An infinitely descending sequence of sound theories each proving the next consistent (Brief technical note)

Benja Fallenstein

This document is part of a collection of quick writeups of results from the December 2013 MIRI research workshop, written during or directly after the workshop. It describes work done by Paul Christiano and Benja Fallenstein, building on previous work by Marcello Herreshoff, Jacob Hilton, Eliezer Yudkowsky, and others.

Yudkowsky & Herreshoff (2013) introduce a sequence $\mathcal{T}_n$ of consistent theories extending PA such that each $\mathcal{T}_n$ proves the soundness of $\mathcal{T}_{n+1}$, i.e. $\mathcal{T}_n \vdash \Box_{\mathcal{T}_{n+1}} \forall x. \ulcorner\varphi(x)\urcorner \to \varphi(x)$ for all formulas $\varphi(x)$, where $\Box_{\mathcal{T}}\ulcorner\varphi\urcorner$ is the provability predicate for $\mathcal{T}$ in the language of arithmetic. Unfortunately, while consistent, these theories are unsound.

Here, we construct a different sequence $\mathcal{T}_n$ such that each $\mathcal{T}_n$ proves the *consistency* of $\mathcal{T}_{n+1}$, i.e., $\mathcal{T}_n \vdash \mathrm{Con}(\mathcal{T}_{n+1})$, or more explicitly, $\mathcal{T}_n \vdash \neg\Box_{\mathcal{T}_{n+1}}\ulcorner\bot\urcorner$. These systems are defined as follows:

$$\mathcal{T}_n := \mathrm{PA} + \psi(n) \to \mathrm{Con}(\mathcal{T}_{n+1}), \quad \text{where} \quad \psi(n) :\leftrightarrow \neg\mathrm{Proves}_{\mathrm{ZFC}}(n, \ulcorner\bot\urcorner), \tag{1}$$

where $\mathrm{Proves}_{\mathrm{ZFC}(n,\ulcorner\varphi\urcorner)}$ is the proposition stating that n is the Gödel number of a proof of φ in ZFC. Note that as long as ZFC is consistent, we have $\mathcal{T}_n \vdash \mathrm{Con}(\mathcal{T}_{n+1})$, since for any particular numeral n, PA can show $\psi(n)$ by mechanical checking. But on the other hand, we cannot use this reasoning *inside PA* to show that $\forall n. \Box_{\mathcal{T}_n}\ulcorner\mathrm{Con}(\mathcal{T}_{n+1})\urcorner$, since PA cannot prove the consistency of ZFC.[1]

A variation of a proof of the consistency of Yudkowsky & Herreshoff's system (relative to Con(ZFC)), which we present below, shows that these new systems are consistent as well. It follows directly from this that they are also sound: Since $\mathcal{T}_{n+1}$ is consistent, $\psi(n) \to \mathrm{Con}(\mathcal{T}_{n+1})$ is true in the standard model, and this is the only axiom of $\mathcal{T}_n$ that is not already in PA.

Moreover, it is easy to see that each $\mathcal{T}_n$ proves the Π_1-soundness of $\mathcal{T}_{n+1}$; that is, we have $\mathcal{T}_n \vdash \forall x. \Box_{\mathcal{T}_{n+1}}\ulcorner\varphi(x)\urcorner \to \varphi(x)$ for every Π_1 formula $\varphi(x)$. This is because $\neg\varphi(x)$ is Σ_1, so PA can show that if there is an x such that $\neg\varphi(x)$, then $\Box_{\mathrm{PA}}\ulcorner\neg\varphi(x)\urcorner$, by constructing a proof that specifies a counterexample to $\varphi(x)$ and mechanically checks that this is indeed a counterexample. This means that the systems $\mathcal{T}_n$ can be used to construct tiling agents in the sense of Yudkowsky & Herreshoff, as long as the goal $\mathcal{G}$ is a Π_1 formula, since then the instance of

[1] In fact, $\mathrm{PA} \nvdash \forall n. \Box_{\mathcal{T}_n}\ulcorner\mathrm{Con}(\mathcal{T}_{n+1})\urcorner$, since it's possible to show by a Löbian argument that otherwise $\mathcal{T}_0 \vdash \bot$, and we show later that $\mathcal{T}_0$ is in fact consistent.

the soundness schema required for tiling is $\mathcal{T}_n \vdash \forall b.\, \Box_{\mathcal{T}_{n+1}} \ulcorner \overline{b} \to \mathcal{G} \urcorner \to (\overline{b} \to \mathcal{G})$; if $\overline{b}$ can be written as a Δ_0 formula, then $\overline{b} \to \mathcal{G}$ is a Π_1 formula as long as $\mathcal{G}$ is, and the soundness schema holds. This is the case, for example, for a goal like "do not destroy the world", as long as "the world is not destroyed before time t" can be written as a Δ_0 formula (because the world is computable, say, and only a bounded amount of computation can happen before time t).

Finally, while there are theorems of $\mathcal{T}_n$ that $\mathcal{T}_{n+1}$ does not prove (since $\mathcal{T}_n \vdash \mathrm{Con}(\mathcal{T}_{n+1})$, and $\mathcal{T}_{n+1}$ cannot prove this since we know it is consistent), there is an n_0 such that all $\mathcal{T}_m$ and $\mathcal{T}_n$ with $m, n \geq n_0$ have the same proof-theoretic ordinal (according to any sensible definition of proof-theoretic ordinal): First, note that $\mathcal{T}_n$ proves anything that $\mathcal{T}_{n+1}$ does, since $\mathcal{T}_{n+1}$'s additional axiom $\psi(n+1) \to \mathrm{Con}(\mathcal{T}_{n+2})$ is a Π_1 formula. Thus, for any reasonable notion of proof-theoretic ordinal, we will have $\alpha_n \geq \alpha_{n+1}$, where α_k is the ordinal of $\mathcal{T}_k$. Since there is no infinitely decreasing sequence of ordinals, we can choose n_0 large enough that $\alpha_m = \alpha_n$ for all $m, n \geq n_0$. Thus, an agent using $\mathcal{T}_{n_0}$ will be able to tile to an arbitrarily long sequence of successor agents all using systems of similar mathematical strength, in the sense of having the same proof-theoretic ordinal. (It is an obvious conjecture that $\alpha_n = \epsilon_0$ for all n, the same proof-theoretic ordinal as that of PA and PA + Con(PA), but we haven't had time to check this, yet.)

We now prove that $\mathcal{T}_n$ is in fact consistent for all n. Work in ZFC and assume that ZFC is inconsistent. Then there is a least n such that $\neg\psi(n)$, and hence $\Box_{\mathrm{PA}} \ulcorner \neg\psi(n) \urcorner$. Thus, $\Box_{\mathrm{PA}} \ulcorner \psi(n) \to \mathrm{Con}(\mathcal{T}_{n+1}) \urcorner$; i.e., $\mathcal{T}_n$ proves the same theorems as PA. By induction, it follows that $\mathcal{T}_{n-1}$ is equivalent to PA + Con(PA), $\mathcal{T}_{n-2}$ is equivalent to PA + Con(PA + Con(PA)), and so on up to $\mathcal{T}_0$; since ZFC knows all theses systems to be consistent, it follows that $\mathcal{T}_0$ is consistent. Hence, we have shown in ZFC that $\neg\mathrm{Con}(\mathrm{ZFC}) \to \mathrm{Con}(\mathcal{T}_0)$, or equivalently, $\neg\mathrm{Con}(\mathcal{T}_0) \to \mathrm{Con}(\mathrm{ZFC})$.

Now step outside ZFC and assume that $\mathcal{T}_0$ were inconsistent. Then ZFC would show this, and hence show Con(ZFC), meaning that it would be inconsistent. This establishes the consistency of $\mathcal{T}_0$ relative to the consistency of ZFC. Finally, if any $\mathcal{T}_{n+1}$ were inconsistent, then $\mathcal{T}_n$ would be inconsistent as well, since it would show both $\mathrm{Con}(\mathcal{T}_{n+1})$ and $\neg\mathrm{Con}(\mathcal{T}_{n+1})$; this establishes the consistency of all $\mathcal{T}_n$.

Remark. Essentially the same trick can be used to establish the consistency of the alternative definition of $\mathcal{T}_n$ using $\psi(n) := \mathrm{Proves}_{\mathcal{T}_0}(n, \ulcorner \bot \urcorner)$; we can then show in PA that if $\mathcal{T}_0$ were inconsistent, then it would be equivalent to one of the systems P_n defined by $P_0 :=$ PA and $P_{n+1} := P_n + \mathrm{Con}(P_n)$, which would then have to be inconsistent as well; thus, if we believe in the consistency of all P_n, then all $\mathcal{T}_n$ are consistent as well.