Strategic Workforce Planning with Deep Reinforcement Learning

Yannick Smit[1⋆], Floris den Hengst[2,3⋆[0000−0002−2092−9904]], Sandjai Bhulai[2[0000−0003−1124−8821]], and Ehsan Mehdad[3]

[1] Universiteit van Amsterdam yannick.smit@xs4all.nl
[2] Vrije Universiteit Amsterdam s.bhulai@vu.nl
[3] ING Bank N.V., Amsterdam, the Netherlands
{Floris.den.Hengst,Ehsan.Mehdad1}@ing.com

Abstract. This paper presents a simulation-optimization approach to strategic workforce planning based on deep reinforcement learning. A domain expert expresses the organization's high-level, strategic workforce goals over the workforce composition. A policy that optimizes these goals is then learned in a simulation-optimization loop. Any suitable simulator can be used, and we describe how a simulator can be derived from historical data. The optimizer is driven by deep reinforcement learning and directly optimizes for the high-level strategic goals as a result. We compare the proposed approach with a linear programming-based approach on two types of workforce goals. The first type of goal, consisting of a target workforce, is relatively easy to optimize for but hard to specify in practice and is called *operational* in this work. The second, *strategic*, type of goal is a possibly non-linear combination of high-level workforce metrics. These goals can easily be specified by domain experts but may be hard to optimize for with existing approaches. The proposed approach performs significantly better on the strategic goal while performing comparably on the operational goal for both a synthetic and a real-world organization. Our novel approach based on deep reinforcement learning and simulation-optimization has a large potential for impact in the workforce planning domain. It directly optimizes for an organization's workforce goals that may be non-linear in the workforce composition and composed of arbitrary workforce composition metrics.

Keywords: Deep Reinforcement Learning · Optimization · Simulation · Strategic Workforce Planning.

1 Introduction

In order to achieve their strategic goals, organizations need to have the right people in the right place at the right time. *Strategic workforce planning* (SWP) is the business process in which the required actions to meet an organization's workforce needs are identified [1]. SWP has been recognized as an important problem

⋆ Authors contributed equally.

across sectors [4,3,5] and is expected to grow in importance with knowledge and human capital becoming increasingly important drivers of economic growth [18]. Workforce planning helps organizations with forecasting their workforce needs given a range of possible business scenarios and includes predicting the impact of various programs and policies on talent attraction and retention, showing how the impact varies across different segments of the workforce, modeling the impact of employee attrition and movements within the organization, and quantifying the financial impact of workforce decisions [1].

SWP problems are challenging since they require a deep understanding of the organization's high-level strategic goals and constraints on the one hand and technical knowledge to express these as an optimization problem on the other. The problem formulation should correctly capture the organization's workforce goals and constraints into its objective, address the aforementioned aspects of uncertainty, and be both actionable and computationally tractable. As a result, achieving impact with SWP typically requires careful collaboration between experts from the HR and analytics domains.

The SWP problem has attracted substantial interest from researchers as a result. Historically, these have focused on relatively simple and specific settings, e.g., problems of a relatively small scale [16], with a homogeneous workforce [4,20], and an objective function linear in the workforce composition [9,10]. Recently, researchers have addressed some of these limitations with more advanced techniques that explicitly include uncertainty of the workforce dynamics [13], that include employee attributes, such as age, skill, and position [3,6], and that use a piece-wise linear objective [7]. Although more general than previous methods, these still rely on problem specifics to cast the organizations' goals and constraints into a tractable optimization problem. This limits their applicability and comes at a significant analysis and modeling burden.

In this work, we propose a generic and widely applicable approach. In our approach, a policy that optimizes a strategic workforce objective is derived with deep reinforcement learning (DRL). Since DRL does not depend on the specifics of the objective, it can be defined as a non-linear combination of high-level workforce metrics. The optimal policy is determined with DRL in a simulation-optimization loop. The optimization step in this loop does not depend on the internals of the simulator, so that the approach can be applied to a wide range of simulators. We also describe how a simulator can be estimated from data on historical workforce compositions so that only the objective and a data set are required as inputs. Additionally, our approach is capable of handling large problems and fine-grained decision-making as a result of the usage of neural networks in estimating the optimal policy. Our approach improves the usability, granularity, and quality of SWP decision support.

1.1 Related Work

The application of different simulation paradigms in finding the optimal workforce planning decisions is very popular; see [2,15,14] and also see [1] for a discussion of simulation in workforce planning in industry. The adoption of deep

reinforcement learning for simulation-optimization has recently become popular in academia and industry; see [11], Pathmind and project Bonsai by Microsoft. To the best of our knowledge, however, no studies have proposed to address SWP with DRL, which brings various benefits to this domain: it does not require any specific domain knowledge, scales well to large problems, and makes no assumptions on, e.g., linearity of the objective function.

This work is organized as follows. We first introduce SWP as an optimization problem, including the modeling of the workforce dynamic and the formulation of optimization objectives. We then introduce the simulation-optimization loop and detail the DRL optimizer. We describe the experimental setup and results, which show that our approach finds suitable policies for high-level objectives for both a synthetic and real-life organization. We conclude that our approach enables direct optimization of strategic workforce goals.

2 Strategic Workforce Planning as Optimization

In this section, we present a quantitative framework for SWP. We first detail a descriptive model of the workforce. This model factors the total workforce into groups of individuals with similar attributes of interest called *cohorts*. Attributes, such as productivity, skills, and manager status, can be included based on the goals and constraints of the organization. We then detail how the dynamics are modeled. Finally, we describe how strategic workforce goals and constraints can be formulated as optimization objectives.

2.1 Cohort Model

We define employee attributes as a set of variables $Y = (Y_1, \ldots, Y_m)$ so that each employee with attributes $(Y_1 = y_1, \ldots, Y_m = y_m)$ can be described by values $(y_1, \ldots, y_m)$ and all employees with the same values can be grouped into the same cohort $C_i \in \mathcal{C} = \{C_1, \ldots, C_n\}$. The number of cohorts n depends on the number of attributes m and the cardinalities $|Y_i|$ of these attributes, i.e., $n = |\mathcal{C}| = \prod_{i=1}^{m} |Y_i|$. Note that n grows as a combination of attributes, so that more fine-grained modeling results in a larger number of cohorts quickly.

We now turn to a model of the evolution of a workforce over time. Specifically, we consider discrete time steps of an arbitrary fixed length (e.g., monthly, quarterly, or yearly) $0 < t \leq T$ for some finite horizon $T < \infty$. At each time point t, the number of employees for a particular cohort C_i is defined as a random variable (r.v.) $X_{i,t} \in \mathbb{N}_{\geq 0}$ and the total workforce as a combination of all cohorts $X_t = \left(X_{(1,t)}, \ldots, X_{(n,t)}\right) \in \mathbb{N}_{\geq 0}^{n}$. The dynamics of these so-called *headcounts* can now be modeled as a Markov chain. Its state space consists of all possible headcount compositions. We assume a scalar $X_{\max} < \infty$ for the maximum number of employees per cohort and define the state space of the Markov chain $\mathcal{S} = \{s \in \mathbb{N}_{\geq 0}^{n} | s \leq (X_{\max})^n\}$.

For any organization and for any time step, we know that an individual can either (i) leave the organization organically due to, e.g., retirement, voluntarily

leaving etc. (ii) leave the organization as a result of a management decision, (iii) move from one cohort to another cohort organically, (iv) be moved from one cohort to another cohort by the organization and (v) enter the organization. With this knowledge, the transition function can be factorized into components, so that for every t:

$$X_{t+1} = X_t - O_t - L_t + \mathbb{1}^n M_t - \mathbb{1}^n M_t' + \mathbb{1}^n N_t - \mathbb{1}^n N_t' + H_t, \tag{1}$$

where $\mathbb{1}^n$ is an n-dimensional vector of ones and (i) O_t an n-dimensional r.v. representing organic leavers per cohort, (ii) L_t an n-dimensional r.v. representing organization-initiated leavers, (iii) M_t an $n \times n$ random matrix of employees moving between cohorts organically, (iv) N_t an $n \times n$ random matrix of moves between cohorts initiated by the organization, and (v) H_t an n-dimensional r.v. of new hires. This model describes how the workforce changes over time and it allows to easily formalize strategic workforce goals as optimization objectives as described in the next section.

2.2 Optimizing the Cohort Model

In this section, we cast the SWP problem as an optimization problem. The first step is to identify the actions available to the organization. We assume that these are direct and indirect controls on the Markov chain in Equation 1. In general, the transitions L_t, N_t, and H_t are controlled by the organization directly. Additional controls may be in place to affect the other r.v.'s indirectly. For example, an employee retention plan can be included to affect the attrition O_t. The cohort model supports both direct and indirect controls, and these can be included based on the organization's needs.

The organization should take those actions that result in the most suitable workforce at every time step. We here formalize the organization's actions as some set $\mathcal{A}$ and a particular action at time t as $A_t \in \mathcal{A}$ and refer the reader to Section 4 for examples. We assume that each workforce composition X_t and A_t can be assigned a numerical value corresponding to the particular SWP goal of the organization with some function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The objective, now, is to maximize this value over time by sampling appropriate states and actions in the system in Equation 1 until some horizon T:

$$\begin{aligned} A^* &= \underset{A_0,\ldots,A_T}{\arg\max}\, \mathbb{E}\left[\sum_{t=0}^{T} r(X_{t+1}, A_t)\right] \\ \text{s.t. } X_{t+1} &= X_t - O_t - L_t + \mathbb{1}^n M_t - \mathbb{1}^n M_t' + \mathbb{1}^n N_t - \mathbb{1}^n N_t' + H_t, \\ &\text{and } O_t, L_t, M_t, N_t, H_t \text{ dependent on } A_t \\ &\text{for } t = 0, \ldots, T, \text{ and a given } X_0. \end{aligned} \tag{2}$$

Having defined the general optimization objective, we now turn to examples of suitable reward functions. A reward function should reflect the strategic workforce goals of the organization accurately. Because of the strategic nature

of SWP, a goal is usually composed of multiple terms. General terms such as headcount and budget, SWP-specific terms such as average span of control[4], job level[5] and manager status, and finally, organization-specific metrics such as productivity, skills, and diversity may all be included.

Strategic Workforce Goals The example strategic workforce goal is composed of three components, here presented by decreasing importance. The primary component consists of bounds for headcounts for each cohort. The second component contains a target average span of control across the organization. In general, such a target span of control is attained by multiple workforce compositions. The third component, therefore, specifies that minimal salary costs are preferred. We formalize this strategic goal by formalizing each component and then combining the components in an overall objective.

To formalize the objective based on headcount bounds, we penalize cohorts that are out of bounds:

$$r_{\mathrm{b}}(X_t) := -\sum_{i=1}^{n} \mathbb{1}_{\{X_{i,t} \notin [\ell_i, u_i]\}}, \tag{3}$$

where $\ell, u \in \mathbb{N}^n_{\geq 0}$ are lower and upper bounds for all n cohorts. Next, we define a component for achieving the target span of control. It is similar to the objective for target headcounts in Equation (9):

$$r_{\mathrm{soc}}(X_t) := \exp\left(\frac{-\alpha_{\mathrm{soc}}\left(\mathrm{soc}(X_t) - G_{\mathrm{soc}}\right)^2}{G^2_{\mathrm{soc}}}\right), \tag{4}$$

where $G_{\mathrm{soc}} > 0$ is a target average span of control, $\alpha_{\mathrm{soc}} > 0$ a precision parameter, and $\mathrm{soc}(X_t)$ a function that returns the average span of control for X_t:

$$\mathrm{soc}(X_t) := \frac{X_{(n/2+1,t)} + \cdots + X_{(n,t)}}{X_{(1,t)} + \cdots + X_{(n/2,t)}}. \tag{5}$$

The third and final component can be formalized based on a function $\mathrm{sal}(X_t)$ that returns the estimated total salary cost for a workforce X_t. This final component has the lowest priority. Therefore, we only assign a positive value based on salary if the span of control component is sufficient, as expressed by a lower bound $\ell_{\mathrm{soc}} \in [0, 1]$:

$$r_{\mathrm{sal}}(X_t) := \begin{cases} r'_{\mathrm{sal}}(X_t), & \text{if } r_{\mathrm{soc}}(X_t) > \ell_{\mathrm{soc}}, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

for a salary normalized to $[0, 1]$ based on the cohort bounds ℓ, u:

$$r'_{\mathrm{sal}}(X_t) := \mathrm{clip}\left(\frac{\mathrm{sal}(X_t) + \mathrm{sal}(\ell)}{\mathrm{sal}(\ell) - \mathrm{sal}(u)}, 0, 1\right). \tag{7}$$

[4] The average number of direct reports of managers in the organization.

[5] A metric to express responsibilities and expectations of a role in the organization, usually associated with compensation in some way.

The *strategic* objective is composed of the sub goals in Equations (3)-(6). We combine the components to reflect all sub goals states earlier:

$$r_{\mathrm{s}}(X_t) := r_{\mathrm{b}}(X_t) + r_{\mathrm{soc}}(X_t) + r_{\mathrm{sal}}(X_t). \tag{8}$$

The simulation-optimization approach proposed in this work targets the direct optimization of objectives that reflect an organization's *strategic* workforce goals and that may be non-linear and composed of arbitrary workforce metrics.

Operational Workforce Goals Another type of workforce goal is to meet a particular known demand for employees in each cohort. This type of goal is relatively easy to optimize for but hard to specify in practice. For this goal, a reward can be assigned based on a distance between the current workforce X_t and the known target composition $X^* = (X_1^*, \ldots, X_n^*)$ for all n cohorts. To ensure that the cohorts contribute uniformly to this reward, headcounts need to be scaled to $[0, 1]$. Now, the following rewards an observed headcount $X_{i,t}$ for a single cohort i based on its target X_i^*:

$$r_{\mathrm{c}}(X_{i,t}) := \begin{cases} \exp\left(\frac{-\alpha(X_{i,t}-X_i^*)^2}{\left(X_i^*\right)^2}\right), & \text{if } X_i^* > 0, \\ \exp\left(-\alpha X_{i,t}^2\right), & \text{if } X_i^* = 0, \end{cases} \tag{9}$$

where the so-called *precision* parameter $\alpha > 0$ specifies how strictly to penalize sub-optimal headcounts. A simple operational reward averages over all n cohorts:

$$r_{\mathrm{o}}(X_t) := \frac{1}{n} \sum_{i=1}^{n} r_{\mathrm{c}}(X_{i,t}), \tag{10}$$

These *operational* workforce goals are generally easy to optimize for using established optimization techniques since they can be cast as linear optimization problems. However, defining the required headcounts for all cohorts to meet high-level workforce goals is very hard in practice.

3 Simulation-Optimization with Deep Reinforcement Learning

We propose a simulation-optimization loop for solving SWP problems. Figure 1 contains a visualization of this loop. First, the user specifies the strategic workforce goals of the organization as a reward function to maximize. This function may be any arbitrary, e.g., a non-linear function defined over a cohort representation of the workforce. Next, a policy is learned by a DRL agent by interacting with a simulator. This simulator can be any suitable black-box simulator that outputs a cohort representation of the workforce and can take into account the decisions made by the agent. By using DRL for optimization, the strategic goals are optimized for directly, and, hence, the resulting policy informs the user in taking the right workforce decisions for their strategic workforce goals. If historical data of the workforce is available, then this simulator can be learned from data as described in Section 3.2.

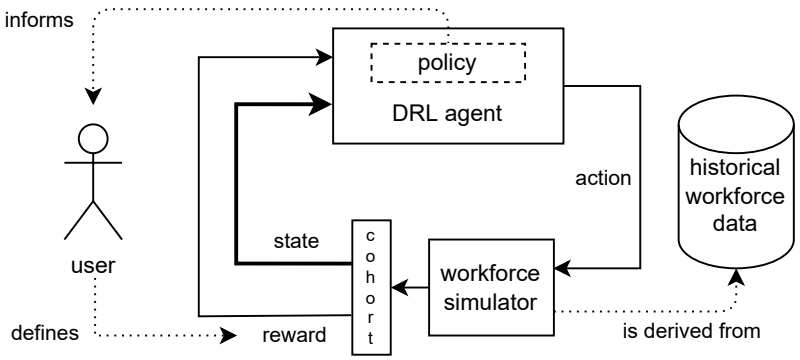

Fig. 1. Overview of the simulation-optimization approach. A user specifies the organization's strategic workforce goal. A black-box workforce simulator is then used to find a policy that directly optimizes for the goal with DRL. This policy helps the user making informed workforce decisions.

3.1 Deep Reinforcement Learning for Workforce Planning

Formally, we cast the SWP problem as a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ an action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ a transition function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ a reward function, and $\gamma \in (0, 1]$ a discount factor to balance immediate and future rewards. The decisions of the agent are defined by its policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which depends on a parameter vector θ which can, e.g., be a neural network. The goal of the agent is to maximize the expected discounted return $J(\theta) := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r_{t+1}\right]$, which can be done by tuning parameters θ with an algorithm that alternates simulating experience in the environment and optimizing the policy. Here $r_{t+1} = r(s_t, a_t)$ and $\mathbb{E}_{\pi_\theta}$ indicates that $s_{t+1} \sim \mathcal{P}(\cdot, a_t, s_t)$ and $a_t \sim \pi_\theta(\cdot|s_t)$.

In the proposed framework, the state space of the MDP is equal to the state space of the Markov chain over headcounts, i.e., $\mathcal{S} = \{s \in \mathbb{N}_{\geq 0}^n | s \leq (X_{\max})^n\}$. The optimization algorithm uses a neural network to evaluate the value of each state. To help the convergence of the network and significantly reduce training time, the inputs to the network are normalized. Hence, we implement the state space of the MDP as a continuous space $\hat{\mathcal{S}} = [0, 1]^n$, where states are defined as $s_t = \left(\frac{X_{1,t}}{X_1^{max}}, \ldots, \frac{X_{n,t}}{X_n^{max}}\right)$ for training. The action space is given by the controls over the workforce as described in Section 2, for example, a multi-discrete set of numbers of employees that enter or leave the organization for each cohort. For the purposes of optimization, the dynamic model $\mathcal{P}$ is assumed to be unknown so that any suitable simulator can be used. The reward function is defined by an end user based on the organization's strategic workforce goals. It can be composed of arbitrary and non-linear workforce metrics of interest to the organization, see Section 2.2 for details and examples.

In the optimization step a policy is updated to optimize the given objective. This update is performed by approximate gradient ascent on θ, i.e., iteratively update $\theta_{k+1} = \theta_k + \eta\widehat{\nabla_\theta J(\theta)}$. The gradient $\nabla_\theta J(\theta)$ is estimated by $\hat{\mathbb{E}}_t\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t\right]$, where $\hat{\mathbb{E}}_t$ denotes an empirical estimate over a batch of samples collected over time and $\hat{A}_t$ is an estimator of the advantage function. While our approach is generic to various optimization algorithms, we propose to use Proximal Policy Optimization (PPO) as it has shown to be suitable in high-dimensional settings with non-linear rewards [19].

3.2 Simulating the Workforce

This section details how the dynamics of a cohort model from Section 2.1 can be estimated from data. Estimation is necessary for two reasons. Firstly, the dynamics may simply not be available to the organization. Secondly, it may be problematic to fully elaborate the dynamics up-front due to the complexity of the problem. Specifically, the size of the state space of the cohort model Markov chain grows exponentially in the number of cohorts. As a result, it becomes infeasible to analytically define it fully for reasonably large organizations.[6] Hence, we estimate the dynamics from data with simplifications that apply to the cohort model.

In many cases, Equation (1) can be simplified by assuming limited control of the workforce by management. For example, if we only model management-controlled hires and leavers, N_t becomes equal to the zero matrix and $A_t := H_t - L_t$ for the combined movement of hires and leavers by the organization. The part of the transition function that is out of management control is now given by $X_{t+1} = X_t - O_t + \mathbb{1}M_t - \mathbb{1}M'_t$. Note that the diagonal entries of M_t can be chosen arbitrary, since $(M_t - M'_t)_{i,i} = 0$ for all $i = 1, \ldots, n$. By realizing that the numbers of employees that remain in cohort i is equal to the headcount of cohort i minus the number of employees that move to any other cohort or organically leave the organization, we may set $M_{i,i,t} = X_{i,t} - \sum_{j=1, j\neq i}^{n} M'_{i,j,t} - O_{i,t}$, or $X_t = \mathbb{1}M'_t + O_t$. It follows that we can then simply write $X_{t+1} = \mathbb{1}M_t$ for the stochastic dynamics of the cohort model in general.

We observe that all employees within a certain cohort are indistinguishable for the purposes of SWP. Hence, approximation of the dynamics of Equation (1) at cohort level is sufficient for the purposes of this work. We, therefore, model the movement of employees between cohorts based on the attributes that describe the cohorts. We define the transition probability matrix $P(t) \in [0,1]^{n\times n}$ by letting $p_{i,j}(t)$ be the probability that an employee moves from cohort i at time t to cohort j at time $t+1$. We additionally assume time-homogeneous transition probabilities, i.e., $P(t) \equiv P$. Under these assumptions, the rows of the random matrix M_t follow a multinomial distribution, i.e., for $i = 1, \ldots, n$, $(M_{i,1,t}, \ldots, M_{i,n,t}) \sim \text{Mult}\,(X_{i,t}, P_i)$, where P_i denotes the i-th row of P.

The transition probability matrix P can be estimated from data that takes record of the cohorts of individual employees over a time period $t = 1, \ldots, T$.

[6] For a model with $n = 30$ cohorts and $X_{\max} = 100$ maximum employees per cohort, the number of transitions in the Markov chain is $|\mathcal{S} \times \mathcal{S}| = \prod_{i=1}^{n}(S_{\max}+1)^2 \approx 10^{120}$.

Let $m_{i,j,t}$ denote the number of employees that are in cohort i at time $t-1$ and in cohort j at time t. Then the maximum likelihood estimator of $p_{i,j}$ is

$$\hat{p}_{i,j} = \frac{\sum_{t=1}^{T} m_{i,j,t}}{\sum_{t=1}^{T} X_{i,t}}. \tag{11}$$

For any time step t and action A_t, the dynamics of the workforce over time can now be simulated by sampling the movement matrix M_t from the multinomial distribution described above and computing

$$X_{t+1} = \mathbb{1}^n M_t + A_t. \tag{12}$$

4 Experimental Setup

This section details the experimental setup, which was designed to answer the following research questions:

1. How does the proposed simulation-optimization approach perform,
   (a) on an operational workforce objective?
   (b) on a strategic workforce objective?
   (c) for a varying employee mobility?
2. Are firing constraints best implemented with a masked policy or an updated objective (penalty for illegal fires)?

We compare the results on a baseline based on linear programming (LP) proposed recently [6]. We evaluate these approaches in two cases. The first is a synthetic organization, and the second is a real-life use case from an international bank to validate the results in practice. We first describe the overall setup, then the baseline, detail the organizations, and include implementation details.[7]

To investigate research question 1a and 1b, we train a reinforcement learning agent for both the operational and strategic tasks in Equation (9) and Equation (8). We evaluate both the trained agent and the heuristic baseline described in Section 4.1 and compare the performance based on the average reward metric, in the manner as described in Section 4.2.

4.1 Baseline

We devise a baseline based on linear programming to compare the performance of the proposed simulation-optimization approach. This baseline was proposed in [6] and makes a number of additional assumptions that allow for efficient solving of the SWP problem. We describe this baseline in detail in this section.

Due to the size of the state space of the Markov chain that describes the workforce dynamics, this stochastic model cannot be used directly with a linear solver. Therefore, we consider a deterministic approximation of Equation (12), by

[7] Code and data for hypothetical use case available at https://github.com/ysmit933/swp-with-drl-release. Real-life use case data will be made available upon request.

replacing the random variables involved with their expectation. This operation, known as mean-field approximation, is justified for large-scale organizations as a result of the functional law of large numbers; see, e.g., [6]. For Equation (12) we obtain

$$X_{t+1} \approx \mathbb{E}\left[\mathbb{1}M_t + A_t\right] = X_t P_{\cdot,i} + A_t, \tag{13}$$

where $P_{\cdot,i}$ denotes the i-th column of the transition probability matrix P. Additionally, we optimize for one time step at a time instead of the whole trajectory $t = 0, \ldots, T$. This is reasonable when the rewards do not depend on time and are given at each time step. In that case, there are no situations where it is required to sacrifice short-term gains for long-term profit.

Consider the target level reward defined in Equation (9) and assume for simplicity that $X_i^* > 0$ for all $i = 1, \ldots, n$. Under the aforementioned assumptions, this version of the SWP problem is given by: find

$$A_t^* = \arg\max_{A_t} \frac{1}{n} \sum_{i=1}^{n} \exp\left(\frac{-\alpha(X_{i,t+1} - X_i^*)^2}{(X_i^*)^2}\right), \tag{14}$$

such that $X_{i,t+1} = \sum_{j=1}^{n} p_{ji} X_{j,t} + A_{i,t}$ for $t = 1, \ldots, T$. Substituting the latter expression in the former, and by noting that each term of the sum is maximized when the term in the exponential is equal to zero, we see that $A_{i,t}^* = X_i^* - \sum_{j=1}^{n} p_{ji} X_{j,t}$. The optimal continuous actions are then mapped to the discrete set of possible hiring options $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. Hence, the decision rule becomes

$$A_{i,t}^* = \mathit{\Pi}_{\mathcal{A}_i}\left(X_i^* - \sum_{j=1}^{n} p_{ji} X_{j,t}\right), \tag{15}$$

where $\mathit{\Pi}_{\mathcal{A}_i}(a) := \arg\min_{a' \in \mathcal{A}_i} |a - a'|$.

To develop a heuristic for the combined reward function, we make the same assumptions as for the target level heuristic. We then consider the state that yields the highest immediate reward, given by $X^* = \arg\max_{x \in \mathcal{S}} r(x)$. We use this as a target level to aim for by applying the target level heuristic Equation (15).

4.2 Training Setup

The reinforcement learning agent is trained for a maximum number of training steps T^{max} specified by the user. At the start of each episode, a random starting state in the neighborhood of X_0 is generated to ensure sufficient exploration of all relevant parts of the state space. This is done by uniformly sampling a state from the interval $[(1-\beta)X_i(0)/X_i^{max}, (1+\beta)X_i, 0/X_i^{max}]$, where $\beta \in [0, 1]$ determines the random spread across the state space. The episode ends after T time steps, at which point the environment resets to a new random starting state. After each T^{eval} number of time steps, the agent is evaluated on an evaluation environment, which is identical to the training environment except for a deterministic start at X_0 and the best performing agent is stored.

When the training process has terminated, we test the trained model on the evaluation environment with a fixed starting state (as a default for 1,000 episodes) and collect several metrics to assess the quality of the model. In particular, during an episode of T time steps, we collect the average reward $\frac{1}{T}\sum_{t=1}^{T} r(X_t)$ and the number of constraint violations $\sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{1}_{\{A_i(t) \text{ is illegal}\}}$.

4.3 Hypothetical Organization

For the hypothetical organization, we consider a model with four cohorts, labeled by M1, M2, C1, and C2 (two cohorts of managers and two cohorts of contributors). We suppose the probability transition matrix is given by

$$P = \begin{pmatrix} 0.98 & 0 & 0 & 0 \\ 0.01 & 0.93 & 0 & 0 \\ 0 & 0.04 & 0.92 & 0.005 \\ 0 & 0.01 & 0.01 & 0.96 \end{pmatrix},$$

and we let $X_0 = (20, 50, 100, 300)$ be the starting state. The hiring options are set to $\mathcal{A}_1 = \{-2, -1, 0, 1, 2\}$, $\mathcal{A}_2 = \{-5, -1, 0, 1, 5\}$, $\mathcal{A}_3 = \{-10, -2, 0, 2, 10\}$, and $\mathcal{A}_4 = \{-25, -5, 0, 5, 25\}$. The maximum cohort sizes are $X^{max} = 2X_0$, the random starting state percentage is $\beta = 0.25$, the time horizon is $T = 60$, and the salary costs are set to $C^{sal} = (10000, 6000, 4000, 2000)$. The target level objective is $X^* = X_0$ (and remain at the same levels as the starting state), with a default precision of $\alpha = 10$. The combined reward parameters are given by $\ell = 0.75X_0$, $u = 1.25X_0$, $G_{soc} = 7$, $\ell_{soc} = 0.9$.

On top of the transitions introduced before, we vary the employee mobility in order to answer research question 1c. In order to answer this question, we evaluate the approach on transition matrices

$$P_\ell = \begin{pmatrix} 1-\ell & 0 & 0 & 0 \\ \ell/2 & 1-\ell & 0 & 0 \\ 0 & \ell/2 & 1-\ell & 0 \\ 0 & 0 & \ell/2 & 1-\ell \end{pmatrix},$$

for mobility rates $\ell \in \{0, 0.01, \ldots, 0.1\}$. For each of these environments, and for both the operational and strategic tasks, a reinforcement learning agent is trained and evaluated. Next, its performance, based on the average reward obtained, is compared to the heuristic baseline.

4.4 Real-life use case

To investigate the performance of our solution method on a real-life use case, we use the following model based on actual headcounts in one particular department of the Bank. The 14,105 employees of this segment of the organization are divided into cohorts based on manager status (manager or contributor) and based on five

Fig. 2. Normalized cumulative training rewards.

distinct job levels, resulting in a cohort model consisting of $n = 10$ cohorts. We label the cohorts as Manager-1, ..., Manager-5, Contributor-1, ..., Contributor-5. The transition probabilities between these cohorts are estimated based on monthly employee data for a period of 48 months. For both tasks, starting state X_0 is set to the workforce at the beginning of the period and target state X^* to the workforce at the end of the period for the operational goal.

For the strategic goal, we use cohort bounds $\ell_i = 0.75X_i(0)$ and $u_i = 1.25X_i(0)$, and the goal for span of control is $G_{\text{soc}} = 7$, with $\ell_{\text{soc}} = 0.9$. Costs associated with salary and management initiated hires and leavers were set in collaboration with an expert in the organization. The hiring options were chosen based on the cohort sizes and include the option to hire or fire zero, a few, many, or a moderate number of employees. The maximum number of employees that could be hired or fired was roughly ten percent of the starting cohort size. For example, the hiring options for cohort Manager-1 were given by the set $\mathcal{A}_1 = \{-25, -5, -1, 0, 1, 5, 25\}$.

To investigate research question 2, we implement three methods to constrain the choices for management-initiated leavers. The first method is a masked policy, for which the illegal actions are removed from the action space by setting the corresponding action probabilities to zero. For the second method, the agent receives a large negative reward for selecting an illegal action. Finally, we constrain the agent to hires only, i.e. in which all leaving employees do so organically. We then train reinforcement learning agents for both the operational and strategic tasks and compare the performance of the unconstrained agent, the masked agent, the penalty-receiving agent, the no-fire agent, and the baseline heuristic.

5 Results

In this section, we look at all results associated with research questions 1a-2 presented in the previous section. We first look at the convergence of the proposed approach in Figure 2 and find that the proposed SO approach converges quickly. Next, we compare the resulting policies with an LP baseline on a test

Table 1. Average normalized cumulative rewards and 95% confidence interval for both tasks on both organizations. **Bold** denotes significant best per task ($p = 0.99$).

	Synthetic		Real-life	
	Operational	Strategic	Operational	Strategic
LP	$\mathbf{0.98 \pm 0.030}$	0.41 ± 0.374	$\mathbf{0.99 \pm 0.010}$	0.12 ± 0.106
SO (ours)	0.94 ± 0.033	$\mathbf{0.83 \pm 0.213}$	0.92 ± 0.015	$\mathbf{0.98 \pm 0.026}$

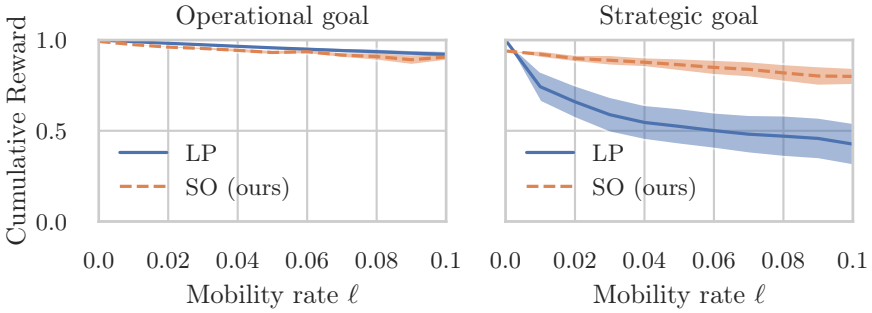

Fig. 3. Normalized cumulative rewards for varying mobility rates.

set. Table 1 shows that the proposed approach performs close to the optimum of the LP baseline on the operational objective and significantly outperforms the baseline on the strategic objective. We move on to research question 1c by looking at the effect of increasing the employee mobility in Figure 3. It shows that the the proposed SO approach is robust against a wide range of mobility levels and that its benefits increase with increasing workforce mobility. The proposed approach shows to be more robust to the stochastic nature of SWP for this nonlinear optimization objective than the LP baseline.

Finally, we compare our approach in a setting with constraints on the organization's control of leavers in Tables 2. Here, we find that we can effectively take the organization's constraints into account using either masking, with a negative reward (penalty) or by only including hires in the action space. Out of these, the 'only hires' variant yields the best results with respect to reward and constraint adherence, with rewards close to its unconstrained counterparts without any constraint violations.

6 Discussion

In this work, we have presented a simulation-optimization approach to strategic workforce planning. The approach optimizes workforce decisions with DRL by interacting with a simulator. Any suitable simulator can be used because the optimization step does not depend on its internals. We propose to use a Markov chain simulator *learned* from historical data. By doing so, the full loop only requires a data set of historical workforce compositions and the organization's

Table 2. Average normalized cumulative rewards and constraint violations (% of total decisions), with 95% confidence intervals. **Bold** denotes significant best ($p = 0.99$).

	Operational		Strategic	
	Reward	# Violations (%)	Reward	# Violations (%)
LP	$\mathbf{0.99 \pm 0.010}$	16.83 ± 3.05	0.12 ± 0.106	13.84 ± 2.96
Unconstrained	0.92 ± 0.015	6.94 ± 1.56	$\mathbf{0.98 \pm 0.026}$	21.72 ± 4.39
Masked	0.74 ± 0.030	$\mathbf{0.00 \pm 0.00}$	0.75 ± 0.135	$\mathbf{0.00 \pm 0.00}$
Penalty	0.87 ± 0.046	0.34 ± 0.07	0.74 ± 0.060	$\mathbf{0.00 \pm 0.00}$
Only hires	0.79 ± 0.041	$\mathbf{0.00 \pm 0.00}$	0.94 ± 0.061	$\mathbf{0.00 \pm 0.00}$

objective as inputs. These objectives may be composed of arbitrary workforce metrics of interest that may be non-linear in the workforce composition. The approach optimizes these objectives *directly*, so that the resulting policy can easily be used to ensure a high impact of the SWP efforts.

We have evaluated the proposed approach on a synthetic and a real-world organization and found that it converges quickly. More so, we compared the quality of the obtained policy to a baseline from the literature. In this comparison, we first targeted an objective composed of workforce metrics. Such objectives are easy to define and accurately reflect the organization's strategic goals. We found that our approach significantly outperforms the baseline on this *strategic* objective and that the difference grows as mobility of the workforce increases. We secondly targeted an *operational* goal, in which the optimal workforce composition is known up-front. Such goals are easy to optimize for with established optimization approaches but hard to define in practice. Our approach performed close to the baseline in this setting. We additionally showed how the approach can take into account realistic constraints by limiting the ability of the organization to control leavers in the organization and found that removing the ability to do so has a very limited impact on overall performance.

We have shown that the proposed simulation-optimization approach is suitable for SWP. Additionally, it opens up various avenues for future work. Firstly, the approach is capable of optimizing for strategic objectives composed of arbitrary workforce metrics. It would be interesting to extend the approach with multi-objective reinforcement learning in order to compute a set of Pareto optimal policies [17]. This will increase the organization's understanding of the trade-offs involved and allow them to fine-tune their strategy. Secondly, the approach currently finds a policy that is optimal on average. While this is suitable for many use-cases, there may be some organizations that prefer a probabilistic guarantee on the minimum number of employees to, e.g., meet service level agreements. Here, risk-sensitive DRL can be employed instead of regular DRL [8]. Additionally, organizational constraints can be formalized and used within approaches that guarantee safety of the resulting policy [12]. We believe that, with the proposed approach, these challenging and interesting research directions that will further increase the impact of SWP have become feasible in practice.

References

1. April, J., Better, M., Glover, F.W., Kelly, J.P., Kochenberger, G.A.: Ensuring workforce readiness with optforce (2013), unpublished manuscript retrieved from opttek.com
2. Banyai, T., Landschutzer, C., Banyai, A.: Markov-chain simulation-based analysis of human resource structure: How staff deployment and staffing affect sustainable human resource strategy. Sustainability **10**(10) (2018)
3. Bhulai, S., Koole, G., Pot, A.: Simple methods for shift scheduling in multiskill call centers. Manufacturing & Service Operations Management **10**(3), 411–420 (2008)
4. Burke, E.K., De Causmaecker, P., Berghe, G.V., Van Landeghem, H.: The state of the art of nurse rostering. Journal of scheduling **7**(6), 441–499 (2004)
5. Cotten, A.: Seven steps of effective workforce planning. IBM Center for the Business of Government (2007)
6. Davis, M., Lu, Y., Sharma, M., Squillante, M., Zhang, B.: Stochastic optimization models for workforce planning, operations, and risk management. Service Science **10**(1), 40–57 (2018)
7. De Feyter, T., Guerry, M., et al.: Optimizing cost-effectiveness in a stochastic markov manpower planning system under control by recruitment. Annals of Operations Research **253**(1), 117–131 (2017)
8. Fei, Y., Yang, Z., Chen, Y., Wang, Z., Xie, Q.: Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. Advances in Neural Information Processing Systems **33**, 22384–22395 (2020)
9. Gaimon, C., Thompson, G.: A distributed parameter cohort personnel planning model that uses cross-sectional data. Management Science **30**(6), 750–764 (1984)
10. Grinold, R., Stanford, R.: Optimal control of a graded manpower system. Management Science **20**(8), 1201–1216 (1974)
11. Heger, J., Voss, T.: Dynamically changing sequencing rules with reinforcement learning in a job shop system with stochastic influences. In: 2020 Winter Simulation Conference (WSC). pp. 1608–1618 (2020)
12. den Hengst, F., François-Lavet, V., Hoogendoorn, M., van Harmelen, F.: Planning for potential: efficient safe reinforcement learning. Machine Learning pp. 1–20 (2022)
13. Jaillet, P., Loke, G.G., Sim, M.: Strategic workforce planning under uncertainty. Operations Research (2021)
14. Jnitova, V., Elsawah, S., Ryan, M.: Review of simulation models in military workforce planning and management context. The Journal of Defense Modeling and Simulation **14**(4), 447–463 (2017)
15. Kant, J.D., Ballot, G., Goudet, O.: Worksim: An agent-based model of labor markets. Journal of Artificial Societies and Social Simulation **23**(4), 4 (2020)
16. Rao, P.P.: A dynamic programming approach to determine optimal manpower recruitment policies. Journal of the Operational Research Society **41**(10), 983–988 (1990)
17. Roijers, D.M., Vamplew, P., Whiteson, S., Dazeley, R.: A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research **48**, 67–113 (2013)
18. Romer, P.: Human capital and growth: Theory and evidence (1989)
19. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
20. Sing, C., Love, P., Tam, C.: Stock-flow model for forecasting labor supply. Journal of Construction Engineering and Management **138**(6), 707–715 (2012)