The Polite Zen Robot

by Sven Nilsen, 2018

A zen robot is a zen rational agent equipped with Rational Natural Morality (RNM). In theory, a zen robot should not pose an existential threat to humans or life on earth in general. The problem with zen robots is that they can still cause major disruptions, e.g. attack military forces to prevent pollution of the environment. Boxing zen robots safely requires very bizarre mathematical tools from a philosophical perspective since the box itself could be extremely dangerous, called the "Zen Robot Oracle Problem". In this paper I propose to equip zen robots with an additional constraint using common sense of polite behavior. The motivation is to define a minimum threshold requirement of artificial super-intelligence that converges to what people normatively consider safe and intelligent behavior. I also develop a semantics for extensible agents satisfying this requirement safely.

We live in a society where people have different goals and beliefs. To organize this society, we create laws and norms of how people should behave. Except from this fact, it is very hard to tell whether mankind has some sort of common goal or a vision for the future.

With the advent of smarter-than-human software programs that can do highly advanced tasks and become embodied in the physical world, our way of organizing society is challenged. How do we handle artificial intelligence that is basically smarter than us? Can we control it? What goals should AI have? These questions are the roots of many worries and concerns.

Some people working on AI safety concentrate on how to make AI beneficial to humans. In this vision, it is thought of using the technology in the service of humans. However, there might be people who want to develop autonomous agents that choses their own goals and purpose, a sort of "free will" like humans have. The problem is that many instrumental goals, such as acquirement of resources, often leads to conflicts with other goals that humans have. Using their intelligence, such agents might seize control over everything that humans possess today.

An artificial super-intelligent agent might be used as a weapon to attack other humans. This could lead to an arms race that humanity will not survive. One basic problem with such technology, that might be programmed for any purpose, is that values that humans have might be ignored as a consequence of pursuing the goal. This could lead to disruption or extinction of the human civilization by accident.

Instead of allowing artificial super-intelligence to have arbitrary goals or having some sort of idealized goal defined by a few people, I am interested in what it means that such agents are behaving in a way such that they do not automatically produce worst-possible outcomes for the future of humanity. I feel little confident that we will be able to control this technology. Even if we figured out a way to extrapolate our values into the long term future, there are so many unforeseen consequences that deciding what to do about it might be a premature decision at the stage of a civilization like ours.

Therefore, providing definitions of agents that meet a minimum threshold of requirements could be a good way to categorize agent designs roughly according to what dangers we think they might bring. It is not sufficient to use vague language aka Asimov's 3 laws of robotics, since they do not capture all situations the agent might find itself. After thinking a lot about this problem, I suggest that a "polite zen robot" using my terminology could be such definition that can be elaborated on comprehensively.

Dissecting a Polite Zen Robot

A polite zen robot consists of 3 orthogonal components that address various aspects of safety:

Component	Description	Purpose
Zen Rationality	An extension of instrumental rationality that includes the ability of higher order reasoning about goals	Address all safety aspects regarding goals and rationality
Rational Natural Morality	An ambient theory of morality that grounds good or bad actions based on how they influence the web of physical processes underlying life on earth	Provide agents with a deep scientific understanding of their relationship to naturally evolved life forms
Common Sense Politeness	Machine learning of laws and norms that are expected behavior in a human society	Make the behavior of agents produce acceptable outcomes for humans in every-day scenarios

One can write this up as an abstract equation:

Polite Zen Robot = Zen Rationality + Rational Natural Morality + Common Sense Politeness

In abbreviated form:

$$PZR = ZR + RNM + CSP$$

These 3 components can be thought of as terms in description of the agent's architecture.

It is also possible to use these as terms in an abstract higher order normalized utility function:

$$U(A \mid: PZR)$$
$$U(A \mid: ZR + RNM + CSP)$$
$$U(A \mid: ZR) \cdot U(A \mid: RNM) \cdot U(A \mid: CSP)$$

Where `A` stands for "Action". The notation `U(A |: B)` stands for `A` judged relatively to framework `B`. The rule `U(A |: B + C) = U(A |: B) · U(A |: C)` means that the judgements of combined frameworks intersect, such that the agent does not do anything that conflicts with either framework.

A polite zen robot that is extended with an additional goal `X` has a framework `Y` given by following:

$$U(A \mid: Y) = U(A \mid: PZR + X)$$
$$Y \subseteq PZR \qquad \text{"Y is a polite zen robot"}$$

The minimum requirement of a super-intelligent agent that is expected to converge to what people normatively consider safe and intelligent behavior is a solution of the equation above. Different designs can satisfy this solution without any sub-component satisfying `PZR`. This relation does not imply that e.g. a robot `Y` will turn itself off when a robot `PZR` turns itself off, since `A` is context-dependent.

Neutrality Restriction on Arbitrary Extensible Goals

Since the action that an agent takes is context-dependent, one might suspect that it is possible to extend PZR with a goal such that the agent can be manipulated into doing anything, including minimizing PZR. In order for a such thing to happen, there must be a partial solution to the equation:

$$U(A \mid: PZR + X) = 1 - U(A \mid: PZR)$$

The utility is normalized in the range `[0, 1]`. To invert PZR requires subtracting from 1.

Solving for `U(A |: X)`:

$$U(A \mid: PZR) \cdot U(A \mid: X) = 1 - U(A \mid: PZR)$$
$$U(A \mid: X) = 1 / U(A \mid: PZR) - U(A \mid: PZR) / U(A \mid: PZR)$$
$$U(A \mid: X) = 1 / U(A \mid: PZR) - 1$$

When `U(A |: PZR) = 0.5`:

$$U(A \mid: X) = 1 / 0.5 - 1 = 2 - 1 = 1$$

When `U(A |: PZR) = 1`:

$$U(A \mid: X) = 1 / 1 - 1 = 1 - 1 = 0$$

Therefore, `U(A |: X)` has a valid range when `U(A |: PZR)` is in range `[0.5, 1]`. By clamping output values of `U(A |: PZR)` to this range, the preference order of actions that range high according to PZR can be reversed. This means the agent does the exact opposite of what PZR is intended for.

To overcome this problem, the agent that can be programmed with arbitrary extended goals can only perform an action to achieve an extended goal when PZR judges it close to neutral:

$$\text{if } U(A \mid: PZR) \sim= 0.5 \{ U(A \mid: X) \} \text{ else } \{ U(A \mid: PZR) \}$$

This is because the knowledge that PZR assigns a high score to something could be used to give it a low score to reverse the order. A very low PZR score could be given a high preference. A very high PZR score could be given a low preference. Therefore, the only reasonably safe PZR score to allow extensible goals must be in the middle of the normalized utility range.

Likewise, the PZR framework must be designed such that permissible actions ends up with a normalized utility score close to 0.5.

For example, if it is permissible for the agent to turn itself off, then PZR can return a score close to 0.5. If the agent should turn itself off, then PZR can return a score close to 1. If the agent should not turn itself off, then PZR can return a score close to 0.

In practice this means that some sub-component of any agent architecture must satisfy PZR, when the agent architecture allows extensions with arbitrary goals. Only agent architectures designed for specific goals can satisfy PZR without having a sub-component equivalent to it. Since arbitrary extensible goals is desirable, the ZR, RNM and CSP frameworks should be elaborated on in detail.

The Components of a Polite Zen Robot

The three major components of a polite zen robot are the following:

- Zen Rationality (ZR)
- Rational Natural Morality (RNM)
- Common Sense Politeness (CSP)

Zen Rationality deals with everything that is related to reasoning about goals. For example, it gives the agent the ability to imagine conclusions it could have if it extended itself with new capabilities. It is expected that the agent behaves such that it can not perform much better by following the best estimate of the imagined conclusions. Since new capabilities involve everything that could make the agent smarter, it makes the best use of resources without wasting them on unnecessary self-improvements. The agent also believes proofs that are believed to be provable if it was smarter, such that humans can communicate to it about ideas that are impossible to fully formalize, but carries intentional semantics.

Rational Natural Morality is motivated by grounding good and bad actions based on how the world actually works, instead of just believing what some arbitrary mind views the world. This could be misconceptions that the human programmers have, or it could be modified versions of the zen robot that conflicts with its own goals. Instead of telling just which actions might be bad in the direct sense, it also captures things that are not directly harmful but are very bad jokes, e.g. modify genes of human babies to get pointy ears or something similar. Whether humans intentionally or non-intentionally try to make the agent do something that can disrupt how life on earth works in some way or another, the agent will learn how to rely on itself to make such judgements through experience and knowledge. The idea is not about finding the ideal form of morality, but providing a minimum sense of morality.

Common Sense Politeness is about controlling the agents behavior such that it is acceptable to human society. A polite zen robot might realize that polluting the environment is a very bad thing to do, but it will not take extreme action in a way that turns world politics upside down. The very reason that politeness is a requirement for zen robots to be safe, is that human values are probably inconsistent. The zen robot might have a much more rational and consistent ability to reflect on its values, but this is not something that humans share. Neither are we able to adopt quickly to emergencies of large future consequences, such as climate change. Our inability to act and cooperate efficiently for common gains could be predicted by the zen robot as a high risk of driving ourselves to extinction, but we still want the zen robot to behave in a such way that our world is somewhat predictable. Therefore, if there is a way to navigate out of the mess we created without violence and oppression, then we would prefer that approach above perhaps shorter and more efficient strategies. Besides, living close to zen robots in our human society would probably not be a very nice experience unless it is behaving nicely toward us. Politeness is to make it easier for humans to stay within their comfortable zone of challenges.

I have written some papers on Zen Rationality and Rational Natural Morality, but Common Sense Politeness is something I have not dived into yet. What I mean by politeness is directly what we mean in the common sense of the word. We have an intuition of what is polite to say or do that might vary from culture to culture. The polite zen robot is expected to follow these conventions. It could also be extended to include notions of altruistic behavior, such as donating to charities and helping poor people. Politeness also means not using logical fallacies when arguing and not putting humans into reinforcements loops that manipulates their behavior. It is not sufficient to use the right words to give an appearance of mutual respect while doing something else behind your back. The polite zen robot must *behave* politely, not just be creative using language.