Differential Utility Complexity

by Sven Nilsen, 2022

In this paper I outline an idea that might help to reason about social higher order utility.
This idea might be important for AI alignment problems.

In Instrumental Rationality[1] (IR), an agent optimises a utility function. One can think about the utility function that takes a type, representing the environment and outputs a real number:

utility : environment → real

In the paper "Zen Rationality"[2] (ZR), I presented an extended version of IR. One can think about Zen Rationality as higher order theory of IR. The correspondence between IR and ZR is the same as the difference between simply typed functions and higher order functions. IR optimizes utility for simply typed functions, which can be well defined in many cases. ZR optimizes utility for higher order functions, which is often not well defined and depends on the constraints one can use to reason about utility.

Differential Utility Complexity is an abstract mathematical language for constraining behaviour of higher order utility functions such that they produce adequate normative higher order social utility. This language has two steps. The first step is an oracle `duc` which takes some environment, a utility function over that environment and produces a differential complexity measure.

duc : environment × (environment → real) → differential_complexity

The second step uses differential complexity. The differential complexity measure can be thought of a function that outputs some complexity measure for relative changes in the utility function:

differential_complexity : real → complexity

The type of complexity is unknown and can vary between applications.

For example, the complexity can depend on which agent is doing the optimization.

With other words, the `duc` oracle can tell how hard it is for an agent to produce outcomes of better or worse utility. The complexity of increasing utility might be used as an indicator to detect reward hacking. The complexity of decreasing utility might be used as an indicator of environment safety.

For example, imagine that you could flip bits directly in the number that represents the balance in your bank account. This would mean that there is very low complexity to reach any new state of wealth or poverty. From a social point of view in economics, it is undesirable that people can flip bits directly like this, because it would ruin the value of money. From a personal point of view, this ability might have very high instrumental utility. However, a person who wants to live in a society where the economy is functional and safe, might inform the bank in the event of a such opportunity.

An agent for which differential complexity is lower in the direction of increasing utility, than in the direction of decreasing utility, might be thought of as existing in some environment that encourages optimizing utility. Thus, differential complexity might be used to reason about how safe an environment is for different kinds of agents. If the environment contains multiple agents, then they might cooperate to increase their utility, which decreases differential complexity of increased utility.

References:

[1] “Instrumental Rationality”
Stanford Encyclopedia of Philosophy
https://plato.stanford.edu/entries/rationality-instrumental/

[2] “Zen Rationality”
Sven Nilsen, 2018
https://github.com/advancedresearch/path_semantics/blob/master/papers-wip/zen-rationality.pdf