Multi-agent Dynamic Resource Allocation: A Reinforcement Learning Approach

Stefania Costantini[1], Giovanni De Gasperis[1], Pasquale De Meo[2], Francesco Gullo[1] and Alessandro Provetti[3]

[1]University of L'Aquila, Italy
[2]University of Messina, Italy
[3]Birkbeck, University of London, UK

Abstract

We investigate the problem of cooperative resource allocation in multi-agent systems, focusing on dynamic scenarios such as hospital networks. In our model, agents (e.g., hospitals) aim to redistribute limited resources, such as medical personnel, in a way that satisfies both local constraints and global equity objectives. We devise a reinforcement learning approach to a dynamic scenario with time-varying resource needs.

We empirically evaluate the proposed approach through extensive experiments. Our results demonstrate the effectiveness of our approach.

1. Introduction

The fair and efficient allocation of resources in decentralised environments has long been a fundamental challenge in Artificial Intelligence (AI) and Economics [1]. In domains where autonomous agents pursue common goals, such as healthcare networks (e.g., the British National Health Service (NHS)) [2], wireless networks [3], or cloud computing networks [4], the ability to enable cooperation without centralised authority is of both theoretical and practical importance.

In this work, we consider a network of agents managing local resources to achieve individual objectives while cooperating toward a collective goal through resource exchanges (e.g., lending). This paradigm aligns with the concept of *fairness*, which ensures equitable outcomes while maintaining system functionality. Our motivating example is an idealized model of the NHS healthcare network: hospitals manage their physician rosters but may temporarily lend doctors to others during localized emergencies (e.g., outbreaks of transmissible diseases). This scenario shares similarities with the interbanking scenario [5, 6], though healthcare systems differ in their universal objective of avoiding hospital failures, unlike banking systems where central banks underwrite systemic stability.

A number of approaches have been proposed so far to allocating resources in multi-agent systems (MARA), including Distributed Constraint Optimization [7], Social Choice Theory [8], and Market-Based Coordination [9]. Among these, *Nash Welfare Optimization (NWO)* stands out as a principled method. NWO models societal welfare as the geometric mean of individual utilities. The NWO formulation balances efficiency and fairness by prioritizing improvements for agents with lower allocations and it comes with no surprise that NWO has been widely studied for its theoretical guarantees and practical effectiveness [10, 11].

However, we identify three critical limitations in its application to healthcare networks, namely:

(a) Minimum vs. Target Staffing: while NWO ensures hospitals meet *baseline staffing* thresholds (that is, we assume that each hospital has at least m_i doctors who ensure its functioning), it ignores *aspirational targets* (that is, we assume that each hospital wants at least t_i doctors) needed for optimal service delivery. In some cases, it is appropriate to concentrate more resources in highly specialised medical centres (such as hospitals specialising in rare diseases or hospitals involved in innovative

International Joint Workshop of Artificial Intelligence for Healthcare (HC@AIxIA) and HYbrid Models for Coupling Deductive and Inductive ReAsoning (HYDRA): HC@AIxIA+HYDRA 2025

stefania.costantini@univaq.it (S. Costantini); giovanni.degasperis@univaq.it (G. D. Gasperis); pdemeo@unime.it (P. D. Meo); francesco.gullo@univaq.it (F. Gullo); a.provetti@bbk.ac.uk (A. Provetti)

clinical trials), even if this may lead to slight inequalities in the distribution of staff. The NWO method cannot meet this requirement.

(b) Constraint Handling: NWO lacks explicit mechanisms to enforce global constraints, such as the fact that the total number of doctors must be constant or hard lower bounds on m_i thresholds. Specifically, the constraint on the overall size of the labour force is a direct consequence of public expenditure management policies, which in some countries impose a freeze on the recruitment of new staff [12].

(c) Dynamic Adaptability: in emergencies (e.g., pandemics), staffing needs fluctuate rapidly. NWO (as well as any other method based on an optimisation algorithm) requires replanning the allocation of doctors from scratch after each change, but this is a computationally prohibitive task for large hospital networks, and it is useless when updates in staffing requirements vary rapidly over time.

In our previous work [13], we addressed the above limitations *(a)-(c)* focusing on a *static scenario* only, i.e., assuming that the staffing needs of each hospital are fixed over time. In this paper, we extend that work by tackling the same problem in the novel *dynamic scenario* in which the demand for staff vary over time. Particularly, we consider two specific settings: *(a) Shifting Targets*, i.e., the staffing needs of a single hospital undergo a small variation (an increase or decrease of two units) at each point in time, and *(b) Shocking Targets*, i.e., we assume that a randomly selected hospital loses half of its staff due to unexpected and exceptional events, but it still satisfies the constraint of minimum operability. We introduce a *Proximal Policy Optimisation* (PPO)-based [14] reinforcement learning (RL) agent. Our agent learns to redistribute staff incrementally, avoiding full re-optimisation. In addition, PPO uses *clipped objective functions* to avoid destructive policy updates. A further novelty of our approach is that it attempts to satisfy constraints by appropriately reshaping its reward function. Experiments reveal that our RL approach responds well to changes in staffing requirements without having to recalculate the optimal solution from scratch, thus it excels in dynamic scenarios such as pandemics or seasonal shifts in demand.

While our approach is grounded in healthcare, its principles generalize to other domains requiring cooperative resource redistribution.

The remainder of the paper is organized as follows: Section 2 presents background on our resource allocation framework in a static settings, originally introduced in our previous work [13]. Section 3 discusses the main contribution of this paper, i.e., a dynamic scenario for staff allocation. Section 4 outlines the experimental setup and benchmarks, followed by an in-depth analysis of the results. Section 5 discusses related literature in multi-agent resource allocation. Finally, in Section 6 we draw our conclusions and highlight directions for future research.

2. Preliminaries

Let us begin with an idealised multi-agent resource allocation framework that models an NHS-style healthcare domain. We are given a set of n hospitals, denoted as $\mathcal{H} = \{\vec{h}_1, \vec{h}_2, \dots \vec{h}_n\}$ where each hospital $\vec{h}$ has three dimensions: the current number of doctors available, c_i, the target rooster, t_i and the minimum number of doctors needed for the hospital to operate, m_i. In a static scenario, the total number of doctors available in $\mathcal{H}$ is fixed, i.e., $\sum_{i=1}^{n} c_i = C$.

Of course, if some hospital $\vec{h}_i$ has critically-low levels of staff, i.e., $c_i \leq m_i$, the best option is to recruit more doctors. Yet, this may not be possible, even for relatively-long interim periods. Hence, we consider transferring doctors from other hospitals to improve the overall efficiency of the hospital system $\mathcal{H}$. The idea is that 'wealthy' hospitals (those operating at or near full rooster) could lend doctors to $\vec{h}_i$. This practice is common in NHS-style health systems, e.g., in summer when population density in tourism areas shows huge alterations. Similarly, emergency situations, such as the outbreak of epidemics and their containment in the geographical areas where they have occurred, may require the emergency re-assignment of medical staff to cope with the spike in hospital admissions.

The goal now is to model the emergency re-distribution scenario so as to compute solutions that are optimal or near-optimal with respect to the global objective of maintaining all hospitals viable

and operating, while fulfilling all the constraints on the capacity of individual hospitals. In [13], we considered a *static scenario*, i.e., we assume that the targets for each hospital are constant over time. We handled such a static scenario by devising three optimisation strategies, based on quadratic programming [15], Progressive Taxation [16], and Nash Welfare Optimisation [17, 18], respectively, along with a hybrid approach that combines the latter two. The focus of this work is instead on a *dynamic setting*, i.e., when sudden, time-varying changes in hospital demand arise. In the following section, we describe such a dynamic scenario and present the proposed reinforcement learning approach to handle it.

3. Multi-agent Resource Allocation: the Dynamic Scenario

In real scenarios, the size of a hospital workforce quickly (and often unexpectedly) fluctuates over time. For example, during pandemics or natural disasters, the number of doctors on duty in a hospital is usually much higher than what is generally considered satisfactory. In *dynamic* scenarios where staffing requirements can vary over time, the approaches described in Section 2 may not be practical because they would require solving an optimization problem from scratch for each change in staffing needs.

An effective (i.e., computationally feasible yet accurate) solution to such a dynamic hospital staff re-allocation problem is given by *Reinforcement Learning* (RL). We point out that a RL approach can be applied to a static scenario. In this regard, we model our problem as a *Markov Decision Process* [19], whose states correspond to the various hospitals. As described above, each hospital $\vec{h}_i$ is described by a three-component vector: the current number c_i of doctors on duty in that hospital, the minimum number m_i of doctors needed to ensure its operation, and the number t_i of doctors (target) that the hospital would like to have.

We employ a reward given by three contributions: *a) Deviation from target*, i.e., the reward penalises configurations where the current allocation of doctors is far from the desired target distribution; *b) Violations of minima*, i.e, heavy penalties are given if any hospital drops below its minimum number of staff units needed to guarantee its operation, and *c) Fairness in transfer procedures*, i.e., uneven transfers of doctors discouraged; the imbalance of the entire re-allocation process is quantified by the variance of the distribution of doctors across the various hospitals. Formally, denoting by y_i the output number of staff units at hospital $\vec{h}_i$ after staff redistribution among hospitals and by a the distribution of the overall staff units transferred from a hospital to another, our reward $\mathcal{R}$ is defined as:

$$\mathcal{R} = -\underbrace{\sum_{i=1}^{n}(y_i - t_i)^2}_{\text{target error}} - \underbrace{\eta \sum_{i=1}^{n} \max(0, m_i - y_i)}_{\text{minima violations}} - \underbrace{\sigma^2(a)}_{\text{fairness}}, \tag{1}$$

where $\eta \gg 1$ is a multiplicative factor whose role is to heavily penalize violations of the minimum staff number requirement. Here, we empirically consider $\eta = 100$. We adopt the well-established Proximal Policy Optimisation (PPO) algorithm of [14] to optimize the above reward. PPO, which is outlined in Algorithm 1 has two key features that make it the best choice here. First, it can operate in continuous spaces (thus allowing fractional resource transfer). Second, it yields a good tradeoff between *exploration policies* (i.e., the generation of new doctor transfer strategies) and *exploitation policies* (i.e., the application of staff transfer strategies that proved to be effective in the past). Next, we describe how the PPO algorithm is contextualized to our dynamic staff allocation scenario.

Our PPO agent learns to redistribute doctors to meet targets while respecting minima and, importantly, trying to minimize resource imbalances among hospitals. The PPO algorithm uses two separate neural networks, that is the *actor* and the *critic*.

The actor network defines the policy of the agent, i.e., it takes the current state s as input and computes the probability distribution $\pi_\theta(a|s)$ of all possible actions a for state s; such a distribution is used to select the next action.

Following the setting of [13], a fractional transfer of resources is permitted in our scenario. Thus, we have that the actor network parametrises a Gaussian distribution (with a predefined mean and standard

Algorithm 1 Proximal Policy Optimization for staff redistribution

Input:
  agent parameters: $\mathbf{c}$ (current), $\mathbf{m}$ (minima), $\mathbf{t}$ (targets)
  PPO hyperparameters: epochs K, clip ϵ, learning rates α_θ, α_ϕ
  Advantage parameters: GAE λ, discount γ
procedure PPO
  Initialize policy π_θ (actor) and value function V_ϕ (critic)
  for iteration $= 1, 2, \ldots$ **do**
    Collect trajectories $\{\psi_i\}$ by running π_θ in environment:
    for each agent $i \in \{1, \ldots, N\}$ **do**
      Observe state $s_i = [c_i, m_i, t_i]$
      Sample action $a_i \sim \pi_\theta(\cdot|s_i)$ (transfer matrix)
      Execute action a_i, get new state s_i' and reward r_i
      Store (s_i, a_i, r_i, s_i') in buffer
    end for
    Compute advantages $\hat{A}_t$ using GAE(λ, γ):
    for each trajectory ψ **do**
      Compute δ_τ (Equation (3))
      Compute $\hat{A}_\tau$ (Equation (4))
    end for
    Optimize surrogate objective for K epochs:
    for epoch $=$ 1 to K **do**
      Sample minibatch from buffer
      Compute $r_\tau(\theta) = \frac{\pi_\theta(a_\tau|s_\tau)}{\pi_{\theta_{\text{old}}}(a_\tau|s_\tau)}$
      Compute L^{CLIP} (Equation (5))
      Update policy (Equation (6))
      Update value function (Equation (7))
    end for
  end for
end procedure

deviation). The critic network is parametrised by ϕ and estimates the *expected cumulative future reward* starting from the state s_0:

$$V_\phi(s) \sim \mathbb{E}\left(\sum_{\tau=0}^{+\infty} \gamma^\tau r_\tau | s = s_0\right) \tag{2}$$

.

The γ coefficient (which we empirically set to 0.99 here) is the *discount factor* and its purpose is to reduce the effect of remote actions (i.e., actions chosen many steps before). The critic network acts as an estimator of the advantages related to a particular action.

Our algorithm collects possible trajectories to be used in the training phase. Specifically, it observes the current state for each time step and uses the distribution $\pi_\theta(a_i|s_i)$ to sample the next action a_i. We call s_i the state after performing action a_i and r_i the reward received.

Our algorithm computes the Generalised Advantage Estimation (GAE), i.e., it quantifies whether (and by how much) a given action produces benefits more than the expected benefits. To compute the GAE, we first consider the correct reward δ_τ obtained from the reward r_τ received by the agent at time step τ, and then adds to it the difference between the critic network at time $\tau + 1$ (weighted by the discount factor λ) and the critic network at time τ:

$$\delta_\tau = r_\tau + \gamma V_\phi(s_{\tau+1}) - V_\phi(s_\tau) \tag{3}$$

Next, we compute the expected advantage $\hat{A}_\tau$ as the weighted average of the corrected rewards δ_τ; the weights used to compute the average are exponentially decreasing and they are obtained by multiplying the discount factor γ by a coefficient λ (which we set equal to 0.95 in our tests). This leads to the following:

$$\hat{A}_\tau = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{\tau+l} \tag{4}$$

The $\gamma\lambda$ coefficient aims to achieve a good trade-off between the bias and variance derived from the trajectories, and, thus, it can be interpreted as a reward reshaping technique.

Our algorithm tries to improve the agent's policy by avoiding large but risky updates that could degrade performance. To do this, it introduces a *clipping* mechanism to limit how much the new policy can differ from the old policy (the one used to gain recent experience). Specifically, our algorithm first computes the ratio $r_\tau(\theta) = \frac{\pi_\theta(a_\tau|s_\tau)}{\pi_{\theta_{\text{old}}}(a_\tau|s_\tau)}$, which quantifies how much the new policy differs from the old one. We note that $r_\tau(\theta)$ could take large values, leading to the acceptance of new policies and the rejection of old ones; however, the new policies could cause a significant degradation in performance. For this reason, our algorithm uses a function called `clip` to constrain the probability ratio $r_\tau(\theta)$ to an interval of the type $[1-\varepsilon, 1+\varepsilon]$, thus constructing the following loss function:

$$L^{\text{CLIP}} = \mathbb{E}_\tau[\min(r_\tau(\theta)\hat{A}_\tau, \text{clip}(r_\tau(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_\tau)] \tag{5}$$

Then the policy is updated via a standard gradient descent step:

$$\theta \leftarrow \theta + \alpha_\theta \nabla_\theta L^{\text{CLIP}} \tag{6}$$

and the critic is updated via mean square minimization to improve value estimates:

$$L^{\text{VF}}(\phi) = \mathbb{E}_\tau\left[\left(V_\phi(s_\tau) - \hat{R}_\tau\right)^2\right] \tag{7}$$

where $\hat{R}_\tau = \sum_{\ell=0}^{+\infty} \gamma^\ell r_{\tau+l}$.

4. Experiments

We evaluated the effectiveness of our approach through a series of experiments mainly designed to address the question about which reward components most significantly influence RL performance. We considered a testbed made up of 150 randomly-generated hospitals; i.e., for each hospital the minimum number of staff needed to function properly, the current number of doctors on the rooster, and the target number were assigned at random. To evaluate the performance of the various methods, we adopt the following metrics metrics:

1. **Target Deviation** (MAE), defined as the average difference between actual staff levels and targets:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - t_i| \tag{8}$$

   Ideally, we would like each hospital to have as close to the target number of doctors as possible; thus, the lower MAE, the more effective the approach to redistributing doctors.

2. **Gini Index** ($\mathcal{G}$) [20], defined as

$$\mathcal{G} = \frac{1}{N}\left(N + 1 - 2\sum_{i=1}^{N}\frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{N} y_j}\right) \tag{9}$$

   The Gini Index is one of the most commonly used inequality measures (e.g., income inequality or inequality in life expectancy). It ranges from 0 (perfect equality) to 1 (perfect inequality). In our case, a $\mathcal{G}$ close to 1 indicates that the available doctors are concentrated in a few hospitals, which is undesirable. Thus, ideally, values of $\mathcal{G}$ close to 0 should be achieved.

4.1. Results

We analyzed a dynamic scenario where the number of doctors per hospital varies over time. This assumption reflects real-world conditions: for instance, tourist destinations face seasonal population fluctuations, leading to temporary mismatches between healthcare demand and the baseline workforce capacity, and, thus, the hospital needs extra resources.

We considered a finite time horizon T partitioned into 50 discrete time steps and evaluated two distinct scenarios: a) *Shifting Targets*, where hospital-specific targets undergo minor adjustments every 10 time steps: specifically, each hospital loses or gains at most two doctors per adjustment; b) *Sudden Shock*, where a randomly selected hospital loses half its workforce at the midpoint of the time horizon $T/2$. We assume that the hospital that suffers the loss of doctors is able to satisfy the minimum operating conditions.

We tested the proposed RL approach. As baselines, we employed the four methods devised in our previous work [13] for a static setting, namely methods QP, Progressive Taxation, NWO and hybrid. For each scenario and tested method, we ran twenty independent simulations and generated time series tracking the evolution of Target Deviation and Gini. The results, averaged over all simulations, are plotted in Figures 1a and 1b.

One immediate conclusion from the experiments is that the results for NWO, QP, and Hybrid are almost *stable* in both scenarios. We argue that this stability in performance is due to the fact that the three strategies (NWO, QP and Hybrid) dynamically re-plan staffing *from scratch,* for each hospital and for each change in staffing needs. Therefore, they are *optimal* in the sense that they immediately react to change and construct the best possible resource re-allocation, albeit with different strategies. Reallocation from scratch, unfortunately, is unrealistic for several reasons. First, computational cost grows quickly with the number N of agents (hospitals). Second, the actual delays involved with staff/resource re-allocation may fall behind the rapid pace of updates, thus rendering intermediate solutions *stale* before they are executed. Third, there might be hidden size effects with large re-allocations, due to institutional, logistical, or ethical constraints.

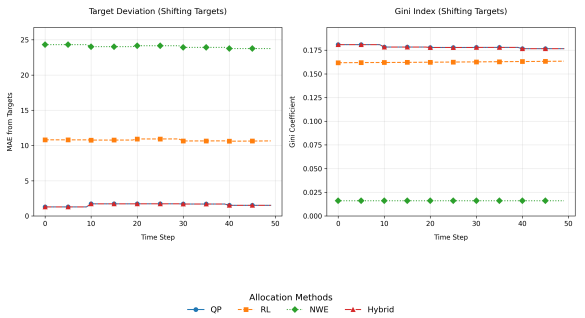
The second takeaway message from the experimental results is that RL is able to effectively *learn* profitable re-allocation strategies, rather than recalculating optimal allocations from scratch. That becomes apparent in the Shifting Target scenario: RL reacts well to small variations in staff levels and achieves a relatively stable Target Deviation and Gini Index.

In the *Sudden Shock* scenario, the RL strategy initially experiences a sharp decline in both metrics following the drastic loss of staff at one location. However, the RL agent has learned an effective resource redistribution strategy that allows it to stabilise Target Deviation and Gini, indeed at slightly lower levels that before the shock.

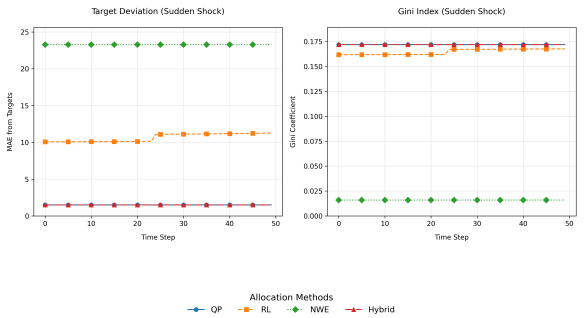
4.2. Ablation study

A controlled ablation study was designed to systematically evaluate how individual reward components influence the performance of reinforcement learning in our scenarios. We trained four different RL agents: (1) the proposed agent equipped with the full reward formulation (Target Deviation + minimum staffing penalty + fairness), (2) an agent without the fairness term, (3) an agent without the minimum staffing penalty, and (4) an agent without the Target Deviation component. Each configuration was evaluated over five independent trials to ensure statistical robustness.

The results of this ablation study are reported in Figure 2. The figure shows that the Target Deviation component is the most significant factor for both Target Deviation and Gini Index. The ablation study also reveals how the removal of the minimum and fairness component leads to a lower Gini Index than in the case of full reward, but at the price of a consistent deterioration of Target Deviation. All in all, the full reward allows for reaching the best tradeoff between Target Deviation and Gini Index.



(a) Average Target Deviation (MAE) and Gini Index of the RL (proposed), NWO, QP, Hybrid (baselines) strategies in the Shifting Targets scenario.



(b) Average Target Deviation (MAE) and Gini Index of the RL (proposed), NWO, QP, Hybrid (baselines) strategies in the Sudden Shock scenario.

Figure 1: Comparison of RL (proposed), NWO, QP, Hybrid (baselines) strategies in the dynamic scenario

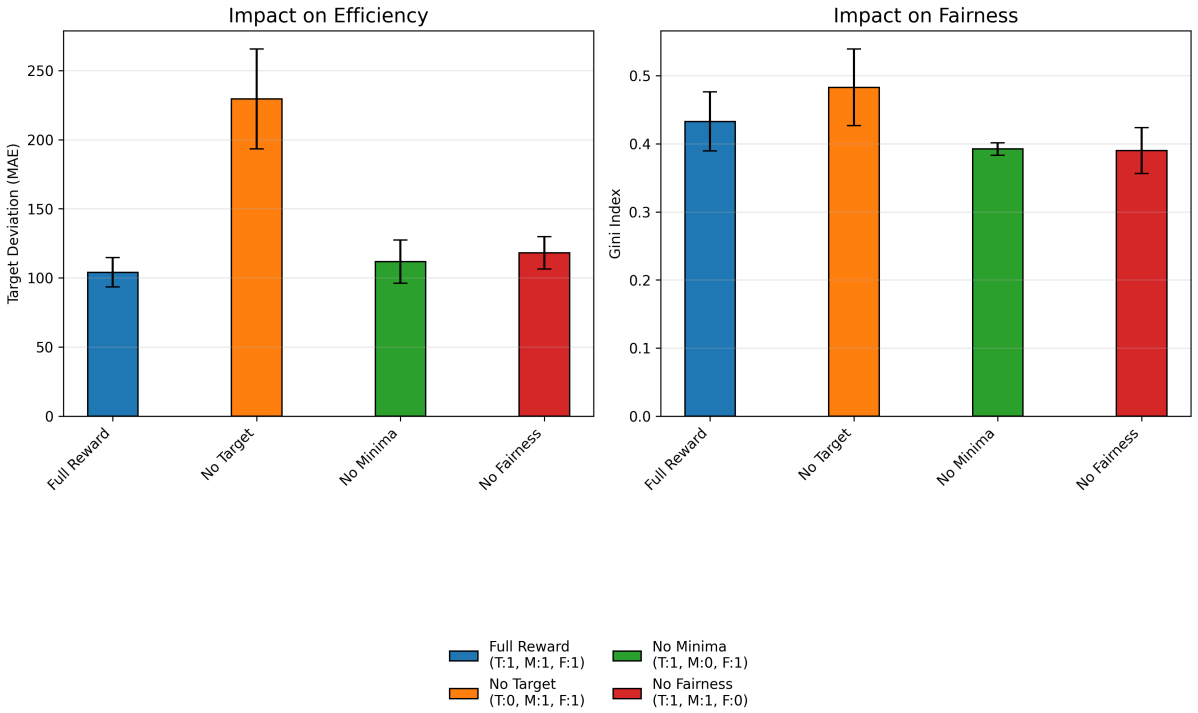


Figure 2: Ablation study of the proposed RL strategy for each component of the reward.

5. Related Work

Multiagent resource allocation (MARA) constitutes a fundamental challenge in multiagent systems, requiring autonomous agents to distribute limited resources in ways that balance efficiency and fairness. Such problems arise ubiquitously [1]. As systems grow in scale and complexity, designing mechanisms that reconcile individual agent incentives with collective welfare becomes increasingly critical.

At its core, MARA involves agents negotiating resource distributions, often encountering dilemmas where self-interest conflicts with group optimality. These interactions necessitate formal frameworks that model agent behavior while guaranteeing allocations beneficial to all stakeholders [21]. Central to this challenge is the concept of *fairness*, a principle vital to both human societies and artificial systems.

In applications ranging from traffic control to cloud computing, equitable resource distribution directly impacts system stability, productivity, and long-term performance [22]. The dual challenge lies in formalizing fairness mathematically while enabling practical implementation in dynamic environments. Two dominant fairness paradigms guide allocation strategies, namely *proportional fairness*, according to which resources are allocated proportionally to agents' contributions or needs, and *envy-freeness*, which ensures no agent prefers another's allocation over their own [22, 23]. While extensively studied in static settings, recent work extends these criteria to dynamic environments where resources and agent populations evolve over time [24]. This shift reflects real-world demands for adaptive solutions that maintain fairness guarantees amid uncertainty.

Our work can be positioned under the broad umbrella of MARA. Specifically, we propose novel strategies for the re-allocation of resources in a multi-agent, cooperative setting.

Centralized vs. decentralized approaches. Traditional centralized methods, such as the Hungarian and Gale-Shapley algorithms, rely on a central authority with full system knowledge to compute optimal allocations [9]. *Nash Welfare Optimization (NWO)* offers a principled framework for fair resource distribution by maximizing the geometric mean of agent utilities. The NWO approach inherently balances efficiency and equity, prioritizing improvements for agents with lower initial utility [8]. It occupies a middle ground between *utilitarian welfare* (maximizing total utility) and *egalitarian welfare* (maximizing minimum utility) and satisfyies key axiomatic properties such as scale invariance, Pareto efficiency, and independence of irrelevant alternatives [11, 10]. Due to these strengths, NWO has been applied successfully in collective decision-making, project funding allocation, and fair division problems [10]. For multi-agent bandits, NWO provides fair allocations when multiple agents have heterogeneous preferences [11].

In this work, we provide a novel contextualization of NWO to the multi-agent, cooperative resource re-allocation setting, Specifically, in our scenario, the utility of an individual agent is defined as the logarithm of the difference between the number of doctors y_i allocated to hospital $\vec{h}_i$ and the minimum required m_i (possibly incremented by one to ensure positivity in the logarithmic function). Consequently, the hard constraint $y_i \geq m_i$ is embedded within the objective function. However, NWO provides no guarantee that allocations will achieve target staffing levels t_i, which represent aspirational goals for optimal service delivery. To address this limitation, we propose a novel staff reallocation method based on quadratic programming that explicitly minimizes deviations from target staffing levels while enforcing minimum operational constraints.

Dynamic resource allocation. In real-world applications, resource allocation evolves over time, necessitating adaptive strategies. *Constrained multi-agent Markov decision processes (MDPs)* [25] formalize sequential allocation problems, modelling resource constraints that couple otherwise independent agents. *Reinforcement Learning (RL)* methods have proven effective in dynamic environments by continuously learning from system feedback: agents refine policies through trial-and-error interactions, achieving both higher overall utility and reduced violation of fairness requirements compared to traditional strategies [26, 4]. Recent applications span cloud computing [4] and healthcare services [2], where RL adapts to fluctuating demands without manual intervention.

Our work advances this line of research by proposing the first RL-based approach in multi-agent, cooperative resource re-allocation scenarios. We employ the Proximal Policy Optimization (PPO) algorithm [14], ensuring operational feasibility by limiting the number of doctors transferred at each step. To address competing objectives—minimizing Target Deviation, avoiding minima violations, and ensuring fair workforce distribution—we design a composite reward function that integrates these goals. This formulation enables the agent to learn adaptive redistribution strategies while respecting hard constraints, overcoming the limitations of static methods.

6. Discussion, Conclusions and Future Work

This paper addressed the problem of cooperative resource re-allocation in dynamic multi-agent environments, focusing on the redistribution of hospital staff. By combining classic optimization techniques

with Reinforcement Learning (RL) approaches, we demonstrated that learning-based strategies, in particular Proximal Policy Optimization (PPO), offer superior adaptability and scalability in dynamic settings. Our experimental evaluation showed that RL can effectively maintain fairness and minimise resource mismatches under uncertain and evolving conditions.

While our results are encouraging, several limitations remain. First, the study focuses on PPO, but other alternative RL algorithms such as DDPG, SAC, or A3C are available. Future work will extend the evaluation to a wider range of RL methods and assess their relative strengths and weaknesses. Furthermore, our model currently assumes fully cooperative agents; however, in realistic scenarios, agents may behave selfishly or misreport their needs. Investigating mechanisms for dealing with partial cooperation and strategic behaviour is an important direction for future research. For example, we could experiment with Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [27], a method for learning in partially cooperative environments, to deal with semi-cooperative hospitals. Another limitation concerns the lack of formal guarantees regarding fairness and optimality under dynamic reallocation policies. While empirical results are promising, deriving theoretical performance bounds for RL-based allocation remains an open challenge. In addition, expanding the evaluation to include health-care-specific metrics – e.g., resilience to recurrent shocks, hospital load variance, and service continuity – will make the performance assessment more comprehensive.

Looking ahead, we see several promising developments. One direction involves the integration of online learning mechanisms to make agents continuously adapt their resource allocation strategies without offline retraining. Another direction is to embed explainability in allocation policies, making them transparent and interpretable to human operators–a crucial aspect in sensitive domains such as healthcare. Finally, hybrid frameworks that combine offline planning with online reactive adjustments may provide an effective tradeoff between optimality and adaptability.

Acknowledgments

Research partially supported by the PNRR Project CUP E13C24000430006 “Enhanced Network of intelligent Agents for Building Livable Environments - ENABLE”, and by PRIN 2022 CUP E53D23007850001 Project “TrustPACTX - Design of the Hybrid Society Humans-Autonomous Systems: Architecture, Trustworthiness, Trust, EthiCs, and EXplainability (the case of Patient Care)”.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly, solely to spell check and improve the grammar. The tool was not used to alter, generate, or influence the semantic contents of the paper. The authors retain full responsibility for the accuracy, originality, and integrity of the entire work.

References

[1] N. Jong, P. Stone, M. Taylor, Multiagent resource allocation: A review of mechanisms and applications, Autonomous Agents and Multi-Agent Systems 22 (2008) 1–29.

[2] Y. Zhao, N. Behari, E. Hughes, E. Zhang, D. Nagaraj, K. Tuyls, A. Taneja, M. Tambe, Towards a pretrained model for restless bandits via multi-arm generalization, in: Proc. of the International Joint Conference on Artificial Intelligence, (IJCAI 2024), Jeju, South Korea, 2024, pp. 321–329.

[3] J. Cui, Y. Liu, A. Nallanathan, Multi-agent reinforcement learning-based resource allocation for UAV networks, IEEE Transactions on Wireless Communications 19 (2020) 729–743.

[4] Y. Zhao, Y. Liu, B. Jiang, T. Guo, CE-NAS: an end-to-end carbon-efficient neural architecture search framework, in: Proc. of the International Conference on Advances in Neural Information Processing Systems 38 (NIPS 2024), Vancouver, BC, Canada, 2024.

[5] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, G. Caldarelli, Debtrank: Too central to fail? financial networks, the fed and systemic risk, Scientific Reports 2 (2012) 541. URL: https://doi.org/10.1038/srep00541. doi:10.1038/srep00541.
[6] J. Tong, B. de Keijzer, C. Ventre, Reducing systemic risk in financial networks through donations, in: Proc. of the European Conference on Artificial Intelligence (ECAI 2024), volume 392, IOS Press, Santiago de Compostela, Spain, 2024, pp. 3405–3412.
[7] S. de Jong, S. Uyttendaele, K. Tuyls, Learning to reach agreement in a continuous ultimatum game, J. Artif. Intell. Res. 33 (2008) 551–574. URL: https://api.semanticscholar.org/CorpusID:13248455.
[8] L. Zhang, Y. Xu, F. Fang, Y. Yang, Online nash social welfare maximization in multi-agent systems, in: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI), 2024.
[9] A. Nongaillard, H. Sohier, V. Hilaire, Centralized and distributed approaches for resource allocation: A comparative study, Journal of Intelligent Manufacturing 27 (2016) 789–803.
[10] T. Delemazure, F. Durand, F. Mathieu, Aggregating correlated estimations with (almost) no training, in: Proc. of the European Conference on Artificial Intelligence, (ECAI 2023 - 26th), volume 372 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, Krakow, Poland, 2023, pp. 541–548.
[11] S. Hossain, E. Micha, N. Shah, Fair algorithms for multi-agent multi-armed bandits, Advances in Neural Information Processing Systems 34 (2021) 24005–24017.
[12] I. Papanicolas, L. R. Woskie, A. K. Jha, Health care spending in the united states and other high-income countries, Jama 319 (2018) 1024–1039.
[13] S. Costantini, G. De gasperis, P. De Meo, A. Provetti, Resource allocation with cooperative agents, in: Proc. of the 32nd RCRA workshop on Experimental evaluation of algorithms for solving problems with combinatorial explosion (RCRA 2025), held in conjunction with the 41st International Conference on Logic Programming (ICLP). TO APPEAR, 2025.
[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).
[15] S. P. Boyd, L. Vandenberghe, Convex optimization, Cambridge university press, 2004.
[16] J. E. Stiglitz, J. K. Rosengard, Economics of the public sector: Fourth international student edition, WW Norton & Company, 2015.
[17] H. Moulin, Fair division and collective welfare, MIT press, 2004.
[18] J. F. Nash, et al., The bargaining problem, Econometrica 18 (1950) 155–162.
[19] E. A. Feinberg, A. Shwartz, Handbook of Markov decision processes: methods and applications, volume 40, Springer Science & Business Media, 2012.
[20] F. A. Farris, The gini index and measures of inequality, The American Mathematical Monthly 117 (2010) 851–864.
[21] D. Ceragioli, F. Rossi, B. Venable, Fairness-aware distributed planning for resource allocation in multiagent systems, in: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), 2024.
[22] A. Jiang, K. Leyton-Brown, Fairness in multi-agent systems with reinforcement learning, Artificial Intelligence 275 (2019) 25–64.
[23] Y. Chen, D. C. Parkes, Envy-free allocation in combinatorial auctions, Games and Economic Behavior 70 (2010) 1–14.
[24] I. A. Kash, A. D. Procaccia, N. Shah, No-envy learning in symmetric auctions, in: Proceedings of the 14th ACM Conference on Electronic Commerce, 2013, pp. 297–314.
[25] V. Nijs, N. Vlassis, M. De Weerdt, Dynamic multi-agent resource allocation under constraints, Autonomous Agents and Multi-Agent Systems 35 (2021) 1–34.
[26] E. Y. Yu, Z. Qin, M. K. Lee, S. Gao, Policy optimization with advantage regularization for long-term fairness in decision systems, in: Proc. of the International Conference on Neural Information Processing Systems (NIPS 2022), New Orleans, LA, USA, 2022.
[27] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, Advances in Neural Information Processing Systems (NeurIPS) 30 (2017).