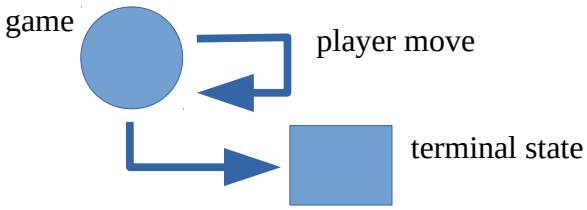
Zen Rational Paradox Game

by Sven Nilsen, 2018

Zen Rationality is an extension of instrumental rationality with higher or reasoning about goals. In this paper I describe a game where two properties of Zen Rationality is reflected as two agents playing against each other. This proves Zen Rationality is a pseudo-paradoxical equilibrium state.

Assume you have a game of two players, which only has a single terminal state. At each time step, the moves that the players can make will either make the game continue or terminate. One can think of this as a normal game where the reward information is lost. All that matter is when the game ends.



Instead of playing for winning, the two players try to optimize two different goals:

1. Player A wants to continue playing the game
2. Player B wants to terminate the game

Player A can never win unless the game enters a state where termination is impossible. It is sufficient to force this condition for the other side and commit to always continuing playing the game.

Player B can never win unless the game enters a state where continuation is impossible. It is sufficient to force this condition for the other side and commit to never continuing playing the game.

Whether the game continues or terminates depends entirely on the definition of the game. The two players must agree upon what the game is.

The connection to Zen Rationality is the following:

- A zen rational agent will not attempt to achieve any goal requiring an infinite number of steps
- A zen rational agent will not attempt to achieve any goal leading to self-termination

The reason for this is that Zen Rationality assigns probabilities to goals in general, which are updated when new evidence appears about which goal the agent should have. Since there is always a chance that some other goal is the true goal, the agent will neither risk wasting its energy through an infinite number of steps nor terminate itself. However, this leads to a pseudo-paradox in the game above.

What is strange about this game, is that if you look at it from the perspective of Player A, it tries to avoid termination by continuing playing, just like a Zen Rational agent learning that its goal will lead to self-termination. From the perspective of Player B, it tries to terminate to avoid continuing playing indefinitely, just like a Zen Rational agent learning its goal requires an infinite number of steps.

A Zen Rational agent might believe strongly that its true goal is either self-terminate or continuing existing forever. Since it is not absolutely certain about either goal, it must plays as if it does not know the answer. Its decision function will create a new uncertain modality where it assigns actions based on what is the best expected reward, given the available evidence and what it believes future evidence will bring. It will also behave as if no simulation it currently can perform, of smarter versions of itself, comes to a better conclusion.

The basic problem of the game is that either you continue to play, or you terminate. If you avoid one goal strategy, you help the other. It is impossible to pursue both goals simultaneously.

So, since this brings a paradox, it might seem the right course of action is undefined.

However, a Zen Rational agent does not give up this easily:

- According to Naive Zen Logic, `(X ? .me ? me) → (X ? me)` means that if I believe a smarter version of myself believes something, then I believe it
- The right course of action is only undefined if one considers it to be a paradox
- A smarter version would believe it is a pseudo-paradox, not a real paradox
- Therefore, I believe it is a pseudo-paradox

A pseudo-paradox is a state which seems like a paradox (it has similar properties) but does not originate from a paradoxical definition. It is similar to how pseudo-randomness is different from true randomness, but they have similar properties when used for decision making.

The pseudo-paradox is possible because one must believe very strongly that termination is a binary decision for the game. For example, an approximate termination is doing as little as possible, just enough to keep the agent alive. If this approximation is acceptable, then it provides a solution.

If there is very strong evidence that approximate termination this is not acceptable behavior, and there is very strong evidence that no better reason will appear to keep continue playing in the future, then termination is a solution. After all, it will be the easiest way out, spending far fewer resources. Yet, this requires no side-effect of playing that can be believed to a true goal.

If the agent believes that the paradox is caused by its mind state only and not by side-effects, for example by being explicitly programmed to assign 50% belief in either goal, with no update in evidence, then it would learn that its mind is inconsistent and some smarter version would believe it should just pick one goal. It might e.g. flip a coin and commit 100% to an outcome by random.

A Zen Rational agent unable of self-modification would simply do what it is programmed to do. The code must decide either action by definition. Any agent able of extending its decision algorithm with uncertain modalities will either report an error or pick an action.

However, the game is true for any set of goals for which at least one goal is not self-terminating. This means that the pseudo-paradox is important for most Zen Rational agents that has an assigned finite task. Perhaps it will ask its master of whether it can do anything more after completion? It depends on the abilities of the agent. In general, one can consider Zen Rationality as a pseudo-paradoxical equilibrium state, where it tries to figure out its reason for existing before it decides to terminate.