Decreasing mathematical strength in one formalization of parametric polymorphism (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 an answer by Will Sawin to a question asked by Benja Fallenstein.

One way of formalizing the essential idea of "parametric polymorphism" described in the tiling agents paper of Yudkowsky and Herreshoff (2013), different from the formalization presented in that paper, is to build an agent using the following decision criterion:

$$\overline{a} \implies \text{PA} \vdash \overline{a} \to \forall n.\, \Box_{\text{PA}}^{n+1}\ulcorner \mathcal{G}(n)\urcorner, \tag{1}$$

where $\Box_{\text{PA}}^{n+1}$ is the $n+1$-fold iterated provability predicate, i.e., $\Box^{2}\ulcorner\varphi\urcorner = \Box\ulcorner\Box\ulcorner\varphi\urcorner\urcorner$ etc., and $\mathcal{G}(n)$ is a family of goals such that the goal n can be achieved by the n'th successor of the current agent.[1]

In the informal metalanguage (which we could formalize as ZFC), we know that $(\text{PA} \vdash \varphi) \iff (\text{PA} \vdash \Box_{\text{PA}}\ulcorner\varphi\urcorner)$ for all formulas φ, where the right-to-left direction follows from the soundness of PA: If PA proves that it proves φ, then it does indeed prove φ. Thus, an agent which takes action a if it can prove $\overline{a} \to \mathcal{G}$ in PA uses a system of exactly the same mathematical strength as an agent which takes action a if it can prove $\Box_{\text{PA}}\ulcorner\overline{a} \to \mathcal{G}\urcorner$ in PA. This leads to the question whether parametric polymorphism using (1) similarly has non-decreasing strength, in the sense that

$$(\text{PA} \vdash \forall n.\, \Box^{n+1}\ulcorner\varphi(n)\urcorner) \iff (\text{PA} \vdash \forall n.\, \Box^{n+2}\ulcorner\varphi(n)\urcorner) \tag{2}$$

for all formulas $\varphi(n)$. This is not the case.

As a counterexample, let T_n be an enumeration of all Turing machines such that $\text{PA} \vdash \forall k.\, \Box_{\text{PA}}^{k+1}\ulcorner T_n(k) \text{ halts}\urcorner$. Let $\tilde{T}$ be the Turing machine such that $\tilde{T}(n)$ evaluates $T_n(n)$. Then by construction, $\text{PA} \vdash \forall k.\, \Box_{\text{PA}}^{k+2}\ulcorner\tilde{T}(k) \text{ halts}\urcorner$.

Suppose we also had $\text{PA} \vdash \forall k.\, \Box_{\text{PA}}^{k+1}\ulcorner\tilde{T}(k) \text{ halts}\urcorner$. Then there would be a $\tilde{n}$ such that $\tilde{T} = T_{\tilde{n}}$. But then, $\tilde{T}(\tilde{n})$ would evaluate $\tilde{T}(\tilde{n})$, i.e., go into an infinite loop; contradiction to $\tilde{T}(k)$ halting for all k.

[1] An agent of this design can tile if $\overline{a}$ is a Δ_0 formula for all actions a, so that $\text{PA} \vdash \overline{a} \to \Box_{\text{PA}}\ulcorner\overline{a}\urcorner$.