The Cursed Diamond Problem

by Sven Nilsen, 2020

Imagine that I tell you there is an expensive, but cursed, diamond under your desk.

Knowing that finding the diamond would change your life significantly,
but perhaps not for the better, would you look under your desk and pick it up anyway?

Or, would you pretend that you never saw the diamond and continue living your life as normal?

A "cursed" diamond is an idea invented by humans to tell stories, where pursuit of wealth leads to more downsides than living a modest, but somewhat satisfying, life. It does not mean that the diamond is actually cursed, but that, although the diamond consists of a relative simple configuration of matter, it changes the environment around it, through interaction with humans. This change can be unpredictable.

This thought experiment is meant to illustrate an important difference between traditional utilitarianism and higher order utilitarianism (which I also call "Zen Rationality").

In traditional utilitarianism, a goal can be expressed as a function. The optimal rational behavior is determined by maximizing the output of the function. From this definition there are some hidden assumptions, such as you need to be able of carrying out plans, there are no limits on time to think, and so on. These hidden assumptions makes certain problems not solvable with traditional utilitarianism.

In higher order utilitarianism, or zen rationality, the goal is not necessary expressible as a function. The goal might be a higher order function where parts of it might be currently unknown. The computational complexity of optimizing for such goals is not always decidable.

Sometimes, there is no way to tell the correct answer to a problem.

The cursed diamond problem has no universal solution.
What is optimal depends on *who is making the decision*.

A zen rational agent can not behave in a such way that it could produce better results by simulating a smarter version of itself, during the available time to think, and determine its actions.

In the case of the cursed diamond thought experiment, a zen rational agent might imagine an improved version of itself, that is better at handling a cursed diamond. However, the agent might prefer to ignore the diamond because it does not trust itself.

At first, this seems to violate zen rationality. If a smarter version can handle the diamond, why does not the agent choose to handle the diamond? Handling the diamond seems to produce better results.

The answer is that a smarter version imagining itself being dumber, put in the agent place, still zen rational, might choose to ignore the diamond. So, ignoring the diamond is still a viable action.

It helps to try formalize what is going on here.

When I formalize this, I will introduce a notion of a proof versus safe moral behavior:

$$\text{proof} : \text{fact} \rightarrow \text{bool}$$

$$\text{safe\_moral\_behavior} : \text{fact} \rightarrow \text{bool}$$

The reason for this is that determining safe and moral behavior is much more complex than finding some sort of proof. By connecting these two using the following law, I am able to simplify the problem:

$$\forall X \{ \neg\text{proof}(X) \Rightarrow \neg\text{safe\_moral\_behavior}(X) \}$$

The `proof` function in this case returns `true` if and only if the agent is able to infer `X` on its own.
If I choose to tell the agent `X`, the agent might choose to ignore it, out of moral safety concerns.
It can imagine itself acting out without knowing `X` and estimate the predictability of the environment.
When the predictability of the environment leads to reduced risks and therefore higher expected utility, the agent can behave as if it did not know `X`, by continue “living its normal life”.

With other words, the agent can “erase” facts from its own mind, using rational decision making.
This ability is not rational to have under traditional utilitarianism, because of its hidden assumptions.
A traditional utilitarian agent will never choose to erase facts after learning them.
However, in zen rationality, this is a rational ability to have, meaning it can be actively used.

The cursed diamond problem belongs to a class of problems where in order to behave safely, after learning a fact `X`, one must compensate for increased complexity by safeguarding against some `X`.

From the laws of logic, the two following expressions have the same truth value:

$$\forall X \{ \neg\text{proof}(X) \Rightarrow \neg\text{safe\_moral\_behavior}(X) \}$$

$$\Leftrightarrow$$

$$\forall X \{ \text{safe\_moral\_behavior}(X) \Rightarrow \text{proof}(X) \}$$

It is not certain that the agent can behave safely by performing these extra tests on itself,
but it is certain *it can not behave safely if it fails to pass these tests*.

By focusing on non-existence of certain proofs, related to safe moral behavior, instead on the definition of safe moral behavior itself, it becomes possible to reason about smarter versions in some way.

A smarter version might use a different `proof` function, using terminology from Naive Zen Logic:

$$\exists X \{ \text{proof}_{me}(X) = \text{false} \wedge \text{proof}_{.me}(X) = \text{true} \}$$

Here, `me` means the agent and `.me` means “smarter me”.

The smarter version is capable of figuring out more stuff on its own. Therefore it might be able to behave safely in contexts such as the cursed diamond problem, by handling higher unpredictability.