The Procrastination Paradox
(Brief technical note)

Eliezer Yudkowsky

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 mainly by Marcello Herreshoff, Jacob Hilton, Benja Fallenstein and Stuart Armstrong, mostly at previous MIRI workshops.

Abstract

A theorem by Marcello Herreshoff, Benja Fallenstein, and Stuart Armstrong shows that if there exists an infinite series of theories $\mathcal{T}_i$ extending $\mathcal{PA}$ where each $\mathcal{T}_i$ proves the soundness of $\mathcal{T}_{i+1}$, then all the $\mathcal{T}_i$ must have only nonstandard models. We call this the Procrastination Theorem for reasons which will become apparent.

1 Review: Hierarchies of trust.

(This section primarily reviews material found in the paper "Tiling Agents for Self-Modifying AI, and the Löbian obstacle."[1])

Notation:

$\mathcal{T}$ is an axiomatic system in classical logic whose consequences are recursively enumerable.

$\mathcal{T} \vdash \phi : \phi$ is a syntactic consequence of the theory $\mathcal{T}$.

$\mathcal{T} \vdash \bot : \mathcal{T}$ is inconsistent, i.e. proves some instance of $P \wedge \neg P$.

$Prv_{\mathcal{T}}(p, \ulcorner\phi\urcorner) : p$ is the Gödel number of a proof from the axioms of $\mathcal{T}$ whose conclusion is the theorem ϕ.

$\Box_{\mathcal{T}}\ulcorner\phi\urcorner \equiv \exists p \colon Prv_{\mathcal{T}}(p, \ulcorner\phi\urcorner)$

E.g. $\mathcal{T} \vdash \phi \implies \mathcal{PA} \vdash \Box_{\mathcal{T}}\ulcorner\phi\urcorner$ states that whenever $\mathcal{T}$ asserts ϕ, first-order Peano arithmetic ($\mathcal{PA}$) proves that there exists a T-proof of ϕ. (This follows from $\mathcal{T}$ being recursively enumerable and is true even if $\mathcal{T}$ is a more powerful system than $\mathcal{PA}$.)

1. Trust. We say that $\mathcal{T}_1$ *trusts* $\mathcal{T}_2$ iff, for every formula ϕ with zero or one variables free, $\mathcal{T}_1$ asserts the uniform reflection principle:

$$\underset{\phi}{\forall} \colon \ \mathcal{T}_1 \vdash \forall x \colon (\ \Box_{\mathcal{T}_2}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)\ )$$

Thus for every formula ϕ, $\mathcal{T}_1$ proves that for every x, if $\mathcal{T}_2$ proves $\phi(x)$ then $\phi(x)$. We can also say that $\mathcal{T}_1$ verifies $\mathcal{T}_2$, or that $\mathcal{T}_1$ proves $\mathcal{T}_2$ sound.

Trust is transitive: if $\mathcal{T}_i$ trusts $\mathcal{T}_j$ and $\mathcal{T}_j$ trusts $\mathcal{T}_k$ then $\mathcal{T}_i$ trusts $\mathcal{T}_k$.

[1] http://intelligence.org/files/TilingAgents.pdf

Löb's Theorem states that in r.e. systems $\mathcal{T}$ which are at least as powerful as $\mathcal{PA}$:

$$\mathcal{T} \vdash \Box_{\mathcal{T}} \ulcorner \phi \urcorner \to \phi \implies \mathcal{T} \vdash \phi$$

Löb's Theorem implies that no consistent $\mathcal{T}$ can trust itself, since then $\mathcal{T}$ would prove $\mathcal{T} \vdash \Box_{\mathcal{T}} \ulcorner \bot \urcorner \to \bot$, which by Löb's Theorem would imply $\mathcal{T} \vdash \bot$.

2. Finite waterfalls. For any ordinal α, including limit ordinals γ of a sequence $\beta_i < \gamma$, we can define a hierarchy of theories which trust all lower theories and base-$\mathcal{T}$:

$$\begin{aligned} \mathcal{T}{+}0 &\equiv \ \mathcal{T} \\ \mathcal{T}{+}(\alpha{+}1) &\equiv \ \mathcal{T} \ \cup \ \underset{\phi}{\forall}\colon \ \forall x\colon (\ \Box_{\mathcal{T}+\alpha} \ulcorner \phi(x) \urcorner \to \phi(x) \ ) \\ \mathcal{T}{+}\gamma &\equiv \ \mathcal{T} \ \cup \ \underset{\phi}{\forall}\colon \ \forall i\colon \forall x\colon (\ \Box_{\mathcal{T}+\beta_i} \ulcorner \phi(x) \urcorner \to \phi(x) \ ) \end{aligned}$$

3. Naive infinite waterfall. Let $\mathcal{T}$ be an r.e. theory at least as powerful as first-order Peano arithmetic. Gödel's diagonalization lemma states that for any formula Γ free in one variable, there exists a fixpoint formula ϕ such that

$$\mathcal{T} \vdash \phi \leftrightarrow \Gamma(\ulcorner \phi \urcorner)$$

E.g. diagonalizing $\Gamma(x) \equiv \neg \Box_{\mathcal{PA}} \ulcorner x \urcorner$ yields the Godel statement G with $\mathcal{PA} \vdash G \leftrightarrow \neg \Box_{\mathcal{PA}} \ulcorner G \urcorner$.

By a similar fixed-point construction we might attempt to build an infinite series of theories $\tilde{\mathcal{T}}{-}n$ with each $\tilde{\mathcal{T}}{-}n$ trusting $\tilde{\mathcal{T}}{-}n{+}1$:

$$\tilde{\mathcal{T}}{-}n \equiv \ \mathcal{T} \ \cup \ \underset{\phi}{\forall}\colon \ \forall x\colon (\ \Box_{\tilde{\mathcal{T}}-n+1} \ulcorner \phi(x) \urcorner \to \phi(x) \ )$$

Theorem. Every $\tilde{\mathcal{T}}{-}n$ is inconsistent.

Proof. For every n, if $\mathcal{PA}$ proves that $\tilde{\mathcal{T}}{-}n{+}1$ is inconsistent we have:

$$\mathcal{PA} \vdash \Box_{\tilde{\mathcal{T}}-n+1} \ulcorner \bot \urcorner \implies \tilde{\mathcal{T}}{-}n \vdash \Box_{\tilde{\mathcal{T}}-n+1} \ulcorner \bot \urcorner \implies \tilde{\mathcal{T}}{-}n \vdash \bot$$

Löb's Theorem allows us to assume that a theorem is provable in order to prove it. Reasoning in $\mathcal{PA}$, suppose it were provable that all $\tilde{\mathcal{T}}{-}n$ are inconsistent. Then all $\tilde{\mathcal{T}}{-}n$ would be inconsistent. Therefore all $\tilde{\mathcal{T}}{-}n$ are inconsistent.[2] $\Box$

4. Sophisticated waterfall. (Herreshoff, previous research.) Let $\mathcal{T}$ be a theory significantly weaker than Zermelo-Fraenkel set theory (e.g. $\mathcal{T} \equiv \mathcal{PA}$), and

[2] When reasoning within a system $\mathcal{S}$ subject to Löb's Theorem, we can freely assume $\Box_{\mathcal{S}} \ulcorner \phi \urcorner$ whenever attempting to prove ϕ. To expand the proof:

$$\begin{aligned} \text{(Assume) } \mathcal{PA} \cup \chi &\vdash \Box_{\mathcal{PA}} \ulcorner \forall m\colon \Box_{\tilde{\mathcal{T}}-m} \ulcorner \bot \urcorner \urcorner \\ \mathcal{PA} \cup \chi &\vdash \Box_{\mathcal{PA}} \ulcorner \forall m\colon \Box_{\tilde{\mathcal{T}}-m+1} \ulcorner \bot \urcorner \urcorner \\ \mathcal{PA} \cup \chi &\vdash \forall n\colon \Box_{\tilde{\mathcal{T}}-n} \ulcorner \forall m\colon \Box_{\tilde{\mathcal{T}}-m+1} \ulcorner \bot \urcorner \urcorner \\ \mathcal{PA} \cup \chi &\vdash \forall n\colon \Box_{\tilde{\mathcal{T}}-n} \ulcorner \Box_{\tilde{\mathcal{T}}-n+1} \ulcorner \bot \urcorner \urcorner \\ \mathcal{PA} \cup \chi &\vdash \forall n\colon \Box_{\tilde{\mathcal{T}}-n} \ulcorner \bot \urcorner \\ \text{(Deduction Theorem) } \mathcal{PA} &\vdash \Box_{\mathcal{PA}} \ulcorner \forall m\colon \Box_{\tilde{\mathcal{T}}-m} \ulcorner \bot \urcorner \urcorner \to (\forall m\colon \Box_{\tilde{\mathcal{T}}-m} \ulcorner \bot \urcorner) \\ \text{(Löb) } \mathcal{PA} &\vdash \forall m\colon \Box_{\tilde{\mathcal{T}}-m} \ulcorner \bot \urcorner \end{aligned}$$

To constructively obtain the desired result $\forall n : \tilde{\mathcal{T}}{-}n \vdash \bot$, just repeat the above reasoning from your own vantage point rather than within $\mathcal{PA}$.

let $\psi(n)$ be the statement that n does *not* Gödel-encode a proof of inconsistency in Zermelo-Fraenkel set theory. We will then have:

$$\begin{aligned} \psi(n) &\equiv \neg Prv_{\mathcal{ZF}}(n, \ulcorner \bot \urcorner) \\ \forall n : \mathcal{T} &\vdash \psi(n) \\ \mathcal{T} &\nvdash \forall n : \psi(n) \end{aligned}$$

Let each $\mathcal{T}\text{–}n$ only trust $\mathcal{T}\text{–}n{+}1$ if $\psi(n)$ is true:

$$\mathcal{T}\text{–}n \equiv \ \mathcal{T} \ \cup \ \underset{\phi}{\forall}\!: \psi(n) \to (\ \forall x\!: (\ \Box_{\mathcal{T}-n+1} \ulcorner \phi(x) \urcorner \to \phi(x)\ )\ )$$

Theorem. If $\mathcal{ZF}$ proves the consistency of every $\mathcal{T}{+}n$, then $\mathcal{ZF}$ proves $\mathcal{T}\text{–}0$ consistent.

Intuition. Suppose that $\mathcal{ZF}$ were actually inconsistent, i.e. $\neg Con(\mathcal{ZF}) \implies \exists k : Prv_{\mathcal{ZF}}(k, \ulcorner \bot \urcorner) \implies \exists k : \neg\psi(k)$. Then $\mathcal{T}\text{–}0$ would be equivalent to the finite waterfall $\mathcal{T}{+}k$ for the smallest such k. We could translate any proof of $\bot$ in $\mathcal{T}\text{–}0$ into a proof in $\mathcal{T}{+}k$ (by translating invocations of $\mathcal{T}\text{–}0$'s axioms trusting $\mathcal{T}{-}1$ into invocations of $\mathcal{T}{+}k$'s trust in $\mathcal{T}{+}k{-}1$, and so on by induction grounding in $\mathcal{T}\text{–}k \equiv \mathcal{T}{+}0 \equiv \mathcal{T}$). E.g. if $\mathcal{T} \equiv \mathcal{PA}$ then $\neg Con(\mathcal{ZF}) \implies \mathcal{T}\text{–}0 \cong \mathcal{PA}{+}k$ for some k. We think every $\mathcal{PA}{+}k$ is consistent. Then if we exhibit a proof of inconsistency in $\mathcal{T}\text{–}0$, we prove Zermelo-Fraenkel set theory consistent and win eternal fame and fortune. It couldn't possibly be that easy to prove $\mathcal{ZF}$ consistent, therefore $\mathcal{T}\text{–}0$ is consistent.

In fact, exhibiting an inconsistency in $\mathcal{T}\text{–}0$ is *such* a cheap way of proving $\mathcal{ZF}$ consistent that, if it worked, we could probably prove $\mathcal{ZF}$ consistent from *within* $\mathcal{ZF}$, meaning that $\mathcal{ZF}$ would prove its own consistency, implying that $\mathcal{ZF}$ would be inconsistent due to Gödel's 2nd Incompleteness Theorem, implying that $\mathcal{T}\text{–}0$ would be consistent.

Proof. Carry out the above reasoning within $\mathcal{ZF}$.

$$\begin{aligned} &\mathcal{ZF} \vdash \neg Con(\mathcal{ZF}) \to (\exists k \in \mathbb{N} : (\Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \leftrightarrow \Box_{\mathcal{PA}+k} \ulcorner \bot \urcorner) \\ &\mathcal{ZF} \vdash \neg Con(\mathcal{ZF}) \to \neg \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \\ &\mathcal{ZF} \vdash \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \to Con(\mathcal{ZF}) \\ &\mathcal{ZF} \vdash \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \to \Box_{\mathcal{ZF}} \ulcorner \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \urcorner \\ &\mathcal{ZF} \vdash \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \to \Box_{\mathcal{ZF}} \ulcorner Con(\mathcal{ZF}) \urcorner \\ \text{(Gödel)}\ &\mathcal{ZF} \vdash \Box_{\mathcal{ZF}} \ulcorner Con(\mathcal{ZF}) \urcorner \leftrightarrow \neg Con(\mathcal{ZF}) \\ &\mathcal{ZF} \vdash \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \to \neg Con(\mathcal{ZF}) \\ &\mathcal{ZF} \vdash \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \to \neg \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \\ &\mathcal{ZF} \vdash \neg \Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner \end{aligned}$$

(The above proof is an improved version produced by Jacob Hilton at the Nov 2013 MIRI Workshop. To summarize the proof within $\mathcal{ZF}$: $\mathcal{ZF}$ shows that if $\mathcal{ZF}$ is inconsistent then $\mathcal{T}\text{–}0$ is consistent, so if $\mathcal{T}\text{–}0$ is inconsistent then $\mathcal{ZF}$ proves its own consistency and therefore is inconsistent, implying $\mathcal{T}\text{–}0$'s consistency; thus $\mathcal{ZF}$ proves $\mathcal{T}\text{–}0$ consistent.)

Herreshoff argued that $\mathcal{T}\text{–}0$ would prove $\exists k : \neg\psi(k)$, and therefore have no standard models, via the following route:

$$\begin{aligned}
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n)) \to \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n)) \to \bot\urcorner \to \Box_{\mathcal{T}\text{–}1}\ulcorner(\forall n : \psi(n)) \to \bot\urcorner \\
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n)) \to \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n)) \to \bot\urcorner \to ((\forall n : \psi(n)) \to \bot) \\
&\mathcal{T}\text{–}0 \vdash \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n)) \to \bot\urcorner \to ((\forall n : \psi(n)) \to \bot) \\
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n)) \to \bot
\end{aligned}$$

However the initial premise was questioned: given $\forall n : \psi(n)$, can we in fact show within $\mathcal{PA}$ that every proof in $\mathcal{T}$–0 maps to a proof of a corresponding theorem in $\mathcal{T}$–1?

An alternate argument (by Armstrong and Fallenstein) then showed a more general result:

2 New result: All infinite sequences of trusted theories have only nonstandard models.

Theorem. For any infinite series of theories $\mathcal{T}_i$ with

$$\underset{\phi}{\forall}: \ \mathcal{T}_i \vdash \ \Box_{\mathcal{T}_{i+1}}\ulcorner\phi\urcorner \to \phi$$

all the $\mathcal{T}_i$ have only nonstandard models (are not sound on $\mathbb{N}$).

Intuition. The theorem was inspired by considering agents which can potentially construct successor agents, and which have a goal of pressing a button at least once over the course of unbounded time. Each agent A_i attempts to prove that constructing its successor A_{i+1} will lead to the button eventually being pressed, via a theorem showing that its successor A_{i+1} will either immediately press the button, or else prove that constructing its own successor A_{i+2} leads to the button eventually being pressed.

If A_0 initially uses a proof system $\mathcal{PA} + \alpha$ for some computable ordinal α, then $\mathcal{PA}+\alpha$ only trusts theories $\mathcal{PA}+\beta$ with $\beta < \alpha$, and A_0 will only be able to carry out its proof if A_1 uses a theory $\mathcal{PA} + \beta, \beta < \alpha$. Thus the chain of agents constructing agents can only descend finitely, and the button will eventually end up being pressed.

But if A_0 begins by trusting $\mathcal{T}$–0, then it may construct an infinite series of successors, each of which proves that the button eventually ends up being pressed. Yet the button never gets pressed after any finite time. If the agents A_i are consistent, their theories must have only nonstandard models in which the button is pressed at a nonstandard time. Hence they are not sound on $\mathbb{N}$.

(Or: Suppose that you and all your future selves believe their later self's promise to finish your paper. Then you just refer the paper to your future self, who refers it to their future self, and so on indefinitely. All of you believe the paper will eventually get done, but there is no finite time at which the paper gets written. We call this the Procrastination Paradox.)

Proof. By diagonalization, let $P_i \equiv \mathcal{T}_i \vdash \exists k > i : \neg P_k$ so that:

$$\mathcal{PA} \vdash P_i \leftrightarrow \Box_{\mathcal{T}_i}\ulcorner\exists k > i : \neg P_k\urcorner$$

Intuitively we could consider P_i a *procrastination sentence* since it asserts that it is fine to procrastinate at time i, because there is some future date $k > i$

when the theory T_k won't be able to prove P_k and will therefore have to stop procrastinating and start getting some work done. Then:

$$\begin{aligned}
&\forall n : \mathcal{T}_n \vdash \neg P_{n+1} \vee P_{n+1} \\
&\forall n : \mathcal{T}_n \vdash \neg P_{n+1} \rightarrow (\exists k > n : \neg P_k) \\
\text{(Diagonalization) } &\forall n : \mathcal{T}_n \vdash P_{n+1} \leftrightarrow \Box_{\mathcal{T}_{n+1}} \ulcorner \exists k > n+1 : \neg P_k \urcorner \\
\text{(Trust) } &\forall n : \mathcal{T}_n \vdash \Box_{\mathcal{T}_{n+1}} \ulcorner \exists k > n+1 : \neg P_k \urcorner \rightarrow (\exists k > n+1 : \neg P_k) \\
&\forall n : \mathcal{T}_n \vdash P_{n+1} \rightarrow \exists k > n : \neg P_k \\
&\forall n : \mathcal{T}_n \vdash \exists k > n : \neg P_k \\
&\forall n : P_n
\end{aligned}$$

Thus any infinite series of theories $\mathcal{T}_i$ trusting $\mathcal{T}_{i+1}$ will all prove their procrastination sentences P_i and also all prove the sentence $\exists k : \neg P_k$.[3] Thus the $\mathcal{T}_i$ must have only nonstandard models.[4] $\Box$

Corollary: No theory sound on $\mathbb{N}$ can verify any theory which trusts an infinitely descending sequence of theories.[5]

We refer to this result as the Procrastination Theorem.

[3] If the base $\mathcal{T}$ proves that every $\mathcal{T}_i$ trusts $\mathcal{T}_{i+1}$, the theorem above becomes provable within $\mathcal{T}$, so $\mathcal{T} \vdash \forall n : P_n \implies \mathcal{T}_i \vdash \bot$. The waterfall theories $\mathcal{T}$–n can be consistent only because the $\mathcal{T}$–n cannot prove the waterfall is infinite; indeed, they each prove that the waterfall halts at some (actually nonstandard) k.

[4] However, consistent $\mathcal{T}_i$ extending $\mathcal{PA}$ are sound for Π_1 sentences on $\mathbb{N}$, since Π_1 sentences with counterexamples can be proven false in $\mathcal{PA}$. Thus an assertion in $\mathcal{T}$–0 of a Π_1 sentence, regardless of how proven, remains trustworthy. A statement that a button is 'eventually' pressed is within Σ_1, hence potentially unsound. But an infinite series of agents proving that a button is pressed on *every* round can be trustworthy.

[5] Thus, despite the apparent ease of constructing infinite descents from base systems far weaker than $\mathcal{ZF}$, an agent using $\mathcal{ZF}$ would not be able to trust a successor agent using any such theory.