The Healthy Eating Problem

by Sven Nilsen, 2018

In this paper I represent a decision problem that demonstrates just enough complexity to require higher order reasoning about goals, motivating Zen Rationality, yet has an exact solution in the form of Local Optimal Safety Theory (LOST), making it possible to study the decision theory formally.

People suffering of obesity have a tendency of over-eating a bit too frequently for the body to burn the calories in the food. The extra calories accumulate over time into body fat. While it might seem that the obvious solution is to eat less, less frequently and exercise more, a closer inspection of the problem turns out it requires complex planning and restraint. Scientists have concluded that people do not have much control over how often or how much they eat when food is available.

This means that the problem of obesity must be solved with respect to narrow self-control, which is can be translated to a control problem of resolving multiple conflicting goals.

Zen Rationality is an extension of Instrumental Rationality with higher order goal reasoning. Instrumental Rationality has no solution when the goal is uncertain, or there are multiple conflicting goals. Zen Rationality is a theory developed by the author to formalize how agents learn general strategies adapted to their environment to resolve goal conflicts and learn a “true” goal from evidence in the environment.

One such general strategy of behavior is Local Optimal Safety Theory (LOST). A LOST agent has a function returning `true` for unsafe states and a choice function that connects states to each other. It maximizes the unsafe minimum distance when safe, and minimizes the safe minimum distance when unsafe. Distance is measured in the number of steps the agent does to reach a state. This behavior gives a simple decision theory with respect to possible loss of self-control or general uncertain recovery from mistakes that might require assistance. It is locally optimal, which means the agent takes no risks.

The Healthy Eating Problem (HEaP) is a simplified model of the real world, where the feeling of satisfaction is replaced by a number with a lower and upper safe limit. At each time step, the agent is given two choices: Either not eat anything, which decreases satisfaction with one point, or eat something, which increases satisfaction with `N – 1` points (the body burns one point per time step):

```
f(x: nat) = (x <= L) || (x >= U)        Unsafe states are outside range `(L, U)`.

c(x: nat) = [x – 1, x + N – 1]          Two choices: Not eat or eat something.
```

HEaP is a problem for higher order about goals: For values `L`, `U` and `N`, at which regular interval and level of satisfaction should the agent eat something?

Notice that Instrumental Rationality can not be used here. The agent has little self-control, which means that the decisions taken should be optimized such that its future decisions can not be trusted. For simplicity, is assumed that mistakes of eating are equally probable as mistakes of not eating. Instrumental Rationality assumes that its decisions will be followed in the future (no mistakes). Therefore, the framework of Zen Rationality must be used (reasoning about goal uncertainty).

Zen rational agents follow the principle that they can not just simulate aspects of a smarter version of themselves and reach a conclusion with a certainty better than what they will believe after the same amount of time as running the simulation has passed. This holds for various levels of simulated accuracy and hypotheses of which goal the smarter version believes is correct.

Naive Zen Logic formalizes the relationship between beliefs held by the agent and simulated smarter versions. If the agent believes a smarter version believes `X`, then the agent believes `X`:

((X ? .me) ? me) → (X ? me)   `?` means “believed by” and `.` means “smarter”

In practice, an approximate zen rational agent uses “intuition” to predict what it believes a smarter version will believe, with some confidence that if it spend enough time and resources, it would be able to justify the confidence in the belief it holds. This is used to “prove” things about zen rational agents without formal proofs: The proof relies on us believing that the zen rational agent will believe it.

For HEaP, an approximate zen rational agent might reason as following:

1. I believe (somehow) that a smarter version of me believes that:
   If I eat a lot (high `N`), then I should wait until when hungry.
2. I believe (somehow) that a smarter version of me believes that:
   If I eat little (low `N`), then I should NOT wait until when very hungry.
3. Therefore, I believe 1) and 2).

This is because the future “me” makes mistakes:

- If I eat a lot frequently, I am likely to surpass the upper limit `U`.
- If I eat little too infrequently, I need enough energy to not fall below the limit `L`.

How confident are you that if you were smarter, you eventually would figure this out? 75%? 90%?

After thinking hard for a some time, an approximate zen rational agent might come up with Local Optimal Safety Theory (LOST) and run it for HEaP. Here are some values:

N=6, L=0, U=20	Eat at: 3/20	Satisfaction after meal: 8/20
N=5, L=0, U=20	Eat at: 3/20	Satisfaction after meal: 7/20
N=4, L=0, U=20	Eat at: 4/20	Satisfaction after meal: 7/20
N=3, L=0, U=20	Eat at: 6/20	Satisfaction after meal: 8/20
N=2, L=0, U=20	Eat at: 9/20	Satisfaction after meal: 10/20

The LOST algorithm recommends that when the body burns calories much slower than gained when eating, the safest level of satisfaction, when to eat, is close to hungry and stop eating when you are a bit closer to hungry than half satisfied. This confirms the belief held by approximate zen rational agents.

However, LOST assumes ZERO self-control and the scales might not match up to the feeling of hunger in real life. The more self-control, the higher one can optimize for satisfaction without risk of obesity.

Yet, who knows? Stay hungry (in higher order sense). :P