containing E_4.

The main result of [2] can also be generalized in this manner. The methods of that paper are easily modified to prove:

> For any r.e. $\mathscr{L}_{\mathrm{EA}}$-theory T that is generic (in the sense of Definition 40), there is an $n \in \mathbb{N}$ such that T' is true, where T' is the smallest K-closed theory containing T along with the schema $\forall x(K\phi \leftrightarrow \langle x, \overline{\ulcorner \phi \urcorner} \rangle \in W_{\overline{n}})$ $(\mathrm{FV}(\phi) \subseteq \{x\})$. Less formally, any such generic knowing machine can be taught its own code and still remain true.

One possible application of this paper is to reverse mathematics [14]. Since the results (except Lemma 52) only use induction up to $\omega \cdot \omega$, suitable versions (minus Lemma 52 and references to $\mathbb{N}$) could be formalized and proved in weak subsystems of arithmetic.

References

[1] Alexander, S. (2013). The Theory of Several Knowing Machines. Doctoral dissertation, the Ohio State University.

[2] Alexander, S. (preprint). A machine that knows its own code. To appear in *Studia Logica*.

[3] Alexander, S. (preprint). Self-referential theories. Submitted.

[4] Benacerraf, P. (1967). God, the Devil, and Gödel. *The Monist*, **51**, 9–32.

[5] Carlson, T.J. (1999). Ordinal arithmetic and Σ_1-elementarity. *Archive for Mathematical Logic*, **38**, 449–460.

[6] Carlson, T.J. (2000). Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic*, **105**, 51–82.

[7] Carlson, T.J. (2001). Elementary patterns of resemblance. *Annals of Pure and Applied Logic*, **108**, 19–77.

[8] Carlson, T.J. (2012). Sound Epistemic Theories and Collapsing Knowledge. Slides from the *Workshop on The Limits and Scope of Mathematical Knowledge* at the University of Bristol.

[9] Lucas, J.R. (1961). Minds, machines, and Gödel. *Philosophy*, **36**, 112–127.

[10] Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford University Press.

[11] Putnam, H. (2006). After Gödel. *Logic Journal of the IGPL*, **14**, 745–754.

[12] Reinhardt, W. (1985). Absolute versions of incompleteness theorems. *Noûs*, **19**, 317–346.

[13] Shapiro, S. (1985). Epistemic and Intuitionistic Arithmetic. In: S. Shapiro (ed.), *Intensional Mathematics* (North-Holland, Amsterdam), pp. 11–46.

[14] Simpson, S. (1982). *Subsystems of Second Order Arithmetic.* 2nd Edition, Cambridge University Press.