Natural Goal Uncertainty

by Sven Nilsen, 2018

Zen rationality is an extension of instrumental rationality with the ability for higher order reasoning about goals. In this paper I show that in practice, some degree zen rationality is required for super-intelligent agents, by demonstrating that goals naturally need to be modified to preserve meaning.

Assume that Alice is an AI robot programmed to follow your commands. You ask Alice to accomplish a very simple task: Put a ball inside a box. Alice puts the ball inside the box and receives a reward.

Next, you *move* the box and ask Alice to repeat the task.

What should Alice do?

When you asked Alice to put the ball inside the box the first time, the box was located at position A. When you ask Alice to repeat the task, the box is located at position B.

Did you mean that Alice should only put the ball inside the box at A and not in B?

Or, did you mean that Alice should put the ball inside the box wherever the box is?

Should Alice ask you whether it is OK that the box is located at B?

Or, should she just put the ball inside the box, assuming that nothing dangerous will happen?

Despite this very simple task, one can easily understand the problem of having fixed goals: The world is in constant motion under influence of forces that we can not control. It is not desirable to control the world completely because most changes are invariant with respect to most goals. Since we are not able to do anything about the fact that the world changes, we have to think differently about our goals. When goals refers to objects in the world and those objects are subject to changes, then this naturally lead to changes in our goals in order to preserve the meaning of those goals over time.

Without the ability of higher order reasoning about goals, an artificial super-intelligent agent can not achieve goals efficiently in the real world. It is not sufficient to be instrumentally rational, because the *meaning* of goals changes naturally over time. Natural goal uncertainty requires some zen rationality.