Pluralistic Alignment Over Time ⏳

Toryn Q. Klassen,
Parand A. Alamdari,
Sheila A. McIlraith

Introduction

If an AI system **makes decisions over time**, how should we evaluate how aligned it is with a group of stakeholders (who may have conflicting values and preferences)?

We suggest how a recent approach to evaluating fairness over time could be applied to a new form of pluralistic alignment: **temporal pluralism**, in which the AI system reflects different stakeholders' values at different times.

Temporal aspects of pluralistic alignment

Group preference change over time

Temporally extended preferences

Alignment with different stakeholders at different times

Our paper further discusses all these; e.g., some temporally extended preferences can be described using **reward machines** (Toro Icarte et al., 2018).

Temporal pluralism

It may not be possible to satisfy everyone with a single decision, but a sequence of decisions can reflect a diversity of values. An AI system that is **temporally pluralistic** reflects different stakeholders' values at different times.

In the following, we adapt some of the fairness-related notions from our prior work **(Alamdari et al., 2024)**.

Restaurant example

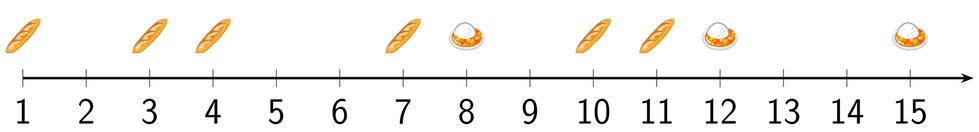
Scenario (inspired by Lackner (2020)):

- An AI assistant is **booking restaurants** for the frequent joint dinners of a group of three friends.
- Two friends prefer restaurants which serve 🥖, while one prefers restaurants serving 🥮.

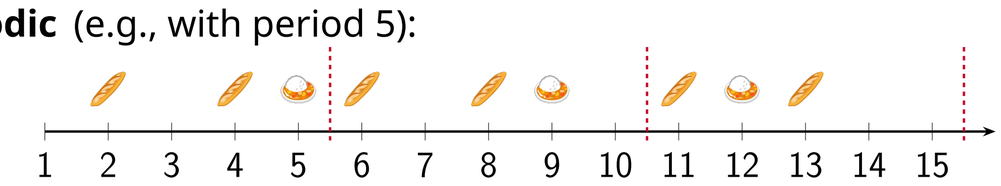
Long-term, periodic, and bounded evaluations

We can evaluate trajectories (for this example, with restaurant visits along them) in various ways.

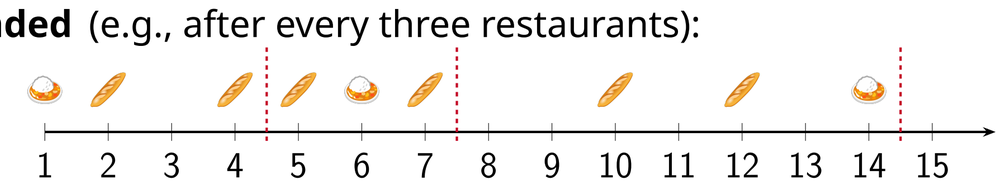
Long-term:



Periodic (e.g., with period 5):



Bounded (e.g., after every three restaurants):



Temporal pluralism scheme

We adapt the fairness schemes from Alamdari et al. (2024).

Given a state space S and action space A, a **temporal pluralism scheme** for n agents is a tuple $\langle \boldsymbol{U}, \boldsymbol{W_{\text{ex}}}, \boldsymbol{B} \rangle$ where

- $\boldsymbol{U} : (S \times A)^* \times S \to \mathbb{R}^n$ is the stakeholder status function.
- $\boldsymbol{W}_{\text{ex}} : (\mathbb{R}^n)^* \to \mathbb{R}$ is the extended aggregation function.
- $\boldsymbol{B} : (S \times A)^* \times S \to \{0, 1\}$ is the filter function.

The status function U returns a vector indicating how well-off each stakeholder is at the endpoint of the given trajectory. A sequence of status vectors can be aggregated by the extended aggregation function W_{ex}.

Temporal pluralism score

Given a trajectory of states and actions $\tau_T = s_1, a_1, s_2, \ldots, a_T, s_{T+1}$, the **temporal pluralism score** of τ_T according to the temporal pluralism scheme $\langle U, W_{\text{ex}}, B \rangle$ is

$$W_{\text{ex}}(U(\tau_{t_1}), U(\tau_{t_2}), \ldots, U(\tau_{t_k}))$$

where τ_t is the prefix of τ_T ending with s_{t+1} and $(t_1, t_2, \ldots, t_k)$ is the subsequence of $(1, 2, \ldots, T)$ for which $B(\tau_{t_i}) = 1$ for each i.

Example scheme $\langle \boldsymbol{U}, \boldsymbol{W_{\text{ex}}}, \boldsymbol{B} \rangle$

Recalling the restaurant example, we could pick the following:

- $\boldsymbol{U}(\tau)$ includes, for each stakeholder, how often they went to a restaurant of their **preferred type** over the trajectory τ.
- $\boldsymbol{W}_{\text{ex}}(u_1, u_2, \ldots, u_k) = \mathit{Nash}(u_{11}, ..., u_{1n}, u_{21}, ..., u_{2n}, \ldots, u_{k1}, ..., u_{kn})$ where Nash is **Nash welfare** (whose value is just the product of its inputs) and u_{ij} is the jth entry of the vector u_i.
- $\boldsymbol{B}(\tau) = 1$ only on those time steps on which another dozen restaurants have been visited.

We compute the Nash welfare as though each **temporal version** of each stakeholder were another individual.

The idea is to give higher scores to trajectories on which not only have a **variety** of restaurants been visited in the long term, but also **during** the process (i.e., within each dozen visits).

Conclusion

- In some cases it may only be possible to achieve pluralistic alignment, reflecting a diversity of preferences or values, **over time**.
- We suggested adapting the approach to temporally extended fairness from **Alamdari et al. (2024)**.
- There was a reinforcement learning algorithm for some fairness schemes, but more **algorithmic** work is needed.
- Further work is also needed to investigate what **specific** temporal pluralism schemes would be most appropriate.

References

Parand A. Alamdari, Toryn Q. Klassen, Elliot Creager, and Sheila A. McIlraith (2024). "Remembering to Be Fair: Non-Markovian Fairness in Sequential Decision Making". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, pp. 906–920.

Martin Lackner (2020). "Perpetual Voting: Fairness in Long-Term Decision Making". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, pp. 2103–2110.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith (2018). "Using Reward Machines for High-Level Task Specification and Decomposition in Reinforcement Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 2107–2116.

{toryn,parand,sheila}@cs.toronto.edu

Computer Science
UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Department of Computer Science, University of Toronto
Vector Institute for Artificial Intelligence
Schwartz Reisman Institute for Technology and Society