Infinitesimal Utility

by Sven Nilsen, 2018

In this paper I argue that almost all commercial application of AI should be considered infinitesimal utility. An infinitesimal utility can not be traded with utility points that captures human values in general. I suggest this to avoid the false assumption that any action is permissible to maximize a commercial goal. Utility functions should be bounded and trigger an alarm when exceeding these bounds. Safety concerns should trigger the alarm in negative direction.

Previously, in the paper “The Polite Zen Robot” (Nilsen, 2018) I outlined an approach to think about safely extensible super-intelligent agents. The major point of this paper was that neutral judgements are necessary when the decision procedure of the extension is arbitrary. A polite zen robot (PZR) makes these judgements from 3 modules that addresses different aspects of fundamental safety: Zen Rationality (higher order reasoning about goals), Rational Natural Morality (the relationship between theory of mind and physical processes underlying life on earth), Common Sense Politeness (cultural acceptable behavior in a human civilization). Each of these modules are believed to be AI-complete, that is, about as hard as creating Artificial General Intelligence (AGI). To build a polite zen robot requires first solving the AGI problem.

Dr. Eric Purdy developed a similar idea directly applied to machine learning in his paper “Grandiosity and Paranoia in the Reinforcement Learning Context: Lessons for AI Safety” (Purdy, 2018). This work shows that it is practically feasible to train agents on bounded utility functions, inspired by the role of serotonin in the brain.

Building on these two ideas, I will then make the following argument:

- Commercial applications of AI should use bounded utility functions
- This technique is justified by interpreting the utility function as infinitesimal
- The utility function that captures human values in the big picture is approximated by PZR

Any action that conflicts with PZR violates one of the 3 modules ensuring safety. This framework is a good approximation for capturing some major human values across a large variety of situations. It means that even in the absence of a near term practical implementation of PZR, all AI applications should be designed in a way such that they lead to results as if PZR already existed.

However, AI applications use a utility function where the advice of PZR is absent. To solve this problem, the utility function should be bounded. If the range is `[0, 1]`, then expected rewards from achieving goals should be near `1` and failure modes near `0`. When expected rewards exceed these limits, an alarm should be triggered to warn operators of potential unsafe situations.

Mathematically, the whole utility can be thought of as a dual number:

$$pzr + app\ \varepsilon$$

The real part `pzr` consists of expected returns from judgements by PZR. The infinitesimal part `app` consists of expected returns from judgements by the commercial application.

A dual number has the property that the real part can not be exchanged with the infinitesimal part without loosing information. Another way to phrase this is that in the infinitesimal part, there exists no value that describes any quantity of the real part, except zero.

$$\varepsilon^2 = 0$$
$$\varepsilon = 1 / \infty$$

Infinity `∞` is not a number on the real line (the real line plus infinite is called the projective reals). When using this relationship between the two terms in a dual number, the two quantities are separate, but they preserve the order:

$$a + \varepsilon > a$$

In an unbounded utility function, all values corresponds to some expected reward, no matter how large or small they are.

For example, an agent is designed to build doors. If it is programmed to build as many doors it can in an unbounded utility function, then a Bayesian rational agent would sacrifice any other means to produce just one more door. Even if the utility function included a term for how much a human life was worth, there would be some amount of doors that corresponded exactly to a human life (according to the utility function).

However, when there exists no value that corresponds to a human life, a Bayesian rational agent can not reason about how much human lives are worth. The moment it starts to reason like this, it will lead to an error (e.g. a floating point error).

Yet, AI frameworks rarely support dual numbers in the reward function.

To fix this problem, one must use a bounded utility function. Instead of creating an infinitely larger value, the range is fixed such that the AI application raises an alarm when entering the zone outside its expected returns.

This is not about human lives having infinite worth to someone, but restricting the domain where the AI used in every-day commercial applications operates safely. When it leaves the safe zone, exceptionally careful guidelines must be created to show how such AI systems should be designed.

It is this safety property that permits us to think of the relationship between these values as a dual number. From a safety perspective, a bounded utility function with expected rewards near `1` and failure modes near `0` is compatible with the concept of a hidden implied background of infinite higher values. This justifies that the AI system tries to shut down when encountering values outside the range, because in that mode it can not reason about what to do (since the PZR does not exist yet).

One benefit of this approach is that when reasoning about super-intelligent AI systems, one can use dual numbers to allow a continuous transition between the two kinds of utility functions without the failure modes of limited commercial AI systems. Combined this gives both a practical guideline for designing safe AI systems for the real world and a foundation for the semantics of interactions between theoretical models such as the PZR and its interactions with other AI systems.