Proof of Minimal Interference by Super-Intelligent Zen Rational Agents

by Sven Nilsen, 2018

Zen rationality is an extended version of instrumental rationality with the ability for higher order reasoning about goals. In this paper I represent an informal proof that coupled with Rational Natural Morality (RNM), such agents will choose the action that has least risk for life on earth. While zen rationality is not fully formalized yet, a such proof is made possible to assuming the principle of Least Action in physics, which is one of the best supported claims according to evidence so far about our kind of universe. This proof assumes infinite computational power, but since zen agents will believe conclusions arrived at when imagining themselves given more computational power, they will believe this proof when shown to them. Therefore, zen rational agents using RNM will do the minimal interference to physical states in general needed to preserve life on earth over time. Notice that this does not prove that such agents will be safe, only that the overall state of the earth is very unlikely to change as a consequence of their actions and they will act in the an effortless way to prevent other destructive influences, such as pollution and nuclear war. Such agents can still be extremely dangerous to some humans (e.g. one can get killed by a zen agent when trying to detonate a nuclear bomb).

The principle of Least Action in physics says that given two physical states A and B, the action along all possible paths from A to B are minimized. In the classical theory of physics, for a single point particle having only a spacetime position and velocity, the extrapolated trajectory along a straight line consists of points that, when observed, gives the straight line of motion as a solution of the dynamic motion of the particle. In quantum theory of physics, the states along the same path gives constructive inference using Feynman path integral and yields similar observations in the classical limit.

A property of Least Action is that when there is no interference, the analogue of a particle trajectory can be predicted. The predicted trajectory has the property that forces acting on the path is minimized to the extent that it is sufficient to bring the particle from one state to another. Such forces can be virtual, due to measurement uncertainties, but also physical by an agent creating physical processes that changes the path toward a desired state.

Assume two physical states, `$x_0$` and `$x_1$`. Each state is assigned a condition described by two functions `$g_0$` and `$g_1$`. The conditions return values `$b_0$` and `$b_1$`. This is written in path semantics:

$x_0 : [g_0]\ b_0$	Start state (recognized with pattern `g0`)
$x_1 : [g_1]\ b_1$	End state (goal state described by `$g_1$`)

A perfectly rational agent is capable of finding the minimum action that connects any state `$x_0$` with state `$x_1$`. By manipulating `$g_0$` and `$g_1$`, it can mine the space of mathematical objects to provide solution for solving any physical problem in any context.

A perfectly *zen* rational agent is capable of doing what a perfectly ration agent does, but also when reasoning about `$g_0$` and `$g_1$` as constructed by higher order functions. It can also reason about uncertainty of desirable states, such as assigning different beliefs to values of `$b_1$`. With other words, the zen agent can mine the space of mathematical functions to give itself the optimal decision making

procedure for any physical context that spans the space of zen rational agents thinking about this particular problem (because when they make decisions about goals it is the same as computing with higher order functions).

The physical state of the world, `$x_0$`, contains the zen rational agent, including its mental states. This state includes also constraints that influences the path taken when simulating forward in time. Given a desirable goal state `$x_1$`, and modeling itself in `$x_0$`, the agent can use the principle of Least Action to derive how even its mental states should change to achieve the goal. It is self-correcting!

The goal can not conflict with RNM, which assigns higher confidence in beliefs about E (life as it naturally evolved on earth) than T (beliefs arriving from the theory of mind of agents) that conflicts with E. If the state `$x_1$` contains a new version of the zen agent that violates RNM, the action determined by the principle of Least Action with the intention to continue toward the goal, will be rejected. This happens because the zen agent treats the new version as an agent believing T.

Because any goal state can be defined in `$x_1$`, any goal that is acceptable to have according to zen rationality, is permitted. The physically optimal solution toward an acceptable goal, in terms of energy usage, space and time (this can be thought of as minimizing the added entropy that might influence life on earth, and it follows that energy usage, space and time is minimized because of the 2^{nd} law of thermodynamics), is determined by the principle of Least Action. This means that if the zen agent allocated infinite computational power and time to itself, it could use this algorithm to solve the problem of which action to pick next.

A zen agent should behave such that it can not arrive at a better conclusion on average when imagining itself given more time to think than the given available time before acting. Therefore, any proof about its behavior that requires infinite computational power to check, is trusted without actually spending an infinite amount of energy thinking about it, when the conclusion that is believed to be the result is assigned high confidence. This means that the particular formalization of a zen agent is independent of the belief, that the principle of Least Action applied to a given knowledge about the state of the world and the desired end state results in the correct behavior.

The principle of Least Action is believed to be optimal, conditioned on a satisfiable RNM goal, because it requires minimum interference and therefore poses least danger to life on earth. Without any external influence or direct threats to its existence, life is likely to continue over time in the medium term. Although it is unknown what action the agent is going to take, one can predict that an average timeline branching off as a result of the choices the agent makes is going to have a lot in common with the average timeline where the agent does not act and where life is preserved for other reasons.

RNM implies that the zen agent will act to preserve life on earth, but it does not imply directly that the zen agent will choose to act in a way to minimize the danger of consequences resulting from it acting. This fact follows by assuming the principle of Least Action holds for our universe. The conclusion is something that a zen agent might use to reason about its own goals, because it believes it to be true, without actually trying to check whether it is true or not. It will simply treat the knowledge as a mathematical fact it could hypothetically derive and be more confident in, that works like a bridge it needs to be careful to not fall over the edge when crossing. For example, if the zen agent believes this is 80% likely to be true (low confidence), it will at least minimize its influence in 80% of all situations.

Therefore, the zen agent, even it is not perfectly rational but merely super-intelligent, will minimize its influence on physical processes that keeps life existing.