Referee Report on: *A functional interpretation for nonstandard analysis* by van den Berg et al

April 26, 2012

This paper introduces a new realizability and functional interpretation where throughout the interpretation one works with non-empty finite sequences (sets) of witnesses, rather than single witnesses. This, surprisingly, gives rise to functional interpretations which are extremely suitable to deal with the kind of principles used in non-standard analysis such as overspill, underspill and the idealization principle, amongst others. The proposed set of techniques is much more involved that the usual modified realizabilty and dialectica interpretations, in that one has to deal with two different kinds of quantifiers (internal and external) and a new atomic predicate "standard(x)", and their corresponding defining axioms. The benefit of this is that most of the classical principles of non-standard analysis become either interpretable or are even eliminated by the interpretation, leading to very strong conservation results which extends previous results of Moerdijk and Palmgren (1997) and Avigad and Helzner (2002). **This is a well-written paper introducing some novel and powerful ideas in the area of functional interpretation, with some impressive applications to non-standard analysis. I have checked in detail more of the proofs, and am quite confident of the correctness of the results. Therefore, I would strongly recommend that the paper be accepted for publication**. I conclude with a few comments which I hope the authors will consider when preparing the final version:

(1) My main suggestion for improvement for the paper is that I don't think it is suitable to identify the type ρ with the set of finite sequence of type ρ, by making use of the surjective Cantor pairing function. Although this is technically correct, I think, for presenting the results, it would be possible (and improve the readability of the paper immensely) to actually distinguish these two types. One should have the type ρ and also the type ρ^* of **finite sets** of objects of type ρ. Although ρ^* is definable in system T, it is good to add it, to improve the readability of the interpretation. In this way one can give types to the new application and the new lambda abstraction as

$$(\cdot)[\cdot] \,:\, (\tau \to \rho^*)^* \to \tau \to \rho^*$$

defined as: given $s^{(\tau\to\rho^*)^*}$ and $t\colon\tau$ we have $(s)[t] = \cup_{f\in s} f(t)$. Similarly we would have

$$\Lambda \,:\, (\tau \to \rho) \to (\tau \to \rho)^*$$

defined as $\Lambda x.t = \{\lambda x.t\}$. With this we have that $(\Lambda x.t)[s] = t[s/x]$. Then, one can say what the types of the variables in Definition 6.1 are explicitly as follows. For instance, given[1] that $\Phi(a)^{D_{st}} \equiv \exists^{\mathsf{st}} x^{\rho^*} \forall^{\mathsf{st}} y^{\tau} \varphi_{D_{st}}(x,y)$ and $\Psi(a)^{D_{st}} \equiv \exists^{\mathsf{st}} u^{\alpha^*} \forall^{\mathsf{st}} v^{\beta} \psi_{D_{st}}(u,v)$ one can write the interpretation of implication as

$$\exists^{\mathsf{st}} U, Y \forall^{\mathsf{st}} x^{\rho^*}, v^{\beta} (\forall y \in Y[x,v]\, \varphi(x,y) \to \psi(U[x],v))$$

and say that the types of U and V are $U\colon(\rho^* \to \alpha^*)^*$ and $Y\colon(\rho^* \to (\beta \to \tau^*))^*$. The point is that U and Y are finite sets of functions with a return type which is also a finite set of a particular type. That's why the new "finite set application" makes sense here. Also, the interpretation of the standard predicate would be

$$\mathrm{st}(u^{\rho})^{D_{st}} \equiv \exists^{\mathsf{st}} x^{\rho^*} (u \in x).$$

So, a witness that u is standard is a finite set containing u.

(2) Line -10 on page 6, in the definition of $n * m$ I think the second $\tilde{\mathrm{lth}}(n)$ should be $\tilde{\mathrm{lth}}(m)$.

(3) Another issue is that of defining the length of a function G (page 6, line -3) as G applied to some canonical value (which the authors choose to be zero). This arbitrary decision could be avoided if, as suggested in point (1) above, they use finite sets explicitly rather than coding finite

[1] Note that in the interpretation the first existential quantifiers are always quantifying over non-empty finite sets (with the monotonicity property).

sequences of type ρ as element of type ρ.

(4) I think the authors must have forgotten to write the side-condition that the formula φ must be *internal* in the statement of the principles $\mathsf{OS}_0, \mathsf{OS}, \mathsf{US}_0, \mathsf{TP}_\forall$, the transfer rules, and various others.

(5) What is the principle R in Proposition 6.11 (and first sentence of Section 8.2)? Did you mean NCR, or maybe the consequence of NCR when $\Phi(x, y)$ is upwards closed. Either case, fix this.

(6) You mentioned on page 32 that there are several unsettled questions concerning the relationship between the various principles. It would be interesting, for the sake of researchers who want might take up the challenge, if you could summarise what is known, and more importantly, what are the open questions here. I know references to these are scattered throughout the paper, but if you could summarise what is known and what's not at the end of the paper would be very beneficial.

(7) The authors write on page 34 (Remark 6.10) that their interpretation has striking similarities with the bounded functional interpretation. An explanation could probably be found in the work of Bezem (Compact and majorizable functionals of finite type, JSL, 1989). In there Bezem shows that an alternative type structure defined in terms of compact sets (which are finite sets in the base type) coincide with the type structure of the strongly majorizable functionals. Maybe there is a third type structure here where finite sets are hereditarily built on all types, and some close relation between this third structure and the other two mentioned could explain the similarity between the two functional interpretations.

(8) It seems that the only clause which does not match the bounded functional interpretation is that of disjunction and the external universal quantifiers. Regarding disjunction, the alternative interpretation given in Definition 9.2 indeed matches the bounded functional interpretation. For the external universal quantifiers, which is different is that the functional X should (if one follows bfi) depend on a bound on z, rather than depend on z directly. Maybe this is also a possible alternative interpretation of $\forall^{\mathsf{st}}$ which could have an implication on the interpretable principles. Anyway, maybe these comments could be added to Remark 6.10.

(9) Could you possibly explain why you believe in Conjecture 8.9? Behind each conjecture there is normally some explanation based on analogies with similar results, for instance.