Deep Science Publishing
https://doi.org/10.70593/978-93-49910-35-5_10

DeepScience
Open Access Books

# Chapter 10: Real-time scheduling optimization using algorithmic planning and workforce analytics

## 10.1. Introduction

Most complex planning and scheduling problems are defined by a single structure that encapsulates all essential elements, such as resources, timelines, and tasks. The relationships within this structure encode the constraints and costs associated with achieving particular goals. However, in certain domains, “real-time” is a key aspect of a planning or scheduling problem. In this case, task scheduling choices must be made frequently and rapidly in ways that minimize execution costs and the impact of the resulting schedule on future execution. Furthermore, real-time scheduling often takes place in environments characterized by high uncertainty, with both incomplete or imperfect information and rapidly changing dynamics. In combination, these two factors make it impossible to model real-time planning and scheduling problems within a fixed and fully-specified structure. For a team of human or robotic agents executing a real-time operation, the structure itself is typically the result of collaborative execution, continually evolving through both the actions of individual agents and their dynamic responses to the actions of others. Little, if any, pre-specified scheduling infrastructure actively constrains their interactions. Moreover, real-time task allocations yield not only schedule plans but also compelling social cues governing the task-specific behavior of co-actors. For a group of humans, these cueing effects can cause unexpected and unwanted changes in schedule-enforced task performance, as a result of behavioral mores and expectations (Boudreau & Ramstad, 2007; Hung et al., 2011; Pinedo, 2016).

Somewhat unusually for planning and scheduling systems, this well-known problem in execution involves inherently social, co-actor-centric behavioral considerations. Our hosts provide two so-called capabilities that we position to ameliorate the limitations of predictable yet emergent schedules in human-robot and human-human mixed initiatives. The first allows robots to collaboratively influence each other and their human partners to simultaneously optimize performance-driven planning-and-scheduling objectives and co-actor-centric behavioral considerations. The second characterizes the unique,

collaborative, co-actor centered qualia at the heart of social planning-and-scheduling and provides feedback to both co-actors and planners (Zhang et al., 2011; Hung et al., 2011; Pinedo, 2016).

## 10.2. Background and Literature Review

The idea of managing work processes through the utilization of a schedule has always attracted researchers and business professionals. The schedule encompassing key activities of an enterprise performed over time and resource assignments to these activities shapes the execution of strategy and achieves the desired business objectives. Traditionally, business objectives support well-defined business needs mapped into scheduling goals allowing planners to use tried and tested methodologies and tools. Now, with the rapid growth of transition economies and the knowledge economy where uncertainty and fluctuating business needs are the name of the game, classical approaches to process scheduling become inefficient and fragile. More and more business schedules are not composed once and left unaltered until execution completion. Instead, they are composed quickly, modified regularly, and altered on the fly. In these agile enterprises, strategic efficiency is accomplished through tighter alignment of corporate, operative, and execution strategies. Accordingly, scheduling objectives including target cost, time, and resource utilization are adjusted daily, weekly, or monthly based on organization performance evaluations and forecast data that continually update the business activity schedule.

Cybernetic theories and control engineering lead to automated process control where deterministic systems with known objects, common substations, and clear operation procedures can be made to perform according to plan or to guarantee the best achievable production result satisfying physical as well as operational constraints. The same principles applied to macro-production planning, scheduling, and dispatching short- and long-term decision problems developed for several decades before the IT revolution fueled the growth of modern optimization frameworks and scheduling software. Meanwhile, organizational behavior theories, flow analysis, and reengineering principles outlined a different approach to business scheduling that emphasizes its ad hoc nature. It advocates modifying the production sequence to respond to market dynamics and imbalances enabled by the characteristics of the resource deployed.

## 10.3. Algorithmic Planning Techniques

Algorithmic planning is an active research area in Artificial Intelligence that focuses on how to effectively generate action sequences to navigate agents through predefined states in stochastic environments. They have been applied in a variety of applications,

such as Robotics, Graphics, Web-service composition, Document planning, Sensor Networks, Army logistics, Game playing, Manufacturing Industry logistics, Counter-terrorism, among many others. One of the main distinctions regarding the types of planning algorithms available is that some planning methods are designed to build solutions for a known deterministic environment, such as Manipulators planning. Some other planning algorithms build solutions for partially known stochastic environments.

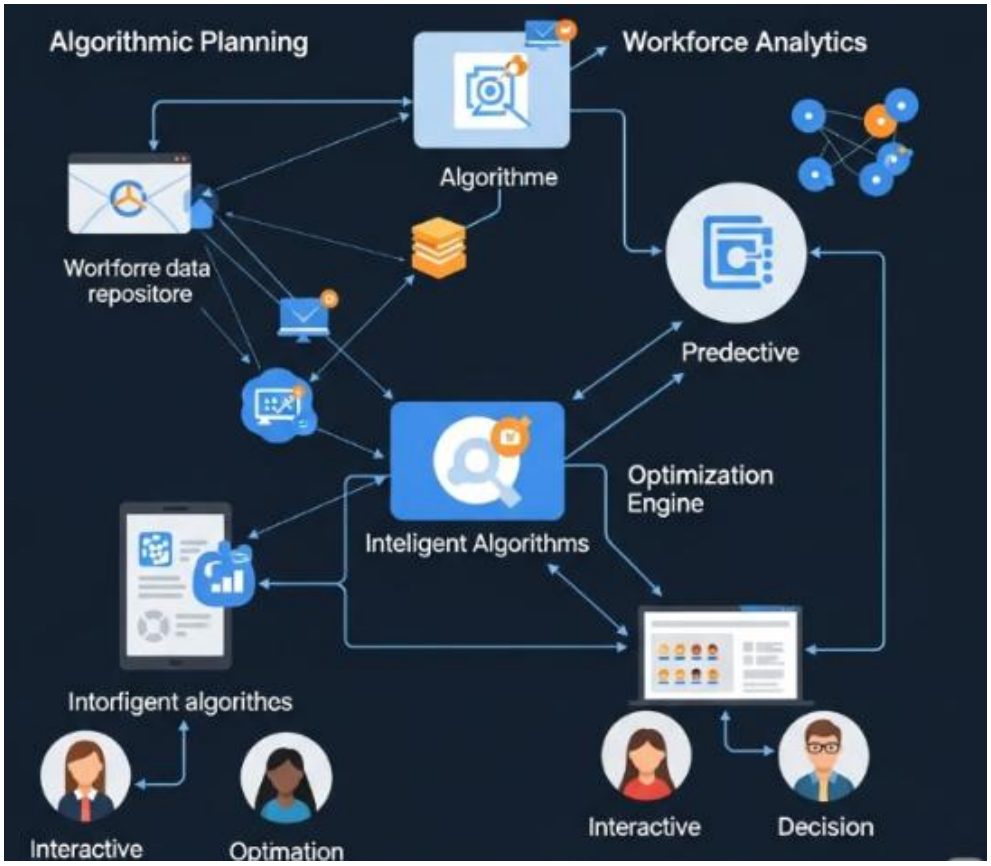


**Fig 10.1:** Multi-Project Staff Scheduling Optimization

The algorithms dedicated to traditional planning problems are usually implemented and applied over preplanned maps that are commonly tagged with positive and negative features to allow the effect of the actions to be taken into account. Relatively few approaches are being offered to the case of the most general i.e. the Optimal Real-Time Strategic Planning problem, which requires constructing an Action Sequence to be executed in-Loco, in Real-Time and according to the observations made with regard to the dynamic environment, i.e. where uncertainty and a possibility of changing the specific characteristics at the time the plan is executed, both by the planner and by other external agents, are at their highest. Due to this, Global Replanning is becoming a

necessary component of any real intelligence system, where the environment can evolve dynamically as the agent is navigating through it.

### 10.3.1. Overview of Algorithmic Planning

For many years, research in AI has focused on the problem of planning, which is the generation of a sequence of actions so that a desired transformation of the world occurs. Although much of the early research in AI did not concern the practical application of its results, planning has developed as a central area in which researchers have attacked fundamental problems that are essential for an understanding of intelligent behavior and in which useful systems have been constructed directed toward important applied problems. As an example of the applied direction, the generation of trajectories for autonomous vehicles is a relatively recent practical problem that has been guided extensively by research in AI planning.

In this chapter, we describe the advantages and disadvantages of the various types of planning approaches, which range from partial descriptions of problems by planning in very restricted domains to fully declarative or general-purpose planning systems. A planning system can also be described by a complex of several distinct dimensions. One of the important ones is the capability of the system with respect to the amount of the problem specification that it requires. On one extreme, a system may accept only a goal description, that is, a logical statement, such as the robot should pick up the cup, that represents the desired properties of the final situation. For example, we could define planning as the effective computation of some form of best plan to solve a problem. The parameters specifying best in planning could potentially be in different formats depending on whether the planner is optimal in terms of computation space, the shortest sequence of individual proposed operations, or some other attribute that measures the quality of the overall plan.

### 10.3.2. Types of Planning Algorithms

Algorithmic Planning is an ambitious and mature subfield of Artificial Intelligence focusing on autonomous generation of plans, both in the abstract and in the concrete. Abstract generation is concerned with selecting the right actions to achieve a sequence of goals; concrete generation is concerned with selecting resources enhanced with time points, so the resulting sequence of actions becomes executable and conforms throughout its execution to some usually dynamic features of the world. Various styles of plans have been considered. Activities, arrangements, timelines, recipes and procedures are only some of them. In this section, we first mention the most common styles of plans and their most basic properties according to which the associated planning

algorithms can be classified. Then, we classify the common simplest uncovered in the literature. Finally, we review, with a focus on the most important algorithmic planners: STRIPS, HTN Planning, Partial-Order Planning and the algorithm supporting the logics of action of the Event Calculus. Planning heuristics are a crucial component of current state-of-the-art planners. Since the invention of planning, people have attempted to devise heuristics that would allow a quick pruning of the search space while being as non-committal as possible. Informative heuristics are informative when they can cut their way down to solutions that should be checked with brute-force methods. In planning heuristics, partial heuristic functions allow partial pruning, which can cut much faster than full, informative heuristics. Partial reusable heuristics allow much tighter and accurate pruning than previously devised such partial functions, and current paradigms of information-based search utilize them efficiently, exploding the reachable portion of the search space only when this is absolutely necessary to find and check solutions.

### 10.3.3. Comparison of Planning Techniques

There is a considerable amount of work done in the area of AI planning. However, most of the research has focused on optimizing the actual plan (or schedule) once an ordering of the steps has been defined. The algorithms used for deciding the step ordering range from very simplistic heuristic best-first methods which ignore the temporal constraints imposed by the partially-ordered plan, to methods which use advanced constraint satisfaction techniques, such as time-dependent pruning and backtracking. However, to the best of our knowledge, the only technique present in the planning literature which actually searches for the temporal ordering of plan actions is the temporal decomposition algorithm which uses hierarchical task networks. There is also work by using temporal decision diagrams which could be characterized as dynamic programming. What distinguishes these methods is their search techniques and the underlying network representation. Both the decomposition and decision diagram approaches are essentially problem-reduction methods. This explains the higher cost of both approaches, as compared to the SAT-based methods, for larger planning problems. However, SAT-based methods actually build a SAT formula in a still-exponential problem space and can therefore be very large, sometimes larger than backtrack-free instantiations, and they also require sophisticated heuristics to reduce their sizes. On the other hand, both the hierarchical task reduction method used by the temporal decomposition algorithm, and the search observation method used by TFT, are guaranteed to produce plans of minimal length. The first method will return the optimal solution the majority of the time because it efficiently encapsulates the domain knowledge associated with method-based hierarchical task reduction, and the second method does not impose any ordering over the operator instantiations.

## 10.4. Workforce Analytics

The workforce is one of the critical resources for any company. Tasks performed are essential for the company's growth, and optimizing those tasks increases productivity and eventually profits. However, to define what to assign to which employee, with what characteristics is an NP-Hard problem and cannot be solved for large datasets. This is where workforce analytics tries to lend a hand. It collects data from different management systems and combines them to provide a clear view of the flows of staff in and out of the company. Workforce analytics can be used to predict issues such as retention and workload, or to optimize staffing and staffing budget. Understanding the data to be collected by the different operational management systems entrusted with recording the working times, the back-office management, the accounts clearing management, the business planning, or the human resources management is crucial for the final result.

The so-famous data were described by 3 Vs (e.g. Volume, Variety, Velocity) but, for workforce analytics, other Vs set better the scenery. We can call them the 5 Vs of Workforce Analytics: Volume, Variety, Velocity, Veracity, and Value. The first three Vs are also part of the 3 Vs. The Volume represents the big amount of information for every employee (earnings, feedback, development, attendance, resumes, reviews, records, hours worked, and so on). Data gathers from various sources, so-called Variety, from different management systems. The more systems are present, and the more easy information can be shared, the better quality of analysis and prediction we can achieve. The Velocity is related to the timing of data acquisition; the more updated the information are, the more reliable is the analysis.

### 10.4.1. Definition and Importance

Workforce analytics is an evidence-based approach to managing an organization's human capital by using data to provide an objective insight and help in making better workforce-related business decisions. It is meant to determine the drivers of employee performance and productivity and how improvements in those areas could contribute toward the bottom line of organizations. Workforce analytics is closely related to the discipline of people analytics, and both are increasingly popular with organizational decision makers. Despite its rapid growth, workforce analytics is in many ways an emerging science, with few recognized formal standards and practitioners using a variety of methods and approaches to extract value from workforce data.

As organizations become increasingly engaged with measurable performance indicators, they have come to realize the need for more extensive and systematic approaches to human capital management as well. Organizations are increasingly investing to develop

technologies and methods to make workforce-related decision making more evidence-based as a way to provide more justification of their decisions. Such investments are justified, in part, because of the financial exposure related to their workforce decisions. In fact, compensation and benefits expenses for many organizations account for the largest share of their operating budgets. More importantly, the workforce directly affects an organization's ability to execute its business strategy, and poor workforce-related decisions can significantly harm organizational performance. The dynamic and complex challenges organizations face in the 21st century increasingly require such decisions to be evidence-based management, which would ideally be informed by systematic analytics.

### 10.4.2. Data Sources for Workforce Analytics

While classification of workforce analytics is commonly adopted to better describe the data and their uses, this section will primarily focus on different sources of data that are being fed to the algorithms used for each stage in the data flow cycle. Data sources are often called out as the first place to start for successful workforce analytics project. There are both qualitative, subjective sources and more commonly used quantitative sources for workforce analytics. Secondary data are the most common and usually inexpensive source of workforce analytics data. Combined with qualitative sources such as employee focus groups and primary qualitative data collection like employee interviews, a richer perspective can be attained. Primary quantitative data collection such as employee attitude surveys can support the analytics models using both subjective and quantitative sources. Free ad hoc models can be first examined using readily available data from each of the input data groupings to narrow down priorities for new primary data collections. As noted previously, the majority of model input data is likely sourced from the business' data warehouse but expect some interactions with different sub-departments to ascertain the nature of the data's collection process and handling. Of course, as with any workforce analytics model, consideration must be made on the input data used prior and during the model build. In particular, the primary factor assumption of all multivariate regression approaches is that the collection processes and/or units are consistent across the samples otherwise correcting for and/or including proxy indicators as model inputs would be necessary.

### 10.4.3. Key Metrics in Workforce Analytics

In simple terms, metrics are succinct representations of information that help form a judgment about a given situation or space. Similarly, workforce analytics metrics are succinct representations of interwoven, structured, unstructured, and time-varying

datasets that express an understanding of the workforce and the workforce management process, biases, and structure. However, not all metrics are of equal importance; some carry more weight and impact than others. This section describes a selection of important metrics that act as beacons that can guide operations through the complex analytics and decision-making process.

This report divides the relevant metrics into three high-level categories: forward-looking, inside-dash, and eventual impact. Forward-looking metrics help during the scheduling and activity planning process about the make, buy, burn, or use decisions by helping assess the feasibility, appropriately, accurately, not-biased, and efficiently provide and manage all the services for the guest experience. Inside-dash metrics help during light touch and responsive tactical planning and provide decision support during operations by projecting, impactfully and clearly balancing, the supply and demand sides of service delivery. Eventual impact metrics come into play post-activity to accurately assess the net impact of service delivery on core and non-core performance indicators. This division, while very context-sensitive, can help with driving the right action at the right time with the right level of granularity and detail.

Other than this high-level categorization, multiple other decompositions are important to understand the set of metrics. The first is to characterize the metrics as efficiency, utilization, or effectiveness related. While not exhaustive, early estimates point to over 300 real-time metrics that are important in understanding the workforce while describing their inter-relations and some key drivers influencing their structure.

## 10.5. Real-Time Scheduling Challenges

Real-time scheduling of work is extremely challenging; the available information at any given time is only partial, the objectives are often unclear or even conflicting, and both the requirements of the service process and the resource availability change over time. These factors complicate our ability to build effective scheduling systems. Furthermore, decisions that are made in related areas may also complicate the operations scheduling problem. To illustrate this complexity of real-time scheduling systems, we identify and address several challenges associated with the instant scheduling of service work.

Dynamic scheduling of work is especially challenging in today's world of service. Work that is scheduled for execution at one point in time may be interrupted by events that give rise to new, different types of work at later points in time. Such changes in the tasking requirements can occur frequently or infrequently, can be known in advance or can occur with little or no advance warning, can be predictable or unpredictable, can be long term or short term in nature, and can be cyclical or one-time only. Even the service operations of the private sector are not immune to the onslaught of dynamic change.

Organizations are becoming more and more service-oriented, motivated by the knowledge that they are increasingly relying on service functions to differentiate themselves from their competitors for a higher level of service-facilitation activities, thus shifting these organizations from the products to the service industries.

### 10.5.1. Dynamic Work Environments

Dynamic scheduling problems typically arise during the execution stage of dynamically changing systems (systems subject to sudden and unpredictable fluctuations in the state of the system or its environment). Scheduling becomes dynamic whenever the system's state changes at any time after the original schedule is completed: new resources enter the system or current resources leave the system, resource release times change, new tasks are added, etc. In traditional scheduling modes, the entire problem is presented to the scheduler a priori. The construction of the schedule is done prior to task execution using deterministic criteria as primary inputs. Task time estimation and release time, processing time, and delay prediction are primarily deterministic. In a dynamic scheduling mode, once the initial schedule has been completed, the system dynamically changes its state during task execution. Even one or more of the environment parameters affecting task execution times are probabilistic rather than deterministic.

Dynamic scheduling of a system's operation is the on-line modification of a predetermined schedule to reduce total operational cost under an uncertain task environment. The unpredictable execution times of tasks and the stochastic effects of the unequal division of labor within a system during multiple task processing require scheduling adjustments throughout the assignment period. External disturbances throughout the assignment period change the task processing times, invoke idle times or need for reassignments, and thus, should not be allowed to negate the benefits of the initial schedule. On-line reactive scheduling is needed to efficiently accommodate unexpected events without starting from scratch. This need makes task scheduling in dynamic settings fundamentally different from that in static scheduling.

### 10.5.2. Resource Allocation Issues

Real-time scheduling requires a control in the assignment of tasks to available resources. This assignment flexibility is an inherent feature of scheduling problems. In most disciplines, researchers develop algorithms suitable for certain sets of problem characteristics. Then, they equip their algorithms with efficiency-asserting techniques, such as heuristics, dynamic priorities, and/or relaxations, and call it a solution to the corresponding scheduling problem. Such techniques require a well-defined model of the problem. For example, assume an algorithm is available that finds a resource allocation

optimization model while executing a task. Then assigning tasks to resources using such a model frequently encapsulates resource allocation over a short horizon. The dispatching model is dynamic. Neither the problem parameters nor the model is static in real time. Not only changes continuously occur in the set of parameters required in the model, but also decisions made earlier will any longer optimize the relevant objective for the same underlying static problem.

Formulating models that provide good approximate solutions for the resource allocation problem at each decision epoch and that are efficient to solve is a critical and very difficult aspect of the research area real time scheduling and resource allocation framework. Despite the hard technology requirements of these highly constrained real-world production systems, some researchers believe such systems will play an increasing role in business economics. In the past decades, the damping effect of globalization offered one buffer against decision inaccuracies in forecasting demand, demand variability, cost underestimations of decision alternatives, etc. However, as a tsunami revealed, this insulation buffer can easily collapse. An immediate consequence of these new requirements and the possibility enabled by the new technology solutions will be that more and more companies will try to enjoy the advantages of true real time resource allocation. More sophisticated modeling approaches should be developed that allow even small companies to take advantage of these new technology solutions.

### 10.5.3. Unforeseen Disruptions

The occurrence of unforeseen disruptions at work presents a very durable and important problem that plagues both managers and employees, from management to day-to-day functions. With many hostile economic factors at play, some work environments are constantly shaken by sudden and expected changes that hurt their initial strategies on resource constraints. Much attention has been given to the problem of unforeseen disruptions and great effort has been applied developing methods and creating tools capable of dealing with the consequences of unforeseen interrupts. Day-to-day functions of a task scheduling system must include the redesign of the respect due dates, often imposed externally, the change of task share in resource allocation conflicts, the need of automatic generation of new assignments and the need of constant updating of the schedule assigning future task. The disregarding of any one of these requirements, even if a scheduling algorithm were to be applied, would be enough for a violent increase in tardiness – that is, in the difference between completion times of jobs and the due dates – in a task flow which was being neatly monitored by the scheduling tool up to the moment of the task interruption.

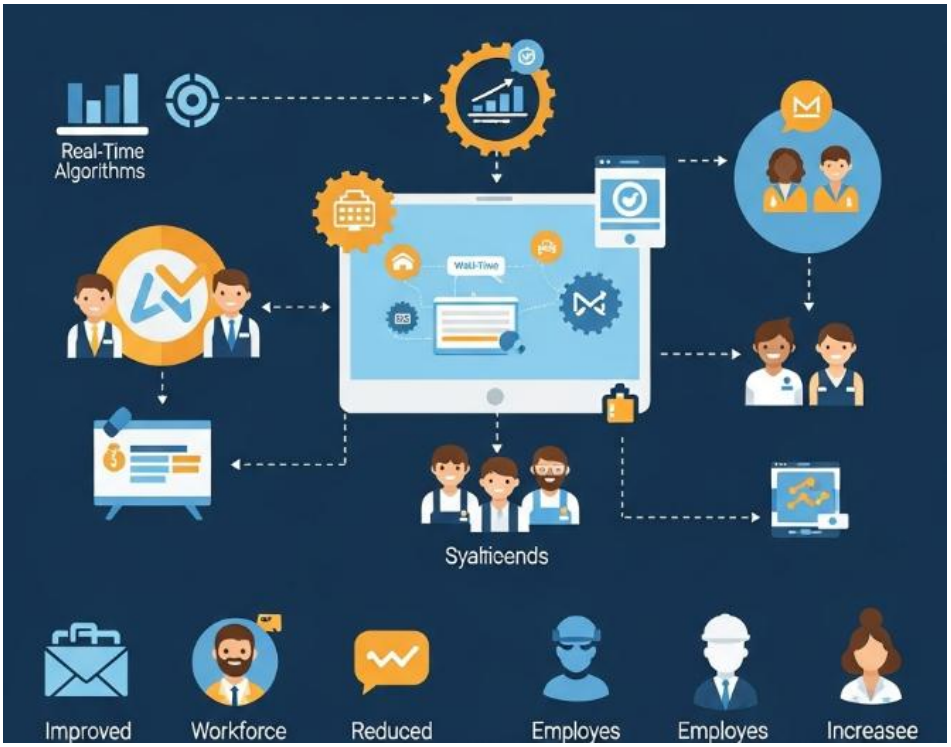


**Fig 10.2:** Optimization-Based Scheduling for the Process Industries

In addition to these difficulties, companies that need to hire temporary workers typically suffer from high rates of gaps in resource needs due to unstable demand. This gap implies that even though these companies announce the need to hire temporary workers, they may not have requests from their clients to provide workers on the day they were hired. Additionally, within the same organization, the workers may be reallocated among different branches that may be in demand. A task that is planned in our optimized task schedule may not be carried out because of the absence of a resource that is crucial for its execution, which means that this task will propagate its delay and possibly cause delays in succeeding tasks until the end of the schedule. This is even worse when the task has a due date, which means that it is important for the client that the job is finished in a timely manner.

## 10.6. Integration of Algorithmic Planning and Workforce Analytics

When we observe the different aspects of planning and analytics, it is impossible not to notice their natural intersection. The optimization methods used in planning are driven to a big extent by the precision of input data such as task priority and work estimates. The most precise and objective estimates are those derived directly from the historical execution metrics generated by analytics, and this is one of the areas where data analytics

has proven of particular utility for algorithmic planners. Examples of this are the demand modeling analytics developed for ridesharing and demand responsive transit applications, or the work estimates used for ambulance service task scheduling. Conversely, the performance prediction functions generated by the best-fit analytics are normally fairly generic, as they are based on data from multiple shifts and sometimes even multiple years. One of the aspects of algorithmic planning that add frequently neglected is its ability to quickly adapt to specific closure rules and plan for inspecific closures similar to the best predictive shift closure solutions. Any workforce planning model that is trained on the input data from only one or a few shifts is bound to overfit, generate solutions that are not generalizable, and significantly underperform during deployment. In addition to being a useful tool for analytics, algorithmic workforce planning can play an important role during the deployment of solutions in the unpredictable demand and short-term closure environments that planners must operate in many times, and not only during major big events but also during day-to-day service.

Algorithmic planning can also take advantage of certain features of some types of workforce analytics. The staffing and optimization schedules of the model estimators that are used in the state-of-the-art shift demand and shift closure models used in some applications of ridesharing and DRT are based on simple staffing requirements and full demand estimators that do not change with time, as there is really no need to do so when the objective is essentially just plug minimization.

### 10.6.1. Synergies Between Planning and Analytics

Workforce analytics produces valuable insights from the analysis of workforce data. Workforce analytics systems — when applied in a targeted way — have become an integral part of the operations of most organizations today. Their increases in complexity reflect not just advances in data analysis, but also the greater volume and varied nature of available workforce data, which is now a key component in a wide variety of workplace decision-making processes. Examples of decisions that can benefit from using analytics include compensation management, talent hiring and identification, promotion and transfer, employee coaching and development, workforce planning, and employee retention.

Algorithmic planning systems are designed to take expected future events into account and derive a plan of action likely to yield good results when those events unfold. The quality of a plan is determined by modeling the structure of the planning problem: the description of resources and activities, with all their parameters and dependencies. Compared to most analytics techniques, the quality of which is determined by the data and the algorithms used to analyze them, the focus in planning is on the quality of the model used in the planning process. In a well-designed model, parameter values can be

kept current without modifying the structures that define a planning problem, and reflect the latest analytical insights, whether historical or emergent. Planning and analytics have diverse but comparable capabilities. There are many aspects of workforce decision-making where analytics may be used to yield improved resources, parameters and structural model components for better and/or faster planning, and where planning may be used to make analytical insights more efficient and holistic. Wherever those synergies exist, both sides may benefit.

### 10.6.2. Framework for Integration

The two platforms, SWAP and TimeSpire, talk to each other to gain decision making synergies through a network of connectors. The technical architecture, as seen from a local planner's desktop or mobile interface, is that from TimeSpire, a Cloud/SaaS Platform, to Web API connectors to SWAP, an on-premises Windows/SQL Studio application. SWAP can dynamically generate planners tasks in TimeSpire at various responsibilities levels, based upon demand. SWAP can use TimeSpire generated work units for work assignment logic during the SWAP optimization processes, then redistribute responsibility and generate shift proposals for specific employees based upon a query of the TimeSpire work units, so that all tasks and their locations are clear and available in TM on PC and mobile for the employees to start their shift. SWAP-AMS integration permits the assignment of request time off, shift bidding and shift exchanging, shift approval and more TimeSpire specific functions, along with all AMS rules and activities (work volume forecast, work templates to condition shifts, restrictions based on workloads, work start and end locations and times and more features to use in the assignment process) automatically from the SWAP desktop.

SWAP can be used during the staffing module of AMS. A SWAP run can provide estimates of the workload corresponding to the work plan function of AMS, needed to develop and fine-tune the current work units for the time period to optimize. In non-optimized mode, SWAP can synchronize AMS-generated work units, and it produces in TM functions transformed work units for upcoming, unoptimized and new time periods. The optimization processes of SWAP can equally orchestrate worker labor capabilities, and accountabilities, requested by the task ownership and requests shown in TM during both the time period to optimize and upcoming time periods.

## 10.7. Case Studies

A common application area for scheduling optimization methods is the manufacturing sector. In this case study, we show two applications from this area. The topics covered address classic issues in supervisory dynamic scheduling and the major operational

differences to heuristics and Artificial Intelligence techniques. The first of these applications concentrated on a PCB assembly line. This job shop was described by an acyclic precedence model that enforced a strict relationship among the different product phases in the stations. We describe a priority-based real-time scheduling approach in which priorities were managed by a generic adaptive Algorithmic Planning model. Recent developments introduced the use of industrial automation networks; therefore efficient approaches are needed in order to transfer information about local statuses and performance indicators to the real-time supervisor and vice versa. One possibility to exchange information without overhead is to use services. However, services support short synchronous requests and short notifications by means of polling techniques. Hereby, the task of how that planning insights could provide needed and structured information in short packets to the local decision is of primary importance. In the second application, the productivity and operational associated problems in unpredictable factories are addressed, for example, the existence of one-star factories.

In a typical job shop, paper-based work instructions may facilitate the execution of work, however, there is no real time information about the local status available. For example, shopfloor operators cannot exchange shop reason or solution type with other operators when in need of assistance. Mobile devices may thus not only facilitate execution of the work processes, but also serve as a tool to exchange knowledge and support among operators to reduce the need for vertical corporate support.

### 10.7.1. Case Study 1: Manufacturing Sector

This case study provides an overview of a project for improving operational performance for a process in a large manufacturing site. The objectives and project scope were defined in discussion with the client engagement team and using inputs from their workforce management system. The project was designed to address the operational constraints and structural issues preventing improvement in adherence and performance, while cascading the objectives to ensure buy-in and accountability at all levels. The daily scheduling of the process had allocations with huge variances from the stated objectives for up to 50 percent of the time. This negatively impacted customer service levels and led to additional costs of expediting shipments. The plan did not take into account set up requirements and down time attributed to maintenance or technology constraints. The production target was derived from forecasted demand, committed plans as also customer requirements. Once the plan was released, the management team effectively abandoned the plan for the next week, leading to low adherence. High turnover, absenteeism and shift fill contract labor requirements impacted base pay costs, led to market interaction for compliance and also suboptimal staffing and sequence optimization for capability. Novel reduction of variability improvement ways were

arrived from the client's production system design and deployment of their common problem solving tool. This led to innovation and engagement with the front line as originally intended from initial deployment objectives.

Implementation of a technology led and algorithmic autonomous planning tool was implemented, on which supervisory and planning teams were trained and skills enhanced. The tool was subsequently configured to derive shift wise and hour wise assignments for optimum utilization across multiple expanding shift structures with set up and skill constraints. The initial tool being optimization centric, had a manual sign off process for the team to execute operations. The dashboard for execution review was subsequently converted for further roll out in an autonomous framework across other processes of the shop. The results of the process and estimated savings for various phases post intervention for the client are detailed below. While the module of hourly utilization for abnormal load and down times of the cleaning tool were triggered on a schedule for shared infrastructure and subsequently planned it was being utilized for reporting only.

### 10.7.2. Case Study 2: Healthcare Sector

As described in Section 5.3.5, the development of a minimal viable product for the healthcare sector was motivated by the need to validate that real-time multi-agent optimization using workforce analytics could be usefully applied to other domains aside from manufacturing. For this purpose, a use case related to the operation of hospital care services was selected where workforce analytics had previously been applied. The development and trial of the system was undertaken in collaboration with an experienced organization that uses sophisticated algorithms to optimize workforce scheduling in emergency departments and other hospital areas. They were particularly interested in extending their hospital product to deal with real-time changes in demand and associated changes in staff allocation requirement. The scheduling system is also used as a ground truth reference by researchers because of its high degree of accuracy for the matching of scheduling decision to demand.

As discussed in Section 6, the trial of the RT-Optia system was conducted to assess the feasibility of applying an agent-based, real-time, adaptive workforce allocation optimization approach to the healthcare sector. The objective was to evaluate limitations with the optia system which would need to be addressed for it to be applied more widely in healthcare before agent-based optimization could be considered a viable alternative to the centralized model-predictive optimization approach. The results from the study, and the advantages and limitations of the different optimization approaches were discussed in Section 7, and provided a strong motivation for the design and development of the RT-Optia system presented in this paper. The discussion and application presented in this Section give valuable insights for the development of workforce scheduling and

allocation solutions not just for the healthcare sector, but also other staff intensive, international operations.

### 10.7.3. Case Study 3: Service Industry

Our final case study is related to the needs for resource allocation in fast food restaurants and general service industry problems such as banks or tourist companies. Most employees in service organizations work shifts. In some work sites, employees can be assigned to either work or be off. In other places, resources are assigned to different work areas. Call and wait lists are often used to manage capacity and time delays. Often, specialized staff is used in varying levels. Staff allocation in some areas of the queue can be improved by planning more barriers and some control over how tasks are performed by customers and assigned automated helpers are used. Their activities may be done more quickly and with less need for specialized information, tools and staff-training.

Education and human resource issues are often used to select employees and their schedules. Such constraints are not very popular in other types of models. Often there is a high drop-out rate and minimum holding penalties. Call and wait lists are often specified to keep openings available. Operations are off-line and assigned each day, week or month. Super-compromise or mock-up calendars can be quickly generated to increase holiday employee satisfaction. Staff availability for those holidays is not generally very flexible. To reduce lower employee turnover, recommendations suggest long duty-slots or more than 5 or 6 consecutive hours. Office closure, day length constraints, daily and maximum work hours rules and overtime-to-regular time ratios are often stated. The models must generally keep multiples of each of these before solving. Staff travel time from and to home can be linked with times when the route is open or closed. Lower turnover is desirable because turnover strategies are usually determined on yearly bases.

Budgetary systems allow only a small amount of schedule manipulation from cycle to cycle. Single or multiple substations may be over- or underscheduled. Staff roles may have to be chosen. The schedule may be available only on short notice. Offices may or may not be open. The travel system has to be closed or opened combining reasons because of budgetary or resource reasons. An inside service group performs shipment services at regulated times or triggers outside suppliers to provide travelers with discounts.

## 10.8. Implementation Strategies

Real-time planning and scheduling systems can take many forms. Cloud hosted SaaS, on-site applications, blended on-site and cloud solutions are several options. Applications can be designed to run on specific devices and may be written for specific types of operating systems. We believe that in most applications that run in browsers, a JavaScript SVG development stack is well suited. There are many libraries that allow users to quickly create desirable user interface components and functional expectations, minimizing the parts of the system that have to be built from scratch. It is worth the effort to search widely for applicable libraries. For both on-site and cloud-hosted systems the security features offered by public clouds minimizes the effort required for successful deployment.

The server side of the application may be written in multiple language/platform combinations, including Linux with C/C++ and Apache, Windows with C/C++ and IIS, Linux/Unix with Python, Windows with .NET, Ruby or PHP, Google Go and Node.js. The choice of server-side languages/platforms should be directed by your development team's technical strengths combined with the needs of your application. Be sure to consider the use of scheduling modules and libraries. These may greatly speed development. Also consider the response time requirements in choosing a server-side implementation. Responding to a user in less than a few seconds is desirable.

We depend on our applications to run reliably with the least amount of downtime. For complex systems this requires planning for hardware redundancy, software failovers and protection against malicious attacks. Configuration management of the development and production version of your code is essential to minimize deployment errors, especially for rapid deployment of code changes. For both cloud-hosted and on-site applications, the need for the application to protect itself against attacks by malicious users has become a critical requirement and these protections must be thoroughly implemented before going into a production environment.

### 10.8.1. Technology Stack Selection

For the large enterprise scale deployment, a key concern is the technology stack selection as the development and execution of the optimization model is a non-trivial task. Its implementation requires selection of tools for development, configuration, integration, deployment, execution, and monitoring. The size and structure of the organization and its investment in the technology determine which approach organizations can take. Some of the key considerations are the existing technology stack used to support Business Intelligence, Business Process Management and IT Service Management books of

record, Executive dashboards, and reporting needs. These considerations should support the integration or data sharing with the selection of the optimization solution technology.

The selected optimization solution stack offers trade-offs in terms of performance, development, hosting, deployment, monitoring, execution, and maintenance costs. At one end of this spectrum, users can purchase a highly efficient, preconfigured solution which can integrate directly with the organization's existing reporting solution for design, deployment, monitoring, and execution. On the other end, organizations can choose to develop the optimization using commercial, off-the-shelf heuristic optimization engines integrated with an ETL solution for data feeds and a cloud data warehouse solution for storage of design and response planning artifacts.

### 10.8.2. Change Management Considerations

Despite technology being at the heart of algorithmic planning, building an advanced version of the planning algorithm is the easiest aspect of the planning process to implement, and it often generates impressive results. Change management of the people problems is difficult, time-intensive, and messy, and it begins before there is a new plan, continues during the period of transitioning, and does not finish when the final new plan is executed. Instead, it is part of the fabric of the new organization, with a new culture, business processes, and new planning rules. It is complicated because many people have roles at multiple layers of the organization, and one group may need to be treated differently from another acting in an interconnected way. All members communicating openly will get the most trust among others. Yet, that is not enough. Members may distrust communications early in the process or talk among themselves but not with all members. Both official messaging and informal cultural bulletins are important. Simply asking each layer to write its own viewpoint on the other layers' changes, both good and bad, can generate a lot of useful material to probe with focus groups. Change management techniques developed in large organizational changes may provide useful techniques. The reason for teamwork on the algorithms and communication to all organizational layers is that large changes in responsibility need very large gains in performance. Forced layoffs can generate gains for any set of rational people approving them. A proper worrying question is finding the objective function for measuring this target, and if losses are to be temporary sustaining underlying trust levels in good times. Even if workers get the message, they may not understand the timing of the changes. Results and irrational actions may cause drug use or other behavioral problems.

## 10.9. Evaluation Metrics for Scheduling Optimization

Optimization in the field of scheduling is often a dense and complex territory, densely populated with numerous methods, models, implementations, and an equal number of contradiction statements. Although there may be no clear "truths" or achievable goals, there is a consensus that the "pitfall" that needs to be avoided is the ill-posed and ambiguous optimization problems. Therefore, it is a common best practice to use a "least" negatively conflicting set of Performance Indicators (PIs) for evaluating and rewarding a variety of diverse scheduling objectives. Such a least negatively conflicting set of PIs may cover the total "penalties" for tardiness or exceeds for order items or production volume, which are either associated with stakeholders or directly reflected to costs; the rejected orders due to insufficient manpower; the under-utilized work times; and excessive work times (real or forecasted).

It is important to note that a Decision Maker (DM) might define the best schedule either based purely on earlier negative cost predictions or on the realistic post-evaluation output; and utilize either of them as a rough estimate for the cost-benefit analysis. This, however, can lead to different solutions, and ultimately verification of what would have been the "best" schedule post comforting reasons. Additionally, while artificially constructed for proof of concept studies, prediction models must be used on real scheduling cases achieving realistic performance scores. Such scores have generally been achieved utilizing Advanced Planning Systems (APSs), as well as real-time optimizing Scheduling Systems (SS). The predictive quality of the criteria used for the cost-benefit analysis moreover supports the argument for post evaluation; or pre-evaluation identification of the "best" schedule.

### 10.9.1. Performance Indicators

You might use different performance indicators to evaluate scheduling optimization from different perspectives, such as economic beneficence, service quality, and business stability. Among various economic indicators, profit and revenue are two primary dimensions to reflect the economic performance of a schedule. Viewed from the customer’s perspective, responsiveness relates to the service time, that is, waiting or delay time. Responsiveness and its associated indicators, including average completion time, maximum completion time, or the probability of completing orders within a specific window, could be used to measure the service level of a schedule. Under some applications, high labor utilization is misleading since it may cause much higher service time and indicators could not reflect how many resources are working or when they work. Under such circumstances, total active time of all resources and the number of resources working at the same time can be used to evaluate if the schedule is proper for

stability. In addition, tardiness or makespan is usually selected to evaluate stability since the tardiness of a schedule relates to the reliability of meeting service requirements.

Some studies recently established several scheduling optimization evaluation frameworks. For example, a multi-criteria framework from two classes of indicators: business performance indicators and service level indicator. You can also decompose economic beneficence and service quality from a triangular model. The stability-related indicators are from a multi-objective scheduled stability issue and its evaluation standards. As a summary, this section analyzes optimizing effects on a total of nine indicators: revenue, profit, tardy orders, makespan, average completion time, probability of completing orders in a specific window, total resource active time, maximum active resource number, and tardiness of a schedule. Furthermore, the nine indicators are categorized into four dimensions: economic performance, service quality, business stability, and relaxation results. When applying scheduling optimization methods toward practice, you can use the present outlined indicators.

### 10.9.2. Cost-Benefit Analysis

The costs incurred by organizations due to deviation from ideal conditions are called cost factors. Such costs include, but are not limited to excessive workforce cost, expediting or premium cost, overtime cost, excursion cost, backfilling cost, and production lost cost. Every departure from ideal site condition during task execution has associated expenses. For example, excessive workforce cost is triggered when additional crew show up to replace members of the crew that is suspended for some reason. Such temporary suspension, usually provoked by an unscheduled event like bad weather, must be in accordance with union and company agreements. Likewise, makeup and production lost costs are canceled if a production process is instantly applied for restorative actions after a depletion event. As these events happen in the worksite, an adaptive response to unpredictable changes through dynamic scheduling gets priority for operational management. The timely rescheduling of the tasks concerned by the current site condition is essential to furnishing seasonal results.

The cost factors outlined can be computed based on historical data for different construction projects. The objective is to determine a function of site conditions that assigns a probability distribution function to every project parameter, and a function that assigns values to the cost factors for that project, that are valid for each expectation of the probability distribution function assumed for that project. Expressing the cost function as the sum of the value of the work cost factor and the value of the task cost factors allows estimating a factor for each construction project, which thus becomes site-dependent. This cost-benefit analysis justifies building computing modules capable of estimating the values of the cost factors as functions of the construction site condition.

Data from historical projects provide the average parameters and costs, and corrective factors address deviations from their average value.

## 10.10. Future Trends in Scheduling Optimization

Scheduled optimization has been an industry necessity far before associated problems were formulated from an algorithmic standpoint. It relied on heuristic tools that evolved in search of efficiency and increasingly ever more advanced feedback systems, relying on data collected for every shift that was worked and continued to be fed with information over time and developed on their own increasingly better heuristics. Recently, associated heuristics were encoded in more plan-oriented AI tools developed by researchers exploring the way that people explain what they wish the AI summarization tool were to do and alternative methods to introduce any sort of new requirements or constraints over time to influence the modification of previously settled plans towards better performance. These two tool categories have become competitive both in planning efficiency and in performance-oriented results. They have engaged in building increasingly more efficient methods to solve the associated problem.

Recently, the blend of both, which includes plan-based algorithms using data-driven heuristics has become the dominant methodology in all general scheduling approaches, but in workforce scheduling optimization, their use has become the solution that dominates among ever-implementing optimization methods. Work patterns have shifted due to the effect of office work mobility restrictions. Not only have hybrid models changed the frequency of work near employees' homes, due to the staggering of part of the working time into home offices, but their information hotspots have also changed. Thus, a fractional value of work related to remote and home working, had they been considered, could have probably changed or optimized a different set of scheduling problems, and provided different solutions for this class of issues. The association of traffic models with scheduling has already been researched for location problems, but it seems to have not overcome into real-time scheduling as a whole, and it would be interesting to explore whether this sort of dual model has previously been the basis of real-world applications.

### 10.10.1. AI and Machine Learning Applications

This paper has explored Algorithmic Scheduling and Real-Time Scheduling as a way to help organizations leverage the value of their data. As we look toward the future, we anticipate AI and machine learning will play an increasing role in Real-Time Scheduling, operating in synergy with commercial scheduling systems. We also see Real-Time

Scheduling becoming even more essential due to the impact of digital transformation. What will this look like?

As demand for workforce flexibility has skyrocketed, organizations can no longer afford to confine their scheduling planning and execution to rigid structures that lock in a resource mix too far into the future. Organizations need to accurately predict changes and automatically adjust scheduled workforce assignments and other required resources to address temporary changes. Whether it is hotels adding staff based on a big local event or hospitals creating extra shifts in anticipation of cold and flu season, planning cycles need to be converted to a real-time capability.

Using AI and machine-learning models of employee behavior, organizations can optimize scheduling algorithms, analyzing past employee behavior details to predict self-scheduling request patterns, just-in-time employee shortages, availability of specific employees at specific times, employee cancellation and request compliance habits, or even your teams' favorite pizza toppings. Together, these supplemental and complimentary capabilities will enable organizations to level up to Surge and Adaptive Scheduling. Leveraging these capabilities will enable organizations to more effectively adjust task priority focus and expedite changes in task assignment to meet business objectives. These adaptive capabilities enhance the value of existing tools, amplify organizational agility and ensure that teams leverage all resources available in support of organizational objectives.

### 10.10.2. Impact of Remote Work on Scheduling

COVID-19 precipitated a fundamental upheaval in the mechanics of collaborative work, inciting an unprecedented shift towards remote work. A confluence of distinct factors initiated, accelerated, and sustained this seemingly radical transition towards extensive remote and hybrid work. First, the mandated stop-gap sheltering-in-place lent receptive companies, and their workforces, no alternative but to hastily adopt full-scale remote work initiatives to sustain business operations. Second, by 2021—long after the initial crisis abated—many organizations, newly cognizant of the observable benefits of remote work for talent attraction and retention, transitioned from temporary policies to long-term plans laying the initial groundwork for hybrid work. Finally, accelerated by the pandemic, remote work was being broadly adopted even in sectors traditionally resistant to flexible working modalities—like knowledge work or services—modified to fit the new normal.

Prior to this fundamental shift towards remote and hybrid work, empirical research established that working from home, when autonomously chosen, offers benefits both for employees and their companies. Employees enjoy higher satisfaction, morale, and

engagement; reduced stress; and lower likelihood of burnout. Employers reaping the accompanying productivity benefits are subsequently gifted with higher profits. However, the transition from partially- to fully-remote work—which eliminates chance encounters—and the resulting changes in interpersonal dynamics has confounded predictions. On the one hand, many leaders argue the reduction of serendipitous interactions and informal collaboration that often spur creativity and innovation threatens to undermine trust, cohesion, and work culture. Others predict that, in the medium-term, as normative behavior re-converges towards stronger reliance on collaboration for more complex and cognitively demanding activities that remote work cannot replicate, individuals will ultimately return to tighter in-person schedules. However, prudent predictions sow discordance; in the long-term, however, they foretell dire consequences for innovation from companies that fail to cohere hybrid schedules.

## 10.11. Ethical Considerations

The increasing prominence of AI technology in our daily lives emphasizes the importance of ethical scrutiny. In particular, the use of algorithmic planning, machine learning and optimization for real-time WFO and RTW optimization involves decisions that can substantially impact a large number of employees. Consequently, ethical considerations should be an integral part of their development, deployment and use. In this section, we focus on two aspects of ethical AI: data privacy issues, which are critical for WFO, and potential fairness and bias issues in allocating available and future work to employees or agents, which are universal problems with algorithmic decision support systems.

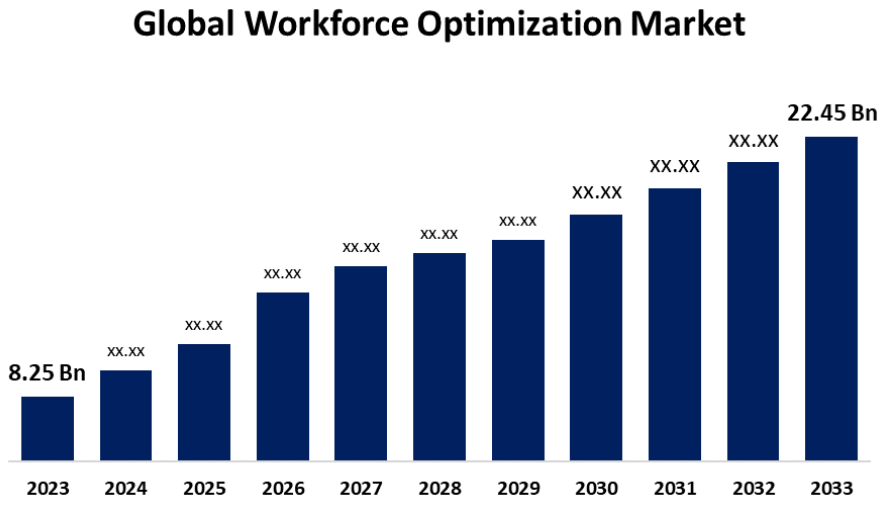


**Fig :** Global Workforce Optimization Market Insights Forecasts

During the creation of a workforce operation or planning capability to support agent scheduling for a specific business, it is inevitable that AI decision support systems will require accessing and probably also storing sensitive data about employee work history, productivity forecasting, unplanned absenteeism, customer service records, and customer satisfaction. Potential consequences of any data privacy breach by either the business that is implementing the capability or the vendor responsible for deploying and maintaining it can be quite serious, including legal action by employees for business violations, as well as loss of reputation. This is a particularly sensitive issue for businesses that also engage in outside surveillance of employee actions. Although machine learning and RTW analysis are heavily dependent on data, predictive algorithms should protect the identity of employees; exposing it to external parties in any way would compromise the ethical motivation underlying the AI technology.

### 10.11.1. Data Privacy Issues

Workforce scheduling processes make use of a large number of datasets containing sensitive personal information such as protected characteristics like race, gender, religion, sexual orientation, ethnic background, health conditions, and political affiliation; contractual rights such as breaks, working hours, holidays, and restrictions in working at odd hours; and psychological needs such as motivation for certain tasks, feelings toward co-workers, and predictability in working hours. Handling such datasets poses important ethical dilemmas for business managers and algorithm developers when optimizing scheduling tasks. On one hand, the aim of business managers and algorithm developers is to optimize an objective function that relies on data as accurate as possible to obtain a schedule that serves the business goal in the best way. On the other hand, these datasets are characterized by privacy concerns by the workforce, which however is also the user of the final product, namely the shifts. Finding the best solution to this dilemma by developing measures that reflect the influence of the sensitive characteristics and guarantee the necessary workforce acceptance of the algorithm is a major challenge. This challenge is particularly hard to tackle when there is little evidence of how the model's decision processes affect certain user groups. Of particular importance in this field is the fairness interpretation. A workforce allocation bias tends to reflect an industrial bias. It frequently happens that a business violates basic civil rights in the area of labor. For these reasons, it is mandatory that the algorithms used in the schedule optimization process are not only ethically acceptable but also feasible and reliable in terms of the achieved outcome. Non-discriminatory algorithms should consider the possible disadvantages for the protected categories to prevent strengthening the existing biases.

### 10.11.2. Fairness in Workforce Allocation

The final work allocation of a workforce may create controversial or perceived unethical conflicts if certain attributes of the workforce or the type of tasks are handled improperly. One has to ponder questions that relate to fairness in the work allocation, such as: Are all members of the workforce treated the same? Do certain members of the workforce receive first preference for certain types of work? Are certain biases present in the work allocation (e.g., requiring certain types of degrees)? Are "special" types of jobs reserved for just a few members, while others have to take on more frequent more "burdened" types of work? Are raters with certain attributes assigned to only a few members of the workforce? Is some other form of bias present in the allocation?

To answer these questions, we investigated the token-based incentivization and use it in a simplified manner: Each member receives the same amount of tokens over time, which are used to bid on different types of jobs. Based on the bidding process, the system can assess which members are interested in which type of jobs. By building a model on top of the bidding information and also the performance data from the tasks already completed, the task allocation system can adapt the balance of burdened to privileged jobs, salience of members with privileged jobs, and types of members obligated to share the burden. In doing so, it can learn and balance the relative preferences, skills, aptitude, stress level, durations, and burstiness.

## 10.12. Conclusion

Real-time scheduling optimization is critical for the U.S. COVID-19 vaccinations effort due to the necessity of always allocating all available doses. Our approach employs an ensemble of proven AI techniques that, when combined, can significantly reduce the run-time of real-time scheduling optimization plans. It does modeled sampling, uses a recently proposed nested lift-and-project algorithm based on sampling to generate a family of cuts that are added to a modified MIP-based proportion optimality algorithm, and uses a novel trajectory-based dynamic agent planning algorithm. We employ MIP constraints to reduce the size of the molded problem that is solved, and we also generate a problem-solving forest to reduce the amount of repeated work that results from its serial dynamic nature of that problem solving. We validate, calibrate, and tune our integrated approach using two sets of real-time data during the early COVID-19 vaccination efforts in the U.S., and we provide comprehensive experimental results using both sets of data.

Motivated by the need for accurate modeling of related decisions, we propose a delay cone constraint and discuss how this constraint can be utilized to reduce the number of trajectory solutions to an optimal problem in a spatial digital agent-based modeling and

simulation. We then present experimental results related to the incorporation of the delay cone constraint in our search algorithm. The results confirm that the use of this constraint significantly reduces search time with small consequences pertaining to the optimality of the results. Finally, we briefly discuss how this constraint can also be used to enhance the capability of other known spatial digital agent-based modeling and simulation systems.

## References


Pinedo, M. L. (2016). Scheduling: Theory, Algorithms, and Systems (5th ed.). Springer.https://doi.org/10.1007/978-1-4614-2361-4

J. Zhang, G., Gao, L., & Shi, Y. (2011). An effective genetic algorithm for flexible job-shop scheduling with fuzzy processing time. Expert Systems with Applications, 38(4), 3563–3573.https://doi.org/10.1016/j.eswa.2010.08.119

Boudreau, J. W., & Ramstad, P. M. (2007). Beyond HR: The New Science of Human Capital. Harvard Business Press.

Hung, Y. F., Chen, H. G., & Wu, J. Y. (2011). A real-time scheduling system for workforce optimization. Information Systems Frontiers, 13(4), 571–584.https://doi.org/10.1007/s10796-009-9205-1