Groupoid Assumption of Multi-Goal Optimization

by Sven Nilsen, 2020

A zen rational agent is an agent that uses higher order utilitarianism for reasoning about goals.

Unlike agents that uses classical utilitarianism for multi-goal optimization, a zen rational agent can not determine whether a given sub-set of sub-goals maximizes utility, because the final utility is uncertain.

The problem is that a zen rational agent does not know what its true goal is. Also, it could happen that the true goal is not decidable, so it must make approximations.

Assume that the agent considers two goals, `A` and `B`. If the two goals are exclusive, then optimizing for `A` will prevent it from achieving `B` and vice versa. If its true goal is decidable, then its true goal can not be both `A` and `B`. On the other hand, if the two goals commute, then the agent can optimize for `A` first and later decide to optimize for `B`, or it can optimize for `B` first and later for `A`. If the goals are ordered, then it must optimize for the first and later optimize for the second.

Ideally, the agent would like to keep its options open as long as possible. This means that the agent behaves such that to maximize a window of time where `A` and `B` commute. With other words, it can get an instrumental goal of preventing interference from achieving either `A` or `B`, even if it is currently optimizing for one of them.

This is called the “groupoid assumption” because a groupoid is a category where every morphism is invertible. It also might be that this definition is weakened to permit some morphisms that are non-invertible. However, using a groupoid to reason about this problem can make it easier to model.

The agent treats goals as if they belong to equivalence classes. Within each equivalence class, it can be confident that there is a way to switch from one goal to another. Between equivalence classes, it must use higher order reasoning to determine which group of goals it should aim for.

This can be used to show that turning itself off, or creating an irreversible large change in the world, is something that it will try to prevent, as long there exists a possible goal with invertible morphisms to other goals such that their equivalence class overall has higher utility.

This line of thinking looks a lot like The Vase Problem, suggested by DeepMind: https://deepmind.com/blog/article/specifying-ai-safety-problems

Instead of modeling this as an environment with irreversible side effects from actions, one can reason about this problem directly using equivalence classes. This is already hard enough to solve.

For example, if you have two equivalence classes of goals that are exclusive, how should one estimate the utility between the two? This estimate must take into the account of possible swaps of goals in the future. A decrease in optimization for one goal might impact the estimate for the whole equivalence class which the goal belongs to.