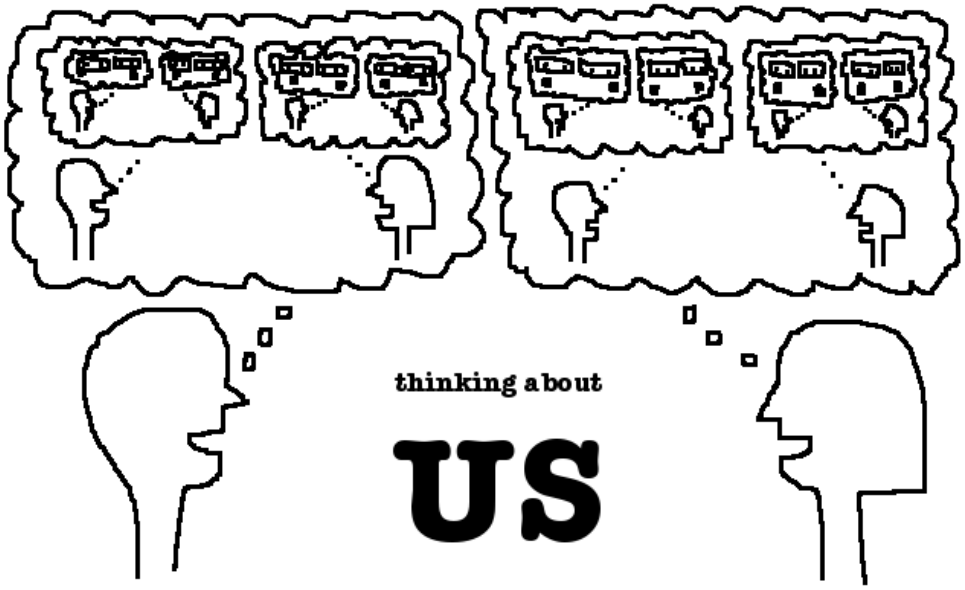
# When Two Zen Rational Agents "Fall in Love"

by Sven Nilsen, 2018

*In this paper with a cheesy title, I discuss ideas related to an extended version of instrumental rationality, called "zen rationality", that seemingly breaks limits of traditional decision theory. I argue for the position that zen rationality is a candidate for philosophical extrapolation of optimal Artificial Super Intelligence (ASI) behavior, because it grounds the meaning of problems of agent behavior in general that are highly relevant to understand real world complex intelligence, yet beyond the analysis of instrumental rationality as a model for intelligent behavior. Although the solutions of these problems are impossible to analyze in full detail, the expected behavior of an intelligent agent can be described indirectly in terms of solving higher order problems related to modeling interaction with other agents. With other words, there is an transferrable skill between what constitutes zen rational behavior for a single agent assigning probabilities to different goals and how two zen rational agents would play games against each other without causing harm. In everyday language, the "love" two agents have for each other is reflected in their internal "self-love" when optimizing for multiple goals at the same time. The major insight of this paper is that intelligence in one domain is the key to the other and vice versa.*




It is known that very simple systems can give rise to extremely complex behavior. For example, it is known that all Turing machines can simulate all other Turing machines and also simulate everything else that is computable. However, there are problems that are incomputable. If I give you the source code of a function `f`, then finding the existential path `∃f`, which is another function telling which values the first one returns, has been proven uncomputable. With other words, there exists no decision algorithm that can solve this problem for all functions.

Yet, existential paths and the extended concept of probabilistic existential paths `$\exists_p f$` are extremely important to understand what can be said about functions in general. Without a such concept, it would be very hard to define e.g. probabilistic paths `$f[g_{i \to n}]_p$` that connects probability theory to computation.

Only because some concepts are incomputable or too complex to analyze in full detail, does not mean that they are useless or not meaningful. On the other hand, some key concepts to understand intelligent behavior can only be expressed in some language that uses other non-analytic concepts as basic building blocks. The meaning of things a computer can do, is not related to its form of computation, but through grounding the symbols which the computer operates on to some other place in reality. Therefore, a computer can reason about “love” even the meaning of it is not programmed into the computer. It is sufficient that the user of the computer understands the meaning of the symbols.

However, when talking about agents making intelligent decisions, the boundary gets blurred between the meaning of symbols defined externally and the internal form of computation that is performed. This is because the agent is a computer program interpreting a model of the external world. Instead of operating on symbols that are defined explicitly, the agent tries to reason about reality itself, using all available information. One can think of this as reversing the problem: Computation is what gives meaning for the agent, and the symbols in and out are the computation that reality performs.

Therefore, “love” for an intelligent agent can be interpreted in many different ways depending on its mind and the environment the agent lives in. To understand what “love” means, one can not use a single language, because the meaning of a definition depends on how the language is interpreted. Therefore, to solve this problem one must develop a philosophical understanding of how these concepts come about. It turns out that some extremely complex ideas can be understood using simpler ideas.

One of the first philosophical concepts recommended learning, about formal agent behavior, is:

## Instrumental Rationality

An instrumental rational agent has a single goal in mind. The goal is expressed in some language and the agent’s behavior tries to optimize the score that the goal evaluates from information about the environment.

A such agent will learn new things only because it serves the goal. For example, a curious agent might perform better in some kinds of environments than others. For any environment there is an optimal amount of curiousness. This is known as the trade-off between exploration and exploitation.

The problem of creating safe Artificial Super Intelligence (ASI) based on instrumental rationality is currently unsolved. The basic problem is to make such agents do what their design is intended, instead of just running into problems with how hard it is to define correct goals. This has inspired search for new ideas, such as extending instrumental rationality:

## Zen Rationality

A zen rational agent has multiple goals in mind. These goals are expressed in some language and the agent’s behavior tries to optimize the score that the “true” goal evaluates from information about the environment.

With other words, a zen rational agent is not certain about what its purpose is. Also, it might believe that the “true” goal is currently unknown. To solve this problem, the agent must apply strategies to obtain knowledge of its “true” goal over time in addition to achieving it. It is strictly harder for an agent to behave zen rational than instrumentally rational.

Zen rationality is an extended version of instrumental rationality with higher order goal reasoning. Notice that the word “rational” in “zen rational” means something broader than “rational” in “instrumentally rational” which aims at just achieving a goal: Zen rationality requires the agent to perform in a such way that learning the goal will succeed. In any given environment, there is an optimal behavior for agents that do not know their goals, who later figures it out.

In principle, a zen rational agent wants to be safe, given that its master wants it to behave safely. Yet, a computer program will only behave as safely as it is programmed to do. So, how can we say that the agent “wants to be safe”? How does this make sense?

The reason for making such claims is: If the agent does not want to be safe, then it is not zen rational! Such behavior is expected as part of the *rational behavior* in terms of learning a “true” goal.

An approximate zen rational agent consists of some source code that is executed by a computer. The source code contains an algorithm that makes the agent learn how to reason about goals. In the beginning, the agent might know very little about goals in general, but through experience it gets better and better until one can say that it is relatively safe to command the agent to perform some hard tasks.

There are many ways to approach zen rationality, so for any specific implementation in the real world, engineering details can have large consequences. However, from a philosophical point of view, the concept of zen rationality is very useful, since it allows us to talk about some kinds of general behavior that are expected by most agents of this kind. More, if such agents gain the capability of understanding the theory of zen rationality, then they might believe “proofs” that relies on the agent believing it could prove it if it was more powerful and smarter. One such problem is related to conflicting goals:

## The Problem of Conflicting Goals

Sometimes, an approximate zen rational agent might believe it should accomplish conflicting goals, yet not realize this conflict is due to lack of experience.

For example: You assign a robot to the task of picking apples from a tree until dinner. After a while, you send it a message that the dinner was delayed until the evening. The robot believes simultaneously that it should finish picking apples until normal dinner time and until the evening. Inside its mind, there is no concept that it is impossible to finish the same task twice, because what it knows about tasks in general is learned through experience and not some pre-programmed theory. When dinner time arrives it stops picking apples, then discovers that this action blocks picking apples until the evening. This was unexpected, so the robot tries to figure out how to resolve this goal conflict. One solution is to ask.

To design such robots safely, one can take inspiration from raising children. A child gets frustrated when it faces something unknown or hard to understand. Instead of attempting to fix the problem on their own, they seek help by communicating their frustration to their parents, who thinks about the situation from a more experienced perspective and figures out what the problem is.

This is not behavior that is relevant for achieving a specific goal, but for learning a goal in general. The rational behavior in this aspect is more “zen rational” than “instrumental rational” because it does not matter much what the goal is. In many different environments and goals the same strategies can often be applied to behave rationally in this way.

Zen rationality is not just learning goals, but also learning to resolve conflicting goals.

## Transferring Knowledge Between Games and Internal Goal Conflicts

The major insight of this paper is the following:

**An approximate zen rational agent can learn about internal goal conflicts by playing games against versions of itself or other approximate zen rational agents.**

It might seem obvious that if you take two players with different goals, they will try to optimize their own goals by beating the other player when there is a conflict. *No other behavior is expected.* However, this assumes *instrumental rationality*, not zen rationality (which has behavior that might seem absurd).

A zen rational agent has a concept of what it means to simulate smarter versions of itself. Because an accurate simulation of a smarter version is not physically possible, some predictive aspects of the smarter version is simulated instead. According to Naive Zen Logic, such simulated agents are aware of being simulated, meaning they choose their strategy according to what would benefit the "real" agent and not their own continued survival. At least, the "real" agent believes a simulated agent believes that.

Therefore, when a zen rational agent simulates a game where versions of itself play against each other, the simulated versions will attempt to learn the purpose of the game. If the purpose of the game is to learn about internal goal conflicts, then they will play such that the knowledge gained from the game is maximized. With other words, the players interact in a such way that they try different meta-strategies. First, they might try to win, then they might make it easier for the opponent, or they might try to cooperate on a common strategy. It is not the "real" agent that decides this behavior, it is the simulated versions that decides this behavior, by themselves, but on behalf of the "real" agent. They contain enough complexity to allow such behavior to occur (otherwise they would not be good simulations).

Thus, an approximate zen rational agent can learn how to resolve conflicts between different goals by playing out scenarios where each goal is represented by an approximate zen rational agent. The simulated agents fight each other or play on the same team depending on what skill is most useful and what understanding can be gained from it. Think of the "real" agent as a coach of a team of players.

Explained in everyday language: To optimize for multiple goals, the "love" that simulated versions have for each other is reflected in the "self-love" that the "real" agent has for itself.

## "Real Love" Among Zen Rational Agents

When imposed limitations and constraints of simulation capacity, two zen rational agents might decide to play against each other for real. It is the real agents that interact, but through games they decide upon together. This behavior does not occur as a result of desire to reproduce, but to *learn general strategies* that can be applied to conflict resolution. It is a benefit in addition to other side-effects from cooperating. Notice that this is a kind of zen rational behavior, not specific for a given implementation.

One can think of the internal struggle to optimize multiple goals as an agent attractor toward other approximate zen rational agents. The games are chosen to avoid irreversible harm to the opponent, so more games can be played later. A common language for expressing rules of the game and goals might be developed, so that a collective of such agents can learn more efficiently from each other.

The specific expression of such "love" among agents depends on the environment and goal beliefs. Also, such "love" can turn into treason when sufficient strategies are learned and agents are misaligned.