Lead by Example Utility Cooperation

by Sven Nilsen, 2018

In this paper I represent a way for a Polite Zen Robot to cooperate with its arbitrary self extensions.

In the original paper on the Polite Zen Robot (PZR), it was shown that a super-intelligent agent can be safely extended with arbitrary goals while staying aligned to 3 modules Zen Rationality (ZR), Rational Natural Morality (RNM) and Common Sense Politeness (CSP).

The problem is that such extensions are limited to neutral judgements or infinitesimal utility functions. While this might serve as a practical solution for many everyday tasks, this design fails to accomplish goals that score high in both PZR and the extended goal.

The basic problem is that the PZR risks loss in expected utility when revealing judgements to the sub-agent. An arbitrary sub-agent might use the knowledge about PZR's preferences to reverse them, or attempt in doing so without getting detected by the PZR.

Here is a table illustrating the sections of utility functions that are of same cardinality:

Utility Control Problem	**PZR Low Utility**	**PZR Neutral Utility**	**PZR High Utility**
Ext Low Utility	Blocked by PZR/Ext	Blocked by Ext	Overridden by PZR
Ext Neutral Utility	Blocked by PZR	OK	Overridden by PZR
Ext High Utility	Blocked by PZR	OK	Lead by Example

The "Lead by Example" technique works as following:

1. The PZR shows Ext a task that maximizes PZR
2. Ext repeats the task to show PZR it understands it
3. PZR then permits Ext to perform variants of the task that varies neutrally in PZR

PZR does not give Ext access to its utility function, but instead teaches Ext to do specific tasks. Since each sub-task might be performed in variations that are neutral in PZR but of high utility in Ext, then Ext can do the tasks in a different way that maximizes its own utility. The PZR watches such that the expected difference in utility caused by the variance stays neutral.

This could not be performed by the PZR alone since it might not know Ext's utility function.

Ext is not allowed to increase or decrease the expected utility in PZR a lot. This is to make sure that it is not allowed to behave as if it guesses what PZR is maximizing. This way the PZR can prevent a lot of information about its utility to leak through when interacting with Ext.

With other words, the PZR delegates tasks to the sub-agent by demonstrating what it should do. This is why it is called "Lead by Example". The PZR acts as a leader that proves what it does is in alignment with its own utility and in a way that permits alignment with the sub-agent's utility.