Friend vs Enemy Asymmetry

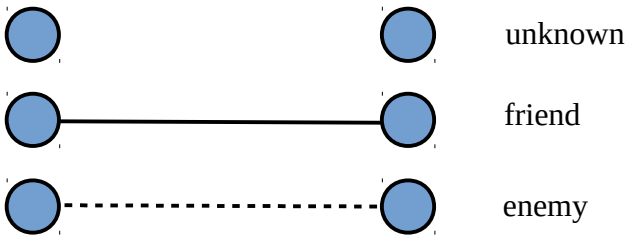
by Sven Nilsen, 2019

In this paper I show that there exists an asymmetry between friend vs enemy relations, which can be used to distinguish between these two classes of behavior. This proof might be used by zen rational agents to safety check the consequences of their estimated utility functions.

Zen Rationality[1] is an extension of Instrumental Rationality[2] with higher order reasoning about goals. One hypothesis about constructing safe super-intelligence is that approximate Zen Rationality is needed, plus an ambient grounding of morality and understanding of human culture.
This hypothesis is called “Polite Zen Robot”[3], developed by the author.

So far, very little work is done on how to make intelligent agents understand human culture.

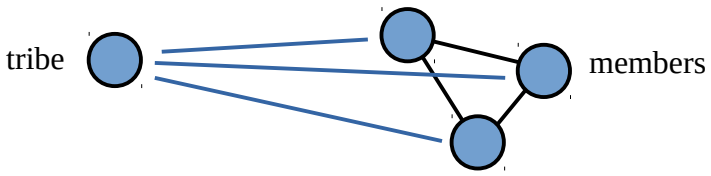
One of the primary features of human culture is that humans model “relations” among themselves. Some humans are friends, while others are enemies. The simplest model, while displaying some realistic features, is one where a relation is modeled as unknown, friend, or enemy:



The nodes are people, while edges are undirected and represent relations.

The complexity of such diagrams is `O($N^2$)`, which is not possible to fit inside human brains for large values. A solution to this problem is to model tribes of people and memberships of tribes, where tribes can have relations among themselves. A tribe is consistent if and only if every member in the tribe are friends among themselves, friends with all members of friendly tribes and enemies with all members of hostile tribes.

In the simple model without tribes, the friend vs enemy relations are symmetric. You get the same theory when swapping these two values. However, when extending the simple model with tribes, one interprets the identity relation of each tribe, which projects onto relations between members of a tribe among themselves. The natural interpretation is to give this relation a value of friendship. This means that there exists an asymmetry between friend and enemy relations in the extended model, which might be used to safety check consequences of estimated utility functions, by requiring that the Polite Zen Robot produces behavior toward humans that is similar to behavior among members of same tribe.



References:

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

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

[3] “The Polite Zen Robot”
Sven Nilsen, 2018
https://github.com/advancedresearch/path_semantics/blob/master/papers-wip/the-polite-zen-robot.pdf