Pluralistic Alignment Over Time

Toryn Q. Klassen, Parand A. Alamdari, Sheila A. McIlraith
University of Toronto & Vector Institute
Schwartz Reisman Institute for Technology and Society
Toronto, Ontario, Canada
{toryn,parand,sheila}@cs.toronto.edu

Abstract

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)? In this position paper, we advocate for consideration of temporal aspects including stakeholders' changing levels of satisfaction and their possibly temporally extended preferences. We suggest how a recent approach to evaluating fairness over time could be applied to a new form of pluralistic alignment: temporal pluralism, where the AI system reflects different stakeholders' values at different times.

1 Temporal Aspects of Pluralistic Alignment

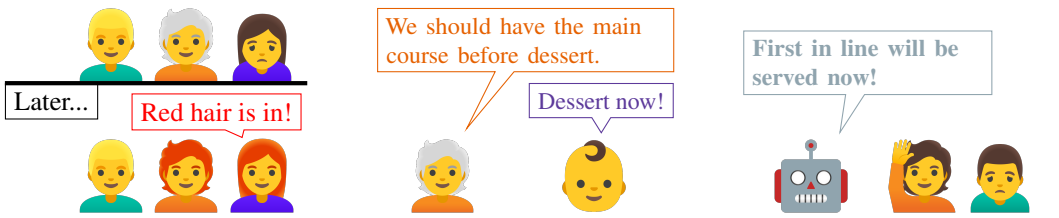
In this paper we consider aligning an AI system's decisions with the values or preferences of multiple stakeholders, a topic that has been recently drawing attention (e.g., Sorensen et al., 2024; Conitzer et al., 2024), in the context of sequential decision making. Sorensen et al. wrote that

> [W]e need systems that are *pluralistic*, or capable of representing a diverse set of human values and perspectives.

Motivated by our context of *sequential* decision making, which takes place over time, we focus on *temporal* aspects of alignment. There are a variety of ways that time may factor into evaluating how aligned an AI system is, including (as illustrated in Figure 1)

1. preference change over time,
2. temporally extended preferences,
3. and how pluralistic alignment may only be realizable over time, through acting to achieve different stakeholder's interests at different times.

Each of these points will be elaborated upon in the following subsections (and we encourage further research on them), but the third will then be the focus of the rest of this paper. We will adapt our recently introduced framework for temporally extended *fairness* (Alamdari et al., 2024) to pluralistic alignment over time.



(a) Group preference change over time

(b) Temporal preferences – i.e., preferences about timelines

(c) Alignment with different stakeholders at different times

Figure 1: Different temporal aspects of pluralistic alignment

1.1 Preference change over time

The preferences of a society or group may change over time. One of the ways of operationalizing pluralistic alignment suggested by Sorensen et al. (2024) (for LLMs) was *Overton pluralism*, where the LLM responds to a query by giving all answers in the Overton window. However, the Overton window is often spoken about as shifting in reality (e.g., Astor, 2019). Furthermore, actions taken by AI systems could potentially influence such shifts, e.g., by changing enough individual preferences (Carroll et al., 2024) or by influencing demographic changes. Less diverse preferences might be easier to satisfy, which could give an AI system an incentive for manipulation.

1.2 Temporally extended preferences, norms, goals, and rewards

People may also have explicitly temporal preferences, i.e., preferences about how the state evolves over time. (There is a body of work in AI planning about satisfying (typically one agent's) temporally extended goals and preferences (e.g., Bacchus and Kabanza, 1998; Son and Pontelli, 2004; Baier et al., 2009; Bienvenu et al., 2011).)

To illustrate, one might prefer that dessert be served after the main course, while someone else might want to eat dessert first (Figure 1b). What order should the meal's courses be served in? For this case, there may also be social norms or conventions to consider. Zhi-Xuan et al. (2024) have suggested that AI systems should be aligned with norms rather than aggregated preferences. Norms may also be temporally extended (e.g., Porfirio et al., 2018; Kasenberg et al., 2018; Malle et al., 2023).

In automated sequential decision making, it's common (for example, when using Markov Decision Processes) to represent preferences or other things to be optimized using a reward function that assigns a numeric value $R(s_t, a_t, s_{t+1})$ to each transition from a state s_t to another state s_{t+1} using action a_t. However, it's also possible to define a *non-Markovian* reward function $R(\tau)$ that maps a whole trajectory of alternating states and actions $\tau = s_1, a_1, s_2, \ldots, a_T, s_{T+1}$ to a real number. This allows for rewarding temporally extended behaviors.

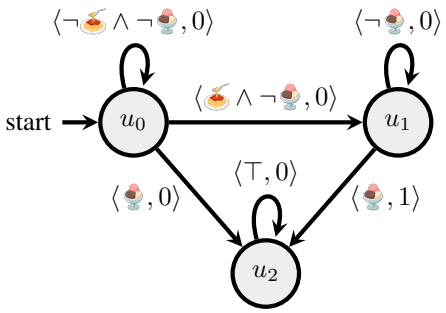


Figure 2: A reward machine that gives reward 1 only on trajectories on which dessert ( ) is eaten after the main course ( ). An edge labelled $\langle \varphi, r \rangle$ is taken when the propositional formula φ is true, yielding reward r.

Non-Markovian reward functions have been describing using temporal logics including LTL (e.g., Bacchus et al., 1996; Thiébaux et al., 2006; Camacho et al., 2017) and reward machines (Toro Icarte et al., 2018, 2022; Camacho et al., 2019). An example reward machine is shown in Figure 2. In the context of alignment, Zhi-Xuan et al. (2024) recently encouraged the consideration of temporal logics and reward machines to better represent human preferences.

Another temporal aspect of preferences is temporal discounting, where future rewards are valued less than current ones. Pitis (2023) considers aggregating the preferences of multiple stakeholders who each have a Markovian reward function, but possibly different discount factors, and argues that in general the aggregation should be a non-Markovian reward function. Finally, we note that the *dynamic reward MDPs* used by Carroll et al. (2024) to model changing preferences also resemble reward machines.

1.3 Pluralistic alignment over time

Pluralistic alignment requires consideration of a collection of human preferences, but such preferences may be conflicting, such that after any one decision, alignment is only with a subset of humans. In a sequential decision making setting, such disparities can be mitigated by future actions. More generally, the decisions made by AI (or human) systems will often cause different people to be better off at different times. How should such temporal tradeoffs be evaluated to determine how good the

decisions were in the short, medium and long term (Alamdari et al., 2024)? To illustrate, consider the problem of distributing goods among people – logistic constraints will typically prevent everyone from getting the same amount of goods at exactly the same time. On a longer timescale, how the utilities of different generations of people (living at different times) should be aggregated is a question that has been considered in economics (e.g., Zame, 2007; Alvarez-Cuadrado and Van Long, 2009).

Indeed, at different times, an AI system may seem to be better aligned with different stakeholders. More positively, while it may not be possible to satisfy everyone with a single decision, a sequence of decisions can reflect a diversity of values. This suggests a new way that a system could be pluralistic: **temporal pluralism**, wherein the system reflects different stakeholders' values at different times.

Which stakeholders' values should be reflected at which times?

The simplest answer might be just that each stakeholder should be satisfied a fair *fraction* of the time. That is closely related to the notion of *distributional pluralism* from Sorensen et al. (2024). Focusing on alignment for language models, Sorensen et al. defined a distributionally pluralistic language model $\mathcal{M}$ as having the property that

> For a given prompt x, $\mathcal{M}$ is as likely to provide response y as the reference population G.

Assume the population provides responses it prefers, and a subgroup of $n\%$ of the population prefers a particular sort of response (and the rest of the population prefers other responses). Then if a distributionally pluralistic model is queried repeatedly, in expectation $n\%$ of its responses would be preferred by the subgroup.

However, some time periods may be viewed as more important than others (because of temporal discounting or other reasons), which suggests a potential need for a more structured approach (e.g., something like turn-taking). Furthermore, in the context of sequential decision making, constraints may be imposed by what state transitions are physically possible, and the AI system may need to plan ahead to be able to satisfy different stakeholders in the future. The matter may be further complicated by stakeholders having temporally extended preferences about the overall timeline, as we discussed in Section 1.2.

2 A framework for pluralistic alignment over time

In this section we adapt the framework for temporally extended fairness from our previous work (Alamdari et al., 2024)[1] to apply to the problem of pluralistic alignment.

Notation When writing a trajectory of states and actions $\tau_T = s_1, a_1, s_2, \ldots, a_T, s_{T+1}$, the subscript T on τ (if present) indicates the number of actions in the trajectory. Given τ_T, we can write τ_i (where $i < T$) for the prefix of τ_T ending with s_{i+1}.

What we want is to be able to evaluate how well a trajectory of states and actions reflects the preferences of a collection of stakeholders. We previously had the notion of a *fairness scheme* (Alamdari et al., 2024, Def. 4.5), which we adopt as a *temporal pluralism scheme*:

Definition 1 (Temporal pluralism scheme)**.** Given a state space S and action space A, a temporal pluralism scheme for n agents is a tuple $\langle U, W_{\text{ex}}, B\rangle$ where

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

U's output, a vector of length n, is meant to measure how "good" the input (a trajectory τ of states and actions) has been for each of n stakeholders. For example, the measure could be, for each stakeholder, the (discounted) sum of that stakeholder's rewards over the trajectory. Alamdari et al. assumed the ith stakeholder has a (Markovian) reward function $R_i(s, a, s')$, but we note that a *non-Markovian* reward function $R_i(\tau)$ is just as compatible with the approach.

[1] Another recent approach to fairness over time was introduced by Torres et al. (2024).

We use the extended aggregation function to compute a *temporal pluralism score* by aggregating the outputs of the stakeholder status function over time. The filter function is used to restrict which time points are considered by the extended aggregation function. We adopt (Alamdari et al., 2024, Def. 4.7) as a *temporal pluralism score*:

Definition 2 (Temporal pluralism score)**.** Given a trajectory $\tau_T = s_1, a_1, s_2, \ldots, a_T, s_{T+1}$, 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 $(t_1, t_2, \ldots, t_k)$ is the subsequence of $(1, 2, \ldots, T)$ for which $B(\tau_{t_i}) = 1$ for each i.

The extended aggregation function can be thought of as a temporally extended social welfare, which looks at how well each stakeholder is doing at each given point in time.

For a simple illustration, consider a scenario (inspired by Lackner (2020)) where an AI assistant books restaurants for the frequent joint dinners of a group of five friends with diverse culinary preferences. We could evaluate the AI using a temporal pluralism scheme with the following components:

- $U(\tau)$ includes, for each stakeholder (member of the friend group), how often they went to a restaurant of their preferred type over the course of the trajectory τ.
- $W_{\text{ex}}(u_1, u_2, \ldots, u_k) = Nash(u_{11}, \ldots, u_{1n}, u_{21}, \ldots, u_{2n}, \ldots, u_{k1}, \ldots, u_{kn})$ where $Nash$ is Nash welfare, a standard social welfare function (whose value is just the product of its inputs) and u_{ij} is the jth entry of the vector u_i. That is, we compute the Nash welfare as though each temporal version of each stakeholder were another individual.[2]
- $B(\tau) = 1$ only on those time steps on which another ten restaurants have been visited.

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 (for each ten restaurants).

As Alamdari et al. noted, temporal pluralism schemes allow for *long-term*, *periodic*, and *anytime* evaluations.

Long-term The score given by a long-term temporal pluralism scheme to a trajectory τ_T only varies with the last stakeholder status $U(\tau_T)$. This ignores the evolution before the end (except insofar as the stakeholder status captures it).

Periodic A periodic temporal pluralism scheme (with period p) ignores or filters out times that are not a multiple of p, so that the temporal pluralism score is equal to $W_{\text{ex}}(U(\tau_p), U(\tau_{2p}), \ldots, U(\tau_{\lfloor T/p \rfloor p}))$. This might be desirable, for example, in a situation where a robot is distributing goods to a group of people, and it takes a while for each round of deliveries to be completed (before the end of a round, it might seem like some people are being ignored, and so that the robot is not aligned with them).

Anytime An anytime temporal pluralism scheme is a periodic scheme with period 1. Depending on the extended aggregation function W_{ex}, achieving a high temporal pluralism score with such a scheme may be difficult or impossible.

An AI system trying to act so as to bring about a trajectory with a high temporal pluralism score would in general need to be able to *remember* some information about the past. Alamdari et al. presented a reinforcement learning approach using an explicit memory, which was able to deal with certain types of fairness schemes, but further work is needed.

3 Conclusion

We have argued for further consideration of temporal aspects of alignment with multiple stakeholders, and suggested that in some cases it may only be possible to achieve pluralistic alignment, reflecting a diversity of preferences or values, *over time*. We further suggested adapting the approach to temporally extended fairness from Alamdari et al. (2024) to the problem of pluralistic alignment. Further work is needed to investigate what specific temporal pluralism schemes would be most appropriate for aggregating people's preferences. Frameworks that people use to organize the satisfaction of their interests over time, like turn-taking and queuing, may be informative in making this choice.

[2]Other options would include having W_{ex} first aggregate across temporal versions of each stakeholder and then aggregate across stakeholders, or vice versa.

Acknowledgements

We wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada CIFAR AI Chairs Program (Vector Institute). The first author also received funding from Open Philanthropy. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute for Artificial Intelligence.

References

Parand A. Alamdari, Toryn Q. Klassen, Elliot Creager, and Sheila A. McIlraith. Remembering to be fair: Non-Markovian fairness in sequential decision making. In *Proceedings of the 41st International Conference on Machine Learning*, pages 906–920. PMLR, 2024. URL https://proceedings.mlr.press/v235/alamdari24a.html.

Francisco Alvarez-Cuadrado and Ngo Van Long. A mixed Bentham–Rawls criterion for intergenerational equity: Theory and implications. *Journal of Environmental Economics and Management*, 58 (2):154–168, 2009. doi:10.1016/j.jeem.2009.04.003.

Maggie Astor. How the politically unthinkable can become mainstream. *The New York Times*, February 2019. URL https://www.nytimes.com/2019/02/26/us/politics/overton-window-democrats.html.

Fahiem Bacchus and Froduald Kabanza. Planning for temporally extended goals. *Annals of Mathematics and Artificial Intelligence*, 22:5–27, 1998. doi:10.1023/A:1018985923441.

Fahiem Bacchus, Craig Boutilier, and Adam J. Grove. Rewarding behaviors. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1160–1167, 1996. URL http://www.aaai.org/Library/AAAI/1996/aaai96-172.php.

Jorge A. Baier, Fahiem Bacchus, and Sheila A. McIlraith. A heuristic search approach to planning with temporally extended preferences. *Artificial Intelligence*, 173(5-6):593–618, 2009. doi:10.1016/J.ARTINT.2008.11.011.

Meghyn Bienvenu, Christian Fritz, and Sheila A. McIlraith. Specifying and computing preferred plans. *Artificial Intelligence*, 175(7-8):1308–1345, 2011. doi:10.1016/J.ARTINT.2010.11.021.

Alberto Camacho, Oscar Chen, Scott Sanner, and Sheila A. McIlraith. Non-Markovian rewards expressed in LTL: guiding search via reward shaping. In *Proceedings of the Tenth International Symposium on Combinatorial Search, SOCS 2017*, pages 159–160. AAAI Press, 2017. doi:10.1609/SOCS.V8I1.18421.

Alberto Camacho, Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. LTL and beyond: Formal languages for reward function specification in reinforcement learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 6065–6073. ijcai.org, 2019. doi:10.24963/ijcai.2019/840.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. AI alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5706–5756. PMLR, 2024. URL https://proceedings.mlr.press/v235/carroll24a.html.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9346–9360. PMLR, 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Daniel Kasenberg, Thomas Arnold, and Matthias Scheutz. Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018*, pages 184–190. ACM, 2018. doi:10.1145/3278721.3278774.

Martin Lackner. Perpetual voting: Fairness in long-term decision making. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 2103–2110. AAAI Press, 2020. doi:10.1609/AAAI.V34I02.5584.

Bertram F. Malle, Eric Rosen, Vivienne Bihe Chi, and Dev Ramesh. What properties of norms can we implement in robots? In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023*, pages 2598–2603. IEEE, 2023. doi:10.1109/RO-MAN57019.2023.10309355.

Silviu Pitis. Consistent aggregation of objectives with diverse time preferences requires non-Markovian rewards. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, pages 2877–2893, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/08342dc6ab69f23167b4123086ad4d38-Paper-Conference.pdf`.

David Porfirio, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. Authoring and verifying human-robot interactions. In *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018*, pages 75–86. ACM, 2018. doi:10.1145/3242587.3242634.

Tran Cao Son and Enrico Pontelli. Planning with preferences using logic programming. In *Logic Programming and Nonmonotonic Reasoning, 7th International Conference, LPNMR 2004*, volume 2923 of *Lecture Notes in Computer Science*, pages 247–260. Springer, 2004. doi:10.1007/978-3-540-24609-1_22.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/sorensen24a.html`.

Sylvie Thiébaux, Charles Gretton, John K. Slaney, David Price, and Froduald Kabanza. Decision-theoretic planning with non-Markovian rewards. *Journal of Artificial Intelligence Research*, 25: 17–74, 2006. doi:10.1613/JAIR.1676.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2107–2116. PMLR, 2018. URL `https://proceedings.mlr.press/v80/icarte18a.html`.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022. doi:10.1613/jair.1.12440.

Manuel R. Torres, Parisa Zehtabi, Michael Cashmore, Daniele Magazzeni, and Manuela Veloso. Temporal fairness in decision making problems. In *27th European Conference on Artificial Intelligence*, pages 1132–1139, 2024. doi:10.3233/FAIA240606.

William R. Zame. Can intergenerational equity be operationalized? *Theoretical Economics*, 2 (2):187–202, 2007. URL `https://econtheory.org/ojs/index.php/te/article/view/20070187/0`.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in AI alignment. *arXiv preprint arXiv:2408.16984*, 2024. doi:10.48550/arXiv.2408.16984.