1 Tying things together: Preferences, Goals, Utilities, and Rewards

The setting we would like to consider is as follows. Suppose α is an agent in an evolving world. The set of all possible states of the world is $\mathcal{W}$, and evolves naturally according to some (non-deterministic) function $\tau : \mathcal{W} \to 2^{\mathcal{W}}$. Suppose that the word is currently at state $w_0 \in \mathcal{W}$.

We will show that the following four things are indistinguishable:

1. Agent α has preferences over infinite future trajectories of the form $[w_0, w_1, \ldots]$ — that is, it has preferences over objects $\mathbb{N} \to \mathcal{W}$