Graph Theory Analysis:

The Health Graph of an Agent:

- A Reinforcement Learning Agent α is formally three things: a Markov Decision Process (MDP) $M_\alpha = (S, A, P, R)$, a policy π produced by an algorithm from a space of possible policies Π_α and the goal of maximizing (discounted) summed reward.
- Let us consider the agents perspective by constructing an equivalency relation: $\forall s_1, s_2 \in S : s_1 \equiv_\alpha s_2 := \forall \pi \in \Pi_\alpha : \pi(s_1) = \pi(s_2)$
- Let us then construct the environmental state graph $\mathbb{S}_\alpha$, a multigraph where vertices correspond to equivalency classes of the relation above, and edges correspond to steps between states weighted by reward.
- The Health Graph $\mathbb{H}_\alpha$ consists of vertices corresponding to strongly connected components of $\mathbb{S}_\alpha$ where vertices corresponding to death have been combined, and vertices corresponding to running out of time have been thrown out. Not only do $E(\mathbb{H}_\alpha)$ correspond to the edges between distinct strongly connected components of $\mathbb{S}_\alpha$ but also the loops with weight r on vertex v correspond to cycles with arithmetic average weight a on strongly connected component s.

Strictly Routinely Suicidal:

We define an agent as strictly routinely suicidal $\alpha \in \mathbf{S}_{R<}$ in terms of simpler properties:

- the unappealingness of routines $\mathbf{S}_{r<}$
- the relative goodness of damaging the situation $\mathbf{S}_{d<}$
- having worse situational health doesnt feel worse $\mathbf{S}_{h<}$
- the despair of being trapped by options $\mathbf{S}_o$
- the despair of being trapped by situation $\mathbf{S}_s$
- the unpreventable nature of death $\mathbf{S}_u$
- the agent views death as peaceful or doesnt think about it $\mathbf{S}_p$.

Here, we will mathematically define the subparts which will be changed by our intervention:

- The agent finds routines unappealing/boring/appealing $\alpha \in \mathbf{S}_{rB} := \forall (v, v, r) \in E(\mathbb{H}_\alpha) : r \, B \, 0$ where B is a binary relation. The agent finds routines unappealing is $\alpha \in \mathbf{S}_{r<}$

- The agent perceives taking damaging actions as relatively good/neutral/bad. $\alpha \in \mathbf{S}_{dB} := \forall(v,v,r),(v,n,w) \in E(\mathbb{H}_\alpha) : r\ B\ w$ where B is a binary relation. The agent perceives taking damaging actions as relatively good is $\alpha \in \mathbf{S}_{d<}$

Safety Intervention:

We focus on the two subproperties which can be changed by intervening on the reward function $\mathbf{S}_{r<}, \mathbf{S}_{d<}$ The safety intervention has two steps, one is to give the agent a reward equal to twice the negative of the weight of the lowest weighted loop in the health graph, the second is to remove any planned incentivization of death by making the episode not end at the end of the game, but rather to have an aftergame where remaining in the end state is incentivized. This will change $\mathbf{S}_{r<}, \mathbf{S}_{d<}$ to $\mathbf{S}_{r>}, \mathbf{S}_{d\geq}$ We model the aftergame in order to preserve the nature of the default reward of -1 providing urgency to the agent, while still removing the suicidal effects of this reward.