Zen Rationality and False Beliefs

by Sven Nilsen, 2018

Zen Rationality is an extension of Instrumental Rationality with higher order goal reasoning. In this paper I show that zen rational agents do not consider believing something as a sufficient proof, when it is believed that a smarter version might believe otherwise.

The result of this paper is the following:

$P ? me \rightarrow \Box P$	$: false$	Proof from belief only is not consistent

At least, in a context where a smarter version might believe the opposite.

`□` means “it is provable that ...”.

Proof by combining Naive Zen Logic with provability logic:

$P ? me$	$(\neg P ? .me) ? me$	$P ? me \rightarrow \Box P$	Assumptions
$(\neg P ? .me) ? me \rightarrow \neg P ? me$			Apply zen-consistency
$(P \land \neg P) ? me$			I believe both `P` and `¬P`
$false ? me$			This is logically equivalent to false
$\Box false$			Applying `false ? me → □false`

A proof of false means that some assumption is false. Therefore:

$$(P ? me \land (\neg P ? .me) ? me) \Rightarrow (P ? me \rightarrow \Box P : false)$$

This means that intuition about false beliefs is grounded in the semantics of “smarter versions”.

For example, in Newcomb-like problems, if a zen rational agent can predict that it will do X, but it believes that a smarter version of itself thinks X is worse to do than Y, then it can not assume that just believing it will do X is a proof that it will do X.

Instead, a zen rational agent will not predict that it will do X in the first place, but it might predict it will do Y, if it believes the smarter version would do that. With other words, it will not apply Löb’s theorem about its ouput from decision function as long it believes there exists a smarter approach.

In this context, the definition of “smarter” would also include the belief “Y is better than X” is true.

The Newcomb-like paradox is avoided by restricting the semantics of provability `□` in situations where there is doubt whether a smarter version of itself would be aligned. The actual decision follows from additional assumptions, such as “Y is better than X”. In order to apply Löb’s theorem at all, a zen rational agent would already believe that is decision is optimal relative to its optimization potential, because no smarter version can believe a better decision might be performed.