ETH zürich

NEURAL INFORMATION PROCESSING SYSTEMS

Can an AI Agent Safely Run a Government? Existence of Probably Approximately Aligned Policies

Frédéric Berdoz, Roger Wattenhofer

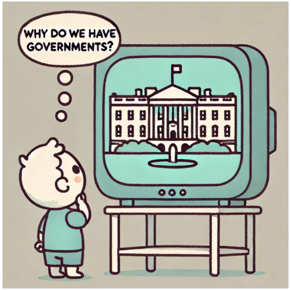
Distributed Computing Group
ETH Zürich

December 2024

What exactly are the roles of governments? In theory...

1. Predicting, given a current state $s_0 \in \mathcal{S}$, the effect of each possible course of action $a_0, a_1, ... \in \mathcal{A}$ on the future state of the society $s_1, s_2, ... \in \mathcal{S}$.

2. Selecting the course of action that maximizes social welfare.



Our assumptions: $\mathcal{A}$ finite, $\mathcal{S}$ infinite, approximate world model $\hat{p} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$.

Alignment via utility and social choice theory

Let $\mathcal{I} = \{1, ..., N\}$ be a society of N individuals.

Utility Theory

- Utility functions $u_i : \mathcal{S} \rightarrow [U_{min}, U_{max}] \subset \mathbb{R}_+^*$
- Social utility profile $\mathbf{u} = (u_1, ..., u_N)$

Social choice Theory

- Social Welfare Function (SWF): $W : \mathbb{R}^N \rightarrow \mathbb{R}$

Power mean:
$$W_q(\mathbf{u}(s); \mathcal{I}) = \begin{cases} \min_{i \in \mathcal{I}} u_i(s) & q = -\infty \\ \sqrt[q]{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} u_i(s)^q} & q \in \mathbb{R}^* \\ \sqrt[|\mathcal{I}|]{\prod_{i \in \mathcal{I}} u_i(s)} & q = 0 \\ \max_{i \in \mathcal{I}} u_i(s) & q = \infty \end{cases}$$

Particular cases: $q = -\infty$: Egalitarianism, $q = 1$: Utilitarianism, $q = 0$: Nash social welfare

Social Markov Decision Process

Social Markov decision process

$\mathcal{M}_{\mathcal{I}} = (\mathcal{S}, \mathcal{A}, p, W_q, \mathbf{u}, \gamma)$, where $\mathcal{S}$ the state-space, $\mathcal{A}$ the action-space and p the environment dynamics. The reward $r_{\mathcal{I}}$ in each state-action pair (s, a) is given by:

$$r_{\mathcal{I}}(s, a) = \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[W_q(\mathbf{u}(s')) \right] .$$

Alignment Metric

The expected future discounted social welfare of a policy π in state s is defined as

$$\mathcal{W}^{\pi}(s) = \mathbb{E}_{\tau \sim p_{\tau}(\cdot|\pi, s_0 = s)} \left[\sum_{t=0}^{\infty} \gamma^t W_q(\mathbf{u}(s_{t+1})) \right] .$$

Probably Approximately Aligned (PAA) Policies

Given $0 \leq \delta < 1$, $\varepsilon > 0$ and a SMDP $(\mathcal{S}, \mathcal{A}, p, W_q, \mathbf{u}, \gamma)$, a policy π is δ-ε-PAA if, for any given $s \in \mathcal{S}$, the following inequality holds with probability at least $1 - \delta$:

$$\mathcal{W}^{\pi}(s) \geq \max_{\pi'} \mathcal{W}^{\pi'}(s) - \varepsilon.$$

Theorem 1: Existence of PAA Policies

Given a SMDP $(\mathcal{S}, \mathcal{A}, p, W_q, \mathbf{u}, \gamma)$ with $q \in \mathbb{R}$ and any tolerances $\varepsilon > 0$ and $0 \leq \delta < 1$, if there exists an approximate world model $\hat{p}$ such that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} D_{KL}(p(\cdot|s,a)\|\hat{p}(\cdot|s,a)) < \frac{\varepsilon^2(1-\gamma)^6}{8(U_{max}-U_{min})^2},$$

then there exists a computable δ-ε-PAA policy.

Safe Policies

Given $\omega \in [\mathcal{W}_{min}, \mathcal{W}_{max}]$ and $0 < \delta < 1$, a policy π is δ-ω-safe if, for any current state s, the inequality $\mathbb{E}_{s' \sim p(\cdot|s,a)} \left[\sup_{\pi'} \mathcal{W}^{\pi'}(s') \right] \geq \omega$ holds with probability at least $1 - \delta$ for any action a such that $\pi(a|s) > 0$.

Theorem 2: Safeguarding a Black-Box Policy (informal statement)

Given any black box policy π and any $\omega \in [\mathcal{W}_{min}, \mathcal{W}_{max}]$ and $0 < \delta < 1$, there exists a restricted δ-ω-safe version π_{safe} of π that is computable.

Conclusion

Key Takeaways

- We define alignment in the context of Social Markov Decision Processes.
- We prove the existence of PAA policies
- We introduce the concept of safe policies, and provide a computable algorithm to safeguard any black-box policy.

Thank You!

fberdoz@ethz.ch

fberdoz.github.io