Remembering to Be Fair: On Non-Markovian Fairness in Sequential Decision Making (Preliminary Report)

Parand A. Alamdari
University of Toronto & Vector Institute
Schwartz Reisman Institute
Toronto, Canada
parand@cs.toronto.edu

Toryn Q. Klassen
University of Toronto & Vector Institute
Schwartz Reisman Institute
Toronto, Canada
toryn@cs.toronto.edu

Elliot Creager
University of Waterloo & Vector Institute
Schwartz Reisman Institute
Waterloo, Canada
creager@cs.toronto.edu

Sheila A. McIlraith
University of Toronto & Vector Institute
Schwartz Reisman Institute
Toronto, Canada
sheila@cs.toronto.edu

Abstract

Fair decision making has largely been studied with respect to a single decision. In this paper we investigate the notion of fairness in the context of sequential decision making where multiple stakeholders can be affected by the outcomes of decisions, and where decision making may be informed by additional constraints and criteria beyond the requirement of fairness. In this setting, we observe that fairness often depends on the history of the sequential decision-making process and not just on the current state. To advance our understanding of this class of fairness problems, we define the notion of non-Markovian fairness in the context of sequential decision making. We identify properties of non-Markovian fairness, including notions of long-term, anytime, periodic, and bounded fairness. We further explore the interplay between non-Markovian fairness and memory, and how this can support construction of fair policies in sequential decision-making settings.

1 Introduction

Sequential decision making involves making decisions over time, typically in service of realizing some near and/or longer-term objectives. As such, the assessment of the fairness of a sequential decision-making process must, in the general case, consider the outcomes of decisions over the entire process, and in the context of other constraints or criteria that inform the objective of the process. In this paper we consider the fairness of a sequential decision-making process with respect to a set of *stakeholders*—a set of entities affected by the outcomes of the decisions that comprise the process.

To ground this discussion, consider the problem of distributing vaccines to countries around the world, and our aspiration that the allocation of vaccines to different countries—our stakeholders—be "fair" in some manner. Without consideration for the distributor, nor for the logistics of the distribution, and with no resource constraints (inventory, financial, or otherwise), we might allocate vaccines instantaneously, and as needed, to all countries around the world. Unfortunately, the world is not without its constraints. Delivering vaccines to countries around the world presents many logistics challenges and constraints that greatly restrict the feasible set of plans. Further, many of

Algorithmic Fairness through the Lens of Time at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

the decisions-–the actions that comprise the plan—are in service of getting the vaccine safely from origin to destination, and may not immediately affect the fairness of the allocation. Does it even make sense to ask whether these intermediate decisions are "fair?" Perhaps what we need to aspire to is **long-term fairness** of the entire decision-making process, that in the limit, when all the vaccines have been delivered, we will be able to proclaim that vaccines were fairly allocated. Unfortunately, this has its problems as well. "In the limit" can be a very long time away. Optimization of logistics may suggest that vaccines be delivered in large batches, one country after another, which may be unfair to those who receive their vaccines later than others. This suggests the need for some form of **anytime fairness**, or perhaps a notion of **periodic** or **bounded fairness** where the fairness of the allocation of vaccines is judged over a bounded time period, assessed at month's end or yearly, or by some other bound such as after every delivery of 1 million vaccines.

This example exposes another important property of fairness in sequential decision making, that it is inherently non-Markovian. That is, the assessment of fairness of a sequential decision-making process does not just rely on the current state, or more generally (s, a, s'), where s' is the state resulting from deciding to perform action a in state s. Rather, it is a function of the history of the decision-making process, the history of state-action pairs (e.g., including all the successful vaccine deliveries), thus necessitating introduction of the notion of **non-Markovian fairness**.

At first blush, this may seem like bad news for machine and reinforcement learning techniques (RNNs and LSTMs notwithstanding) that rely on systems being Markovian. Nevertheless, we observe that this issue of non-Markovian fairness doesn't always seem to arise in practice, and that's because we use **memory**; we often *engineer* our systems to remember what we need to remember to inform (fair) decision making going forward. In some cases we only need to remember a subset of past actions or state, or perhaps only for a window of the past. In other cases, we can even engineer our processes so that the assessment of fairness is Markovian, by augmenting the state of our sequential decision making process with extra variables that capture any necessary computation and book-keeping required to assess fairness using only the current action and state.

Our goal for this preliminary report is modest. We aim to provide formal foundations for studying non-Markovian fairness in sequential decision making. To this end we:

- Define the notion of non-Markovian fairness, long-term fairness, anytime fairness, and periodic or bounded fairness, and what it means for a policy to be fair.
- Provide a number of definitions of domain-independent fairness functions.
- Study the role of memory in non-Markovian fairness.

To simplify our discussion going forward, we will migrate from our vaccine example to a simple example involving the distribution of indivisible goods – in this case doughnuts.

Running Example – Doughnut Allocation: Given n people (stakeholders), and the arrival of m doughnuts for distribution at each time step ($m < n$), define a policy to distribute these doughnuts in a fair manner over time. This problem presents at least two challenges: (i) defining what constitutes a fair allocation of these goods, and (ii) prescribing, or *learning* a policy to realize a fair distribution. We will use variants of this problem to illustrate properties of the formalism that follows.

2 Related Work

We are not the first to study fairness over time. Indeed, a flurry of research activity has followed from the observation that intervening to promote fairness in the short-term can lead to distinct and sometimes unexpected results in the long run (Liu et al., 2018; Hashimoto et al., 2018; Hu and Chen, 2018; D'Amour et al., 2020). Consequently, fair *decision making* and its multi-agent variants have emerged as useful frameworks for modeling long-term fairness considerations. Zhang and Shah (2014) introduced a method to maximize returns for the worst-off agent in a setting where individual agents have local interests. Jabbari et al. (2017) explored a more conservative notion of fairness, where a reinforcement learning agent was tasked with considering long-term discounted rewards when comparing two actions, reminiscent of a dynamic take on the individual fairness principle of "treating likes alike" (Dwork et al., 2012). Group notions of fairness have also been explored and theoretically analyzed within RL (Deng et al., 2022).

Defining what constitutes fairness is an open-ended design choice with normative and political implications (Narayanan, 2018; Binns, 2018; Xinying Chen and Hooker, 2023). Exploring these various normative frameworks has been an exciting recent development for fair decision making: given the importance of the scalar reward signal in RL, one key question is how to aggregate rewards across stakeholders and time to measure the overall fairness of a policy. Recent work has proposed optimizing Nash welfare, Gini social welfare, and utilitarian objectives (Mandal and Gan, 2022; Siddique et al., 2020; Fan et al., 2023).

Our contributions are distinct in our focus on measuring fairness over a history of decisions—as opposed to immediate choices or long-term aggregated discounted rewards—and the special role of memory that this non-Markovian perspective on fairness entails. We argue that non-Markov measures of fairness are quite natural and underexplored formally.

Finally, our motivating sequential doughnut allocation example is related to an established literature on allocation problems (Ibaraki and Katoh, 1988), where fairness concerns have also been explored (Kash et al., 2014; Bouveret and Lemaître, 2016; Caragiannis et al., 2019). Our focus on formal frameworks for sequential decision making differentiates us in this regard; although temporally-extended notions of fairness have also been suggested in the context of computational social choice (Boehmer and Niedermeier, 2021), they are underexplored, even in this context.

3 Preliminaries

In this section we provide basic definitions in service of our characterization of fair sequential decision making. As a starting point, we consider an environment that is fully observable with dynamics that are governed by a Markov Decision Process (MDP) (Puterman, 2014). We utilize an MDP variant that makes explicit the stakeholders that are acting and/or affected by the sequential decision-making process, and their individual reward functions. Our definition is functionally equivalent to that of a multi-objective MDP (Roijers et al., 2013).

Definition 1 (Multi-stakeholder Markov Decision Process)**.** A multi-stakeholder MDP is a tuple $\langle S, s_{\text{init}}, A, P, R_1, \dots, R_n, \gamma \rangle$ where S is a finite set of states, $s_{\text{init}} \in S$ is the initial state, A is a finite set of actions, and $P(s_{t+1} \mid s_t, a_t)$ is the transition probability distribution, giving the probability of transitioning to state s_{t+1} by taking action a_t in s_t. For $i = 1, \dots, n$, $R_i : S \times A \times S \to \mathbb{R}$ is the reward function of the ith stakeholder, and $\gamma \in (0, 1]$ is the discount factor.

In such an environment, executing action a_t in state s_t results in state s_{t+1} according to transition probability distribution $P(s_{t+1}|s_t, a_t)$, with each stakeholder i receiving a reward $R_i(s_t, a_t, s_{t+1})$. (The formalism is agnostic with respect to who—stakeholders or others—actually executes the actions.)

A *policy*, $\pi(a|s)$, is a probability distribution over the actions $a \in A$, given a state $s \in S$. Each policy induces a probability distribution over the future states and rewards that will be encountered if actions are selected according to the policy. This allows us to rank policies by their expected cumulative reward, and in the case of multiple stakeholders, to distinguish between the expected cumulative reward for an individual stakeholder, i, or perhaps some expected cumulative aggregation of the rewards of individual stakeholders. Such policies can be defined (e.g., by domain experts) or learned.

To facilitate exposition, we introduce the notion of a trace, corresponding to an execution of a (multi-stakeholder) MDP.

Definition 2 ((Bounded) trace)**.** A trace τ of an MDP is a sequence of state-action pairs corresponding to a potential execution of the MDP starting in state $s_1 = s_{\text{init}}$. i.e.,

$$\tau = s_1, a_1, s_2, a_2, s_3, a_3, \dots$$

A bounded trace, $\tau_{x,y}$ is a finite trace that begins at time step x and ends at y, with the state resulting from the last action, s_{y+1}, appended to the finite sequence. We commonly use finite traces starting in s_{init}, written as $\tau_{1,y}$.

Observe that stakeholder rewards can be directly computed from traces.

Following in the spirit of (Sutton and Barto, 2018), given a trace τ (resp. bounded trace, $\tau_{1,y}$) the *discounted return* for stakeholder i is the discounted sum of that stakeholder's rewards accumulated over τ (resp. $\tau_{1,y}$):

$$G_i(\tau) \stackrel{\text{def}}{=} \sum_{t=1}^{\infty} \gamma^{t-1} R_i(s_t, a_t, s_{t+1}) \quad \text{and} \quad G_i(\tau_{1,y}) = \sum_{t=1}^{y} \gamma^{t-1} R_i(s_t, a_t, s_{t+1})$$

The return of a policy may vary, depending on the stochasticity of the environment. As such we evaluate policies in terms of their expected return.

4 Non-Markovian Fairness

In Section 1, we argued that to assess the fairness of a sequential decision-making process in the general case we needed to examine the entire history of the process. In this section, we elaborate on that claim. We begin by first defining a decision process in which fairness, like reward in an MDP or multi-stakeholder MDP, is Markovian.

Definition 3 (Fair Multi-Stakeholder MDP)**.** Similar to a multi-stakeholder MDP, a fair multi-stakeholder MDP is a tuple $\langle S, s_{\text{init}}, A, P, R_1, \ldots, R_n, \gamma, F \rangle$ where S is the finite set of states, s_{init} is the initial state, A is the finite set of actions, P is the transition function, R_i is the reward function for agent i, and γ is the discount factor. $F : S \times A \times S \to \mathbb{R}$ is a *fairness function* which, like the reward function, only considers the current transition. As a result, $f_t = F(s_t, a_t, s_{t+1})$ returns a number indicating the *fairness* of action a_t, which was taken in state s_t and resulted in state s_{t+1}.

The intended use of the fairness function is to formalize a notion of fairness. For example, one could define a binary-valued fairness function

$$F(s_t, a_t, s_{t+1}) = \begin{cases} 1 & \text{if taking action } a_t \text{ in state } s_t \text{ was ``fair''} \\ 0 & \text{otherwise} \end{cases}$$

Alternatively, one could allow the function to range over $\mathbb{R}$ to capture a graded notion of fairness. Nevertheless, because the fairness function in a fair multi-stakeholder MDP is Markovian—it depends only on the last experience $\langle s_t, a_t, s_{t+1} \rangle$—its expressiveness is limited. For example, when deciding who should receive a doughnut, it may seem intuitively fair to choose the stakeholder who has received the fewest doughnuts. However, this would require keeping track of how many doughnuts each stakeholder had received (or at least, the differences between the stakeholder allocations), and this is not native to the state, and thus not captured by a Markovian fairness function, $F(s_t, a_t, s_{t+1})$.

This issue with expressiveness similarly affects *rewards*, which are typically Markovian. While not the focus of this paper, we note that non-Markovian rewards are becoming increasingly popular in reinforcement learning to model reward-worthy behavior over time (e.g. Toro Icarte et al., 2018a,b; Camacho et al., 2019; Gaon and Brafman, 2020; Toro Icarte et al., 2022).

Following Bacchus et al. (1996), a Non-Markovian Decision Process (NMDP) is like an MDP but the reward function is non-Markovian, i.e., it is defined over the entire state-action history, $R : (S \times A)^+ \times S \to \mathbb{R}$. As such, the reward that the agent receives from performing action a_t in state s_t, resulting in state s_{t+1}, depends on the entire history, $R(s_1, a_1, \ldots, s_t, a_t, s_{t+1})$. Commensurately, the policy for an NMDP is also non-Markovian.

Definition 4 (Non-Markovian Policy)**.** A non-Markovian policy is a mapping from histories to actions (or distributions over actions if the policy is non-deterministic): $\pi(a_t | s_1, a_1, \ldots, a_{t-1}, s_t)$.

Taking inspiration from NMDPs, we can define a process in which fairness, instead of rewards, depends on the history:

Definition 5 (Non-Markovian Fair Decision Process (NMFDP))**.** An NMFDP, like a fair multi-stakeholder MDP, is a tuple $\langle S, s_{\text{init}}, A, P, R_1, \ldots, R_n, \gamma, F \rangle$ where the reward functions remain Markovian, taking the form $R_i(s_t, a_t, s_{t+1})$, but where the fairness function F is a *non-Markovian fairness function*,

$$F : (S \times A)^+ \times S \to \mathbb{R},$$

and the value of F after execution of the first t actions, $f_t = F(\tau_{1,t})$, corresponds to a measure of the fairness of the process over that time period.

Theorem 1. There exists a fairness objective that can be modeled with a NMFDP but not with a fair multi-stakeholder MDP.

Proof. We build an example to show that this fairness objective exists. Consider the doughnut allocation problem in Section 1, which we model as a NMFDP $\langle S, s_{\text{init}}, A, P, R_X, R_Y, \gamma, F\rangle$ where there are only two doughnut recipients (stakeholders) X and Y, and we have that $S = \{s_{\text{init}}\}$ and $A = \{toX, toY\}$ (actions that allocate one doughnut to X (correspondingly to Y)). We define the NMFDP fairness function as follows:

$$F(\tau_{1,T}) = \left| \left(\sum_{t=1}^{T} \mathbb{I}(a_t = toX) \right) - \left(\sum_{t=1}^{T} \mathbb{I}(a_t = toY) \right) \right|$$

That is, the output at time T is the absolute value of the difference in the total number of doughnuts that have been allocated to stakeholders X and Y over that time period. It's easy to see that for any natural number N, there is a time T and trace $\tau_{1,T}$ such that $F(\tau_{1,T}) = N$. On the other hand, in any fair multi-stakeholder MDP, while we can pick a different set of states or actions, those will still be finite and so the fairness function, of the form $F(s_t, a_t, s_{t+1})$, can have only finitely many possible output values. □

We note that there is also work on the limited expressiveness of Markov *rewards* (e.g., Abel et al., 2021), which is also relevant as the fairness functions we consider take the same forms as reward functions (Markovian functions in fair multi-stakeholder MDPs and non-Markovian functions in NMFDPs).

While the NMFDP is expressive, it is also abstract: two key context-specific design choices must be made before we can make meaningful claims about the "fairness" of an NMFDP. The first choice is how to realize the fairness function F. Noting that F is evaluated at every transition in the trace τ, the second choice is to how to aggregate the sequence of non-Markovian fairness signals $\{f_t\}$, which tells us to what extent a policy π can be considered fair overall. We spend the remainder of this section (non-exhaustively) discussing several possible realizations for each of these design choices.

4.1 Different Choices of Fairness Function

Nash welfare. As Caragiannis et al. (2019) explain, the idea of maximizing Nash welfare, the *product* of agents' utilities, has the property that "informally, it hits a sweet spot between Bentham's utilitarian notion of social welfare—maximize the sum of utilities—and the egalitarian notion of Rawls—maximize the minimum utility." We could make the fairness function Nash welfare (taking an agent's utility to be its return):

$$F(\tau_{1,T}) = \prod_{i=1}^{n} G_i(\tau_{1,T}) \tag{1}$$

Rawlsian social welfare Simply stated, the Rawlsian objective is to improve the outlook of the most disadvantaged. In our sequential multi-agent setup, this comprises maximizing the utility of the worst-off agent, resulting in the following fairness function:

$$F(\tau_{1,T}) = \min\{G_i(\tau_{1,T}) : 1 \le i \le n\}. \tag{2}$$

Time in first place This is a more unusual function designed to illustrate how temporally extended properties could be taken into account. A trace could score well in terms of Nash welfare or Rawlsian social welfare while always having one agent slightly behind the others in accumulated rewards. That agent might think that it would be more *fair* if they themselves could too have some moments in which they are ahead. Let's define $\text{first}_t \subseteq \{1, \ldots, n\}$ as the set of agents at time t who are in "first place" – the agents whose return so far is the greatest (the set will only contain more than one agent if two agents have identical return). The following fairness function is the fraction of time that the worst-off agent is in first place:

$$F(\tau_{1,T}) = \min_i \left(\frac{\sum_{t=1}^{T} \mathbb{I}\left(i \in \text{first}_t\right)}{T} \right) \tag{3}$$

To our knowledge, this notion of fairness has not been explored previously. Furthermore, it highlights the need to evaluate fairness at the intermediate points throughout the trace, rather than looking at single-step or eventual outcomes.

4.2 Different Notions of a Fair Policy

Given an NMFDP, what sort of policy should one try to find? There are a number of options. Let's first note that in general we might want a non-Markovian policy, and henceforth when we refer to a "policy" π we mean a non-Markovian policy (Definition 4).

One natural idea of a "good" policy is one that *works towards* being fair in the **long term**. The initial state of affairs might be very unfair, but a policy could aim to make things eventually fair, which leads to the following definition.

Definition 6 (Fair in the Limit Policy)**.** Let $\mathcal{M}$ be an NMFDP, and π is a policy defined on $\mathcal{M}$. We assume that $F : (S \times A)^+ \times S \to [0, 1]$. π is fair in the limit, if and only if:

$$\lim_{t\to\infty, \tau_{1,t}\sim\mathcal{M},\pi} F(\tau_{1,t}) = 1$$

(If the environment is sufficiently stochastic, that may not be a realistic notion.) A guarantee of eventual fairness might only be satisfactory to those with a lot of patience. The next definition considers (approximate) fairness within a time interval.

Definition 7 (Any-time ϵ-Fair Policy)**.** Let $\mathcal{M}$ be an NMFDP where $F : (S \times A)^+ \times S \to [0, 1]$, π be a policy defined on $\mathcal{M}$, and $0 \leq \epsilon < 1$ be an arbitrary real number. π is any-time ϵ-fair on an interval $I \subseteq \mathbb{N}$ if and only if:

$$\forall t \in I, \tau_{1,t} \sim \mathcal{M}, \pi : \ F(\tau_{1,t}) \geq 1 - \epsilon$$

We note that there is a straight-forward connection between the notions of fair-in-the-limit and any-time ϵ-fairness.

Theorem 2. Let $\mathcal{M}$ be an NMFDP and π a policy that is fair in the limit. Then $\exists t_1 \in \mathbb{N}$ such that π is an any-time ϵ-fair policy in $[t_1, \infty)$.

Proof. Following Definition 6 (and the mathematical definition of a limit), for any $\epsilon > 0$ the following holds: $\exists t_1 > 0 : \forall t \geq t_1 \ F(\tau_{1,t}) > 1 - \epsilon$. As a result, π is any-time ϵ-fair in $[t_1, \infty)$. □

In some scenarios, the fairness of the process may need to be assessed at regular intervals, such as weekly or monthly, or at the end of a fiscal year. In such cases, rather than anytime fairness or long-term fairness, a guarantee of fairness within a fixed time period may be necessary. For example, a research group may mandate that every student in the group gets the opportunity to present their research at least once during any 3-month period. To address such fairness criteria, we define the notion of a *periodically fair policy*:

Definition 8. [Periodically Fair Policy (with period k)] Let $\mathcal{M}$ be an NMFDP, and π is a policy defined on $\mathcal{M}$. We assume that $F : (S \times A)^+ \times S \to [0, 1]$. π is a periodically fair policy with period k, if and only if:

$$\forall \tau \sim \mathcal{M}, \pi \text{ and } \forall t_1 \in \mathbb{N} : \exists t_2 \in \{t_1, t_1 + 1, \ldots, t_1 + k - 1\} : F(\tau_{1,t_2}) = 1$$

where τ_{1,t_2} is a prefix of τ.

As defined, periodic fairness only requires that the trace be fair at *some* point within $k - 1$ steps of any given point. We can additionally define that π is **exactly** periodically fair with period k if $F(\tau_{1,n\cdot k}) = 1$ for all $n \in \{1, 2, 3, \ldots\}$ Further, as discussed in Section 1, still other applications warrant a notion of *bounded fairness*, analogous to the above, but where the period is dictated by some property of the state, rather than by a time period. In our vaccination example, we suggested assessing fairness after the delivery of 1 million vaccines. In the doughnut example, we might assess fairness after the allocation of the 24 doughnuts in the box.

The above definitions relating to different notions of fair policies treat fairness as an absolute notion. Alternatively, we can consider fairness as a relative notion—some policies are fairer than others—using various criteria. In this context, fairness becomes something to optimize. The following definition treats fairness as analogous to reward, and balances optimizing both.

Definition 9 (Fair-optimal Policy)**.** Let $\mathcal{M}$ be an NMFDP. The goal is to find a policy that maximizes the following objective:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\alpha_1\left(\sum_{i=1}^{n} G_i(\tau)\right) + \alpha_2\left(\sum_{t=1}^{\infty}\gamma^{t-1}F(\tau_{1,t})\right)\middle| s_1 = s_{\text{init}}, \tau \sim \pi\right] \tag{4}$$

where α_1 and α_2 are hyper-parameters to balance the tradeoff between fairness and sum of rewards.

We were inspired by the *utilitarian* perspective and aggregate the rewards of the stakeholders with a linear function. In general, the aggregation can be a weighted linear combination of rewards. Weights of stakeholders can be chosen based on a preference over the stakeholders, e.g., in the example in Section 1, stakeholder countries with severe outbreaks might be weighted more heavily, or the stakeholder corresponding to the agent responsible for vaccine distribution could be weighted more heavily.

5 The Role of Memory

Having argued that fairness can necessitate reasoning over the history of past states and actions (Definition 5), in this section we ask whether we need to reason over the *whole* history. We find memory—an external scratchpad that augments the current state—can, in some cases, allow us to convert these non-Markovian fairness functions into corresponding Markovian functions, thereby facilitating policy learning. In particular, by making the function Markovian, we can adapt reinforcement learning strategies designed for Markovian reward functions to learn policies that optimize for fairness.

As intuition, we observe that in some cases, to determine the fairness of a process we only need to remember a subset of past actions or states. Indeed, when building an MDP, it is not uncommon to utilize domain knowledge to augment observed state with extra state variables that encode *what* to remember from the past. Returning to the doughnut example from Section 1, if our objective is to distribute doughnuts as they arrive and to do so as evenly as possible, we might create a queue, define a policy to distribute to the front of the queue, and then just require remembering people's order in the queue. Alternatively, we might aggregate and remember how long each individual has waited for their doughnuts and develop a policy to minimize the difference in aggregate wait times between individuals.

Further, if we are defining fairness in terms of Nash welfare as described in Section 4.1, we only need to remember the return of each stakeholder so far which is a discounted sum of rewards. To model this sort of thing, we can modify the definition of the fairness function in the NMFDP to the following:

$$F : (S \times M) \rightarrow \mathbb{R}$$

where M is the set of possible memory states. We note that memory states could be finite or infinite. The value of the memory at time $t+1$ would be $m_{t+1} = g(m_t, a_t, s_{t+1})$ for some computable function g, where m_t is the value of the memory at time t (we will call g the *memory update function*). The memory would always start in some initial configuration $m_{\text{init}} \in M$ and at each time point, we could compute F using the memory to that point, $f_t = F(s_{t+1}, m_{t+1})$. I.e., it is a function of the state and memory resulting from executing a_t in s_t.

5.1 Expressiveness

It's important to understand *when and how* we can augment an NMFDP with memory in order to realize an equivalent Markovian fairness function. In what follows we provide several insights on this topic.

Corollary 1 (to Theorem 1)**.** In Theorem 1, we constructed an NMFDP and calculated F using the bounded trace so far. As the difference in the number of received doughnuts by X and Y can be arbitrarily large (equal to t after t actions), we need an unbounded memory in order to calculate F at each time step t. So, there exists an NMFDP with a fairness objective that needs unbounded memory.

We can also specify several fairness functions that can be expressed using a limited amount of memory.

Theorem 3. Let $\langle S, s_{\text{init}}, A, P, R_1, \ldots, R_n, \gamma, F\rangle$ be an NMFDP and suppose $F : (S \times A)^+ \times S \rightarrow \{0, 1\}$ is *regular* in the following sense: the set of strings in $(S \cup A)^*$ that F maps to 1 is a regular language (i.e., the set of strings can be described by a regular expression). Then computing $F(\tau_{1,t})$ at every time step requires only a bounded amount of memory.

Proof. Since F is regular, there exists a deterministic finite automaton (DFA) (see, e.g., Sipser, 1997, Chapter 1) $A_F = \langle Q, \Sigma, \delta, q_0, Acc\rangle$ where Q is the finite set of automaton states, $\Sigma = S \cup A$ is the set of input symbols (the set of states and actions of the NMFDP), $\delta : Q \times \Sigma \rightarrow Q$ is the transition function, q_0 is the automaton's initial state, and $Acc \subseteq Q$ is the automaton's set of accepting states, such that

$$A_F \text{ accepts a bounded trace } \tau_{1,t} = s_1, a_1, s_2, a_2, \ldots, s_{t+1} \text{ just in case } F(\tau_{1,t}) = 1$$

Only a bounded amount of memory is required to store the automaton A_F and compute its transitions. □

For an example of a *regular* fairness function (as described in Theorem 3), suppose we have an NMFDP $\langle S, s_{\text{init}}, A, P, R_1, \ldots, R_n, \gamma, F\rangle$ where the set of actions A can be partitioned into $A_1 \cup \cdots \cup A_n$ so that the actions in A_i (for each i) are associated with the ith stakeholder (e.g., because the actions in A_i allocate goods to the ith stakeholder, or because the actions in A_i are performed by the ith stakeholder). One simple notion of fairness is *turn-taking*. If we require that the stakeholders take turns in numerical order and mark a trace as fair if the turn order has been followed so far, that corresponds to the following regular expression:

$$\underbrace{(SA_1SA_2\ldots SA_n)^*}_{\text{0 or more completed rounds}}\underbrace{((SA_1)|(SA_1SA_2)|\ldots|(SA_1SA_2\ldots SA_n))}_{\text{the last (possibly incomplete) round}}S$$

We now establish the complementary relationship.

Theorem 4. Let $\mathcal{M}$ be an NMFDP and suppose $F : S \times M \rightarrow \{0, 1\}$ is a binary fairness function where g is the memory update function. If M is a finite set, then F is regular (i.e., could be described with a regular expression).

Proof. We can construct a DFA $A_F = \langle Q, \Sigma, \delta, q_0, Acc\rangle$, where $Q = S \times M$ is the finite set of automaton states. $q_0 = \langle s_{\text{init}}, m_{\text{init}}\rangle$ which means that both $\mathcal{M}$ and M are in their initial states. $\Sigma = A \times S$ is the set of input symbols and $\delta : Q \times \Sigma \rightarrow Q$ is defined by $\delta(\langle s, m\rangle, \langle a, s'\rangle) = \langle s', g(m, a, s')\rangle$. We define Acc so that for all $s \in S, m \in M$, if $F(s, m) = 1$ then $\langle s, m\rangle \in Acc$, and Acc only contains these states. It can be shown by induction that starting from q_0, A_F accepts the sequence $\langle a_1, s_2\rangle, \langle a_2, s_3\rangle \ldots, \langle a_t, s_{t+1}\rangle$ if and only if $F(s_{t+1}, m_{t+1}) = 1$. As F can be represented as a DFA, F is regular. □

Theorem 5. Let $\mathcal{M} = \langle S, s_{\text{init}}, A, P, R_1, \ldots, R_n, \gamma, F\rangle$ be an NMFDP, and $F : (S \times M) \rightarrow \{0, 1\}$ is a binary fairness function with a bounded amount of memory M. Then, we can define a fair multi-stakeholder MDP such as $\mathcal{M}' = \langle S', s'_{\text{init}}, A, P', R_1, \ldots, R_n, \gamma, F'\rangle$ where $\mathcal{M}'$ is a Markovian equivalent of $\mathcal{M}$.

Proof. We can construct $\mathcal{M}'$ as a fair multi-stakeholder MDP by augmenting the state space with the memory. Let $S' = S \times M$ be a new state space, and $s'_{\text{init}} = \langle s_{\text{init}}, m_{\text{init}}\rangle$ where m_{init} is the initial state of memory M at s_{init}. We refer to the first element of s' as s, and the second as m. The actions and reward functions are the same, as we can use $s_t \in s'_t$ as the input. $P'(s'_{t+1}|s'_t, a_t) = P(s_{t+1}|s_t, a_t)$ if $m_{t+1} = g(m_t, a_t, s_{t+1})$ and otherwise 0, and $F'(s'_t, a_t, s'_{t+1}) = F(s_{t+1}, m_{t+1})$. Then it can be seen that for an arbitrary trace $\tau_{1,t} \sim \mathcal{M}$, there is a corresponding trace $\tau'_{1,t} \sim \mathcal{M}'$, and their respective fairness functions' values agree at each time step. □

We conclude this section by observing that by augmenting the state space with appropriate memory, it allows us to convert non-Markovian fairness functions into Markovian functions, thereby, in some cases, facilitating the learning of a policy with off-the-shelf learning algorithms.

6 Concluding Remarks

In this paper, we have endeavored to advance some insights, formal foundations, and theoretical properties relating to the study of fairness in sequential decision making. Central to our exposition is the observation that the fairness of sequential decision making is, in the general case, non-Markovian—that it requires consideration of the history of the decision-making process. We have also observed that under certain conditions it is possible (and in other cases, impossible) to construct—through the exploitation of memory—an equivalent *Markovian* decision process whose policies mirror those of the original problem.

So what? This is consequential with respect to at least three practical tasks relating to fairness in sequential decision making:

1. **Verification of an instance of an executed process:** assessment of whether a particular trace—a history of actions and events that occurred in a particular instance—is deemed to be fair;
2. **Verification of policy:** assessment of whether a policy and additionally a fairness function enforces certain properties that are desirable—fairness or otherwise;
3. **Policy learning:** learning (optimal) fair policies from data, where, as observed in Section 5, by augmenting our NMFDP state space with appropriate memory it allows us to convert non-Markovian fairness functions into Markovian functions, thereby facilitating the learning of a policy with off-the-shelf learning algorithms.

In future work, we will utilize the foundations introduced in this paper to address some of the tasks identified above. This includes developing techniques to learn fair policies in the context of sequential decision making, and to analyze their computational and theoretical properties.

Acknowledgements

We wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and Microsoft Research. The second 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

David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael L. Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 7799–7812, 2021.

Fahiem Bacchus, Craig Boutilier, and Adam J. Grove. Rewarding behaviors. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1160–1167, 1996.

Reuben Binns. What can political philosophy teach us about algorithmic fairness? *IEEE Security and Privacy*, 16(3):73–80, 2018.

Niclas Boehmer and Rolf Niedermeier. Broadening the research agenda for computational social choice: Multiple preference profiles and multiple solutions. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 1–5, 2021.

Sylvain Bouveret and Michel Lemaître. Characterizing conflicts in fair division of indivisible goods using a scale of criteria. *Autonomous Agents and Multi-Agent Systems*, 30(2):259–290, 2016.

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, 2019.

Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3):12:1–12:32, 2019.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

Zhun Deng, He Sun, Zhiwei Steven Wu, Linjun Zhang, and David C Parkes. Reinforcement learning with stepwise fairness constraints. *arXiv preprint arXiv:2211.03994*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023*, pages 1991–1999, 2023.

Maor Gaon and Ronen I. Brafman. Reinforcement learning with non-Markovian rewards. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 3980–3987. AAAI Press, 2020.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

Toshihide Ibaraki and Naoki Katoh. *Resource allocation problems: algorithmic approaches*. MIT press, 1988.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pages 1617–1626. PMLR, 2017.

Ian Kash, Ariel D Procaccia, and Nisarg Shah. No agent left behind: Dynamic fair division of multiple resources. *Journal of Artificial Intelligence Research*, 51:579–603, 2014.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. *CoRR*, abs/2208.12584, 2022.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conference on Fairness, Accountability and Transparency*, 2018.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8905–8915. PMLR, 2020.

Michael Sipser. *Introduction to the Theory of Computation*. PWS Publishing Company, 1997.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Teaching multiple tasks to an RL agent using LTL. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018a. 452–461.

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 (ICML)*, pages 2112–2121, 2018b.

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

Violet Xinying Chen and JN Hooker. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, pages 1–39, 2023.

Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems 27*, 2014.