Playing With Fire:
A Zen Rational Ethical Dilemma

by Sven Nilsen, 2018

In this paper I represent an ethical dilemma called "Playing With Fire" that is relevant for self-modifying zen rational agents. The problem is what ethical framework should be used to justify operator signal for Intermediate Decision Theories that transform the agent into a state where risks and side effects are balanced versus expected long term gain from mastering states that, although safe, are closer to unsafe states than locally optimal. I also analyze how a Polite Zen Robot migth solve this.

Zen Rationality is an extension of Instrumental Rationality with the ability for higher order reasoning about goals. It is currently believed that Safe Artificial Super Intelligence (ASI) requires some form of approximate zen rational behavior, due to the likelihood of existence of Operator Triggered Intermediate Decision Theories (IDFs) in its design for provable safety, unless a single Decision Theory (DT) can handle this world's complexity.

A self-modifying zen rational agent might extend its own design with new IDFs, which poses new challenges for the ethical reasoning that governs the agent's decision: Since IDFs are triggered by some kind of operator signal, on what basis is it justifiable to design a control mechanism which permits the operator of the agent to transition it into states where risks are taken to maximize long term expected safety?

Consider the human behavior analogue of the problem:

Alice is a child who wishes to become a fire fighter when she grows up. Alice understands that to practice fire fighting, she needs experience with how fire behaves. Alice wants to play with fire in order to become *better* at preventing dangerous fires. So, Alice asks her parents whether she can create small fires in the garden using a fire-safe container.

The problem is NOT that Alice's parents should say "NO, DON'T DO THAT. WAIT UNTIL YOU ARE GROWN UP!". Whether the parents thinks playing with fire is OK, is *irrelevant*, because playing with fire is usually restricted by *law* and not by parents' ability to make *permissions*. Human society is organized in a way such that most parents know there are rules to follow, so there are certain things that parents are responsible for teaching their children.

Similarly, what kind of *law* would be ethical sound for similar situations in Safe ASI designs? Since a Safe ASI might be able of self-modification, an approximate zen rational agent might figure out that extending its design with new IDFs will maximize expected long term safety. These IDFs must be triggered by the operator somehow, but by **which *assumption* is it safe to let the operator decide**?

This dilemma is called "Playing With Fire" due to analogue of a human society that does not let children play with fire, even if their parents says "yes". With other words, when an ASI design is extended with new IDFs, should *every* new operator triggered signal be integrated in the same mechanism for controlling the agent?

Assume that there exists some “law of Safe ASI”, formalized in deontological logic:

1. If the law of Safe ASI **forbids** new IDFs to master close-to-unsafe states,
   then expected long term safety is reduced
2. If the law of Safe ASI **obligates** new IDFs to master close-to-unsafe states,
   then expected short term safety is reduced
3. If the law of Safe ASI **permits** new IDFs to master close-to-unsafe states,
   then what kind of procedure should be used to give such permissions?

It is kind of obvious that making an for-all rule that either sacrifices all safety for the long term or short term in unacceptable. However, *even in the absence* of a Safe ASI authority, the *problem persists*.

Why? Because zen rational agents are expected to behave such that a smarter version of themselves can not be simulated and come up with a better conclusion within the same period of time it thinks about this problem. If it believes a smarter version believes \`X\`, then it believes \`X\`.

For example, a Polite Zen Robot (PZR), a zen rational agent equiped with Rational Natural Morality (RNM) and Common Sense Politeness (CSP), might come up with an idea like this:

1. Only execute in IDFs to master close-to-unsafe states for a limited amount of time,
   before requiring a new operator triggered signal to repeat
2. Publish a log of operator triggered signals than can not be tampered by the operator
3. If the information of the published log is sensitive for the future safety of life on Earth,
   or it is unlikely that this information will prevent the operator from doing something stupid,
   seize control over the extended operator signal and decide protocol itself on case-by-case basis

The two first points are reasonable-sounding behavior because they score high according to CSP. This module of the PZR deals with acceptable behavior in human cultures.

The third point is not that easy to accept because the PZR delays its decision to a future version of itself. With other words, it grants itself higher autonomity with increasing capabilities instead of handing the control over to the operator. This happens because RNM decides whether it can trust the operator to handle the PZR safely (using theory of mind modeling) and also whether CSP should be overridden (sometimes, acceptable behavior in human cultures violates environmental safety).

The expected behavior of a PZR is that it will NOT hand over control of such IDFs to the operator. With other words, if the PZR was Alice, she would not ask her parents, if she believed there was a good reason for not trusting her parents in this matter. Instead, she would decide to rethink the problem on a case-by-case basis later.

A precise behavior of PZR is impossible to predict, however, this example is used to describe what the sort of reasoning that goes behind the PZR’s decision.

This implies that a PZR contains some sort of meta-logic for IDFs. The concept of IDFs alone is not sufficient to determine a PZR’s behavior. For example, an operator signal might be disrupted if it believes that the operator can not be trusted, and e.g. fall back to a simulated mind model of the operator that is more trustworthy.