
| RESEARCH ARTICLE**AI-Driven Predictive Workload Management in Contact Centres: A Review of Forecasting, Staffing, Real-Time Optimization, and Human-AI Governance****Ammar Khalid***Independent Researcher, Techstrike Company***Corresponding Author:** Ammar Khalid, **E-mail:** Mohummad.ammar@gmail.com

| ABSTRACT

The contact center has grown from a “service queue” into an AI-driven Omni channel environment where the needs of the customers, the availability of your agents, service quality, automation, and real-time routing all dynamically interact. Predictive workload management is therefore critical as it is instrumental in managing the workload for a good balance of customer waiting time, service level compliance, operational cost, and agent utilization and employee wellbeing. This work is a literature review focused on key areas identified for AI-driven predictive workload management in a contact center—a few of which have particular relevance for this round—including workload forecasting, staffing optimization, queueing-based decision support, deep learning for time series prediction, automation channeling demand deflection, and responsible human-AI governance. The review reveals that classical queueing and workforce management modelling concepts continue to play a vital role in meeting demand forecasts, adapting staffing levels and service levels, but they are becoming increasingly inactive without the incorporation of adaptive forecasting, uncertainty considerations for forecasts and real-time feedback. In recent years, the application of machine learning, recurrent neural networks, attention-based models, and transformer networks has introduced enhanced capabilities to capture the complex and non-linear nature of workloads with multiple horizons and high frequency. But that's not enough - predictions have to be accurate. For AI-driven workload management to take place, the architecture must be integrated to enable data ingestion, forecast demand, optimize staffing, refocus reforecast in the middle of the day, control routing, monitor automation, and make sure that the automated results are explainable and fair. Some key research gaps highlighted by the review are: multichannel forecasting, concept drift, workload redistribution due to the use of a chatbot, agent well-being, privacy-preserving learning, and interpretable prescriptive analytics. It suggests that creating a new generation of workload management for contact centers must be a socio-technical system with the augmentation and enhancement of human workload operational judgment provided by AI.

| KEYWORDS

AI-driven workload management; contact centers; call arrival forecasting; workforce management; queueing theory; deep learning; predictive analytics; prescriptive analytics; explainable AI; human-AI governance

| ARTICLE INFORMATION**ACCEPTED:** 02 August 2026**PUBLISHED:** 18 September 2026**DOI:** 10.32996/fcsai.2026.5.9.23

1. Introduction

Contact centers are one of the most modern and most information volume-intensive, and most operation-sensitive service systems in the digital economy. Through phone calls, live chat, email, social media, messaging platforms, IVR, chatbots, and more recently (with the advent of Generative AI), with the assistance of GenAI, they facilitate communications between organizations and customers. They work on tight matching of customer demand, which is subject to uncertainty, and their service capacity is limited. With an underestimate of demand, queues grow, service levels decrease, customers are less satisfied and customer abandonment rises. If the demand is on the higher side, the organizations have to invest more on staffing and the agents have to take some time as they are not working. This in turn creates the challenge of workload management, and the key to success lies in prediction, staffing, routing and real-time control working together as one system, and not the sum of its individual parts.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (<https://creativecommons.org/licenses/by/4.0/>). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

This results in workload management being a core issue in contact-center operations, where the elements of prediction, staffing, routing and real-time control need to work together, not as separate functions [1], [2].

This fundamental literature deals with the contact centers as stochastic service systems where the arrival process, service times, abandonment, routing rules and number of agents play a role in describing the performance [1]–[6]. A mathematical approach to estimate delay, service level, occupancy and abandonment in varying staffing levels is provided by queueing theory. Classical modelling (including Erlang-based) has value, as it can provide input into operational choices based on workload forecasts. Today's contact centers are very different from the ones for which many traditional models were created, however. These are multichannel, multi-skill, semi-automated, influenced by changing/revolving promotional offerings, outages, seasonal events, macroeconomic conditions and the use of digital self-service. These are the factors that add volatility, which results in the lack of effectiveness of static workforce planning in modern service environments [2, 11, 17]

The key to predictive workload management is predicting workload in the future. An interval-level workload can be simply represented as:

$$W_t = \lambda_t \times E[S_t]$$

Where, W_t is the expected workload at time t , λ_t is the expected number of arrived contacts at time t and $E[S_t]$ is the expected handling time. In reality, this simple equation needs to be expanded to incorporate abandonment probability, channel mix, agent skill group, after call work, shrinkage, automation deflection, routing policies and service level limitations etc. It, actually, does not represent merely a forecasting process, but a prediction-to-decision pipeline in which a forecast of the workload needs to be translated into a workforce schedule, staffing changes made within a day, and operational interventions [7]–[13].

Previous research in call-center forecasting has involved modelling the daily arrival pattern, updating the forecasts within the day, and quantifying the uncertainty of the volume of arrivals [7]–[12]. Such research found that there is significant volatility in demand in contact centers, both hourly, and on various other time scales; daily, weekly and holiday related patterns. They were also able to demonstrate that updating on an intraday basis is important as the initial deviations from the anticipated arrival times can reflect on workload pressure at a later point in the day. Further studies have demonstrated a link between forecasting uncertainty and the making of staffing judgements, with the results indicating that poor forecasting can lead to either service failures or that of unnecessary labor costs [13]–[18]. This relationship between forecast error and the operational cost is particularly salient, as contact-center managers are not interested in forecasts for the sake of forecasting; they are interested in forecasts that help them to work more efficiently, assign their staff more efficiently, route calls more efficiently and deliver the service they desire with more efficiency.

Artificial intelligence (AI) has now added to the list of the methodologically useful tools for workload prediction. Although certain statistical models could capture non-linear relationships, high dimensional inputs, and complex temporal dependencies, these relationships are too complex to represent by statistical models alone and machine learning and deep learning models could capture such relationships too [20, 21, 22, 23]. Multi-horizon time-series forecasting [21]–[27] has demonstrated the good potential of the following types of neural network models: recurrent neural networks (RNNs); probabilistic deep neural networks (PDL); neural basis expansion models (NBE); attention mechanisms; and transformer-based architectures. These approaches are especially relevant to contact centers as there are factors that influence the workload including factors that interact with one another, including customer intent; customer change of channel; marketing events; services disruptions; agent unemployment; automation performance etc. But they have to be assessed based not just on prediction accuracy but also by robustness, interpretability, ease of computation and being compatible to work force management systems.

What's also needed is a turn towards prescriptive analytics as opposed to predictive analytics, with AI-driven workload management. Beyond that, a change from predictive analytics to prescriptive analytics is needed with AI-driven workload management. Predictive models predict what will happen, prescriptive models recommend what will need to happen [28]. In contact centers, that translates to making a shift from forecasts to staff scheduling and break plans, overtime choices, routing adjustments, chats escalation rules, and cross-skilling. While other operational environments can also integrate demand forecasting and resource optimization with their analytics pipeline the tight coupling between service quality, labor cost, employee experience and customer patience highlights some unique challenges for contact centers related to analytics [13, 15, 28, 29, 35, 36].

Another challenge is that the task changed for AI now that it has to predict the changed workload. Routine contacts can be filtered out by chatbots or virtual assistants, or customer's issues can be resolved with agent-assist, automated summarization or generative AI systems, which reduce handling time or enhance first contact resolution. Meanwhile, they could decide to move easier interactions away from human agents so that those who deal with more complex and emotional or escalated cases are those agents. These not only alter the volume of contacts but also the level of workload and the skills and stress level of agents [30, 35]. Future workload models should then not be based on the premise that past human-operated workload patterns will not change due to automation-only workloads.

However, the opportunities aren't adequate, and it creates substantial governance concerns when using AI workloads. Forecasts and staffing recommendations impact how employees are scheduled, workload is balanced, stress is experienced, service is accessed, and the fairness of service. If managers or agents are not comfortable with a black-box approach, it might be

challenging to give them trust if overtime, breaks, or shift assignment recommendations are involved. SHAP and LIME are transparent methods from the Explanation field of XAI that can be used to identify drivers of the forecasts and boost transparency [31] [32]. Additionally, distributed platforms incorporating sensitive customer, employee, and behavioral data will demand privacy-preserving techniques, like federated learning, especially when being integrated in the contact center [33]. Therefore, it is critical that principles of fairness and human-AI interaction are enforced when building predictive workload systems to guarantee service performance improvements that do not result in biased, non-transparent or negative labor practices [34, 35].

AI-driven Predictive Workload Management in Contact Centers serves as a review of the socio-technical nature of AI for handling contact centers' workloads. This articulates the research on several topics related to queueing models, workload forecasting, staffing optimization, deep learning, automation, explainability, privacy and governance of human-AI. This is the core theme: Workload management research should shift from focused on prototype accurate forecasting and move towards end-to-end frameworks for workload management that are interpretable, adaptable and ethically governed.

2. Review Method and Scope

This review aims to take a structured narrative review approach to pull together the findings and insights from studies on AI for predictive workload management in contact centers. A complete meta-analysis was not suitable given the diverse domains of the literature ranging from call center queueing theory, arrival-rate forecasting and workforce staffing, to time-series machine learning, deep-learning, "assistive" chatbot service automation, explainable AI, fairness and human-AI interaction. In contrast, the review is thematic, ensuring a linkage between operational foundations and recent AI methods and responsible deployment considerations.

Part i) addressed the modelling of workload in contact centers as a problem of queuing and staffing; (ii) has modelling in contact centers changed?; (iii) what is now a more appropriate way to model workload in contact centers?; and (iv) how has the industry responded to this change in modelling? Describe in brief: (ii) what are the changes in workload forecasting by statistical, machine learning and deep learning methods? To answer these questions, we must understand the disciplines involved in forecasting, turning those projections into decisions about staffing, routing and real-time management, and the governance ramifications of using AI to affect service access, labor allocation and operational control. The questions can be interpreted as moving from classical operational models [1]–[6] to questions of arrival forecasting [7]–[18] and using forecasting and uncertainty to make decisions, and finally to the questions addressed in the area of AI enabling forecasting, optimization, and responsible decision support [19]–[35].

Table 1. Review Search Strategy and Inclusion Logic

Review Dimension	Scope Applied in This Review	Representative References
Operational foundation	Queueing models, Erlang-type staffing, abandonment, service levels, multi-skill routing	[1]–[6], [15]–[18]
Workload forecasting	Daily, intraday, density, and uncertainty-aware arrival forecasting	[7]–[12]
Staffing and optimization	Forecast-to-staffing conversion, chance-constrained staffing, simulation, workforce planning	[13]–[18]
AI forecasting methods	Machine learning, deep learning, probabilistic forecasting, attention, transformers	[19]–[27]
Prescriptive analytics	Translation of predictions into staffing, routing, scheduling, and operational actions	[28], [29]
Automation effects	Chatbots, demand deflection, escalation, and AI-mediated customer service	[30], [35]
Responsible AI	Explainability, privacy, fairness, human oversight, and governance	[31]–[35]

This review paradigm can be thought of as a prediction-to-decision pipeline, but rather than that, it is conceptualized as a forecast-to-decision pipeline.

Historical and real-time data → Workload forecast → Staffing and routing decision → Service outcome → Feedback and model updating

This framing is essential since the utility of AI technologies in contact centers is not just about its predictive accuracy, but also how it can help improve service levels, reduce operational expenses, distribute labor more fairly, and facilitate ethical and transparent leadership decisions.

3. Contact-Centre Workload predicted as a Predictive Operations Problem

Contact-center workload is more than just a number of incoming contacts; it's a service-system problem that is stochastic and dynamic in nature. It will be the planning managers' responsibility in each planning interval to estimate the number of contacts that will arrive, the time that the contact takes to be resolved, which skills will be required, which channels the customers will use, and how to allocate available agents. The basic works in call centers then revolve around the issue of workload: the relationship amongst arrival streams, service-time distribution, delay in the queue, abandonment, routing and capacity of staff [1]–[6]. The simplest form of interval workload can be expressed as:

$$W_t = \lambda_t \times E[S_t]$$

As for expected workload in interval t , W_t can be expressed as a product of the expected arrival rate of contacts, λ_t , and the expected contact's service / handling time, $E[S_t]$. But these simplistic formulations are not operational in real contact centers because the presence/absence of after-work, shrinkage, and channel work switching, agent skill differences, customer aversion to wait, and contact automation deflection all impact workload. Instead, a more accurate version would be:

$$W_t^* = \sum_{c=1}^C \lambda_{t,c} (1 - d_{t,c}) \times E[S_{t,c}] + ACW_{t,c}$$

Where c represents channel or contact type, $d_{t,c}$ refers to the deflection or self-service resolution rate, $E[S_{t,c}]$ are expected handling times and $ACW_{t,c}$ refers to after-call work. This expression reveals why the essence of the workload must also be captured and modeled in AI-powered workload management including workload composition and workload effort by channel. Operational link between predicted workload and staffing decisions are provided by classical queueing models. These models are based on an Erlang-type model, which makes estimates about the arrival rate and length of service, whereas other extensions like extensions $M / M / N + M$ include loss of the customer by abandonment, which is essential in the real call-center systems [3],[5],[6]. These models continue to be useful in converting predictions into estimates of services to customers. For instance, a 1,000 call forecast will be of significance to the manager only if broken down into future call waiting time, abandonment risks, agent occupancy, and necessary agents.

Table 2. Core Variables in Predictive Contact-Centre Workload Management

Variable	Operational Meaning	Why It Matters for AI-Driven Management
Arrival rate (λ_t)	Expected number of contacts in interval t	Primary driver of forecasted workload [7]–[11]
Average handling time ($E[S_t]$)	Expected service duration per contact	Determines labour minutes required per interval [1], [4]
Abandonment rate	Customers leaving before service	Indicates service failure and queue pressure [5], [6]
Service level	Percentage answered within target time	Main operational performance target [1], [17]
Occupancy	Share of agent time spent handling work	Balances efficiency against burnout risk [2], [17]
Shrinkage	Paid time unavailable for handling contacts	Converts workload into actual staffing requirement
Skill group	Agent capability for contact type	Essential in multi-skill routing and staffing [15]
Forecast error	Difference between predicted and actual workload	Drives overstaffing, understaffing, and service instability [13]

The multi-skill routing and omni-channel demand of today's contact centers adds to the problem. A single customer journey could go from a chatbot to live chat to voice to back office follow-up. But workload dependence is subject dependent as well as time dependent: Congestion in one channel can influence demand in other ones, and poor automation performance can lead to

an increase in human-agent workload. Some of these complexities can be mitigated using multi-skill staffing models and simulation-based approaches, which enable flexible allocations of agents to the various types of contacts [14, 15].

Therefore, it is essential to interpret the forecasting process as a decision-making context to motivate predictive workload management. Therefore, the forecasting context has to be considered as a decision-making process to motivate predictive workload management. The aim is not only to reduce the error in forecasting the rates of change but to reduce the operation consequences of this forecasting error. Low predictions can lead to long queues and leads to losses of abandonment while over prediction will lead to unnecessary labor cost. Thus, efficient AI systems should capture at least two kinds of uncertainties: uncertainty regarding workload and uncertainty.

$$\hat{W}_{t+h} = f(X_t, X_{t-1}, \dots, X_{t-k}) \pm \epsilon_{t+h}$$

The model predicts the workload at horizon h , W_{t+h} , which is equal to the sum of the historical and contextual predictors (represented by X) and the forecast uncertainty (ϵ_{t+h}). Such a framing with uncertainty is key to staffing under uncertain demand [13] and to the general trend from static workforce planning to a more “operational” way of controlling the workforce with AI systems, [17] [18].

Contact-center workload should always be treated as a state of operation rather than as a quantity measure, as a whole. The amount of work seen during any period is a result of the interaction between the demands, duration of the service, routing and abandonment patterns, agent timings and service-level targets. Thus, predictive workload management will need to be integrated with the forecasting models, with the queueing logic, and with the staffing optimization and live operational feedback. Accordingly the use of statistical, machine learning techniques and deep learning techniques to predict workload in the contact-center domain, and their differences in supporting decision making for adaptive workforce management are discussed next [7-12], [19-27].

4. Forecasting Contact-Centre Workload

The analytical foundation behind all aspects of predictive workload management, staffing, routing, break planning, overtime allocation and service level control requires accurate and believable predictions of incoming demands. Unlike traditional demand forecasting, workload forecasting has immediate operational implications in contact centers, affecting issues such as surge in queues, abandonment and service levels violations on the former end of the scale, and labor expenses and decreased agent productivity on the other [1, 2, 13]. So an effective forecasting model needs to forecast not only expected volume of contacts but also uncertainty, variations within the day, changes in handling time, channel mix and a shift in workload because of automation.

4.1 Statistical and Queue-Informed Forecasting

The early research into forecasting arrival rates in contact centers revealed the presence of a strong time-dependence of arrivals along with several seasonality within the down-to-the-minute window, as well as day-of-week, holiday, etc. seasons, and event-triggered peaks [10]–[15]. Avramidis et al. [7] modeled daily flows of arrivals to a telephone call center and showed how it was crucial to obtain the systematic arrivals. Taylor [8] compared univariate time-series approaches to the intraday arrivals, and found that for high-frequency recurrent demand in contact-centers, methods capable of capturing the repeating nature of the intraday demand are required. Shen & Huang [9] also argued that forecasts have to be revised during the day since any early forecast deviations that are actually clues might carry useful information to other, later parts of the day.

One of the common objectives for forecasting is:

$$\hat{\lambda}_{t+h} = f(\lambda_t, \lambda_{t-1}, \dots, \lambda_{t-k}, Z_t)$$

where λ_{t+h} is a prediction of the number of arrivals at the horizon h , historical temporal structure comes from previous arrivals and Z_t are factors to be considered beyond the model, such as holidays, campaigns, outages, weather, billing cycles, or chatbot activity. But staffing predictions are not enough for workforce management. They must be translated to workload forecasts taking into consideration the handling time that is expected:

$$\hat{W}_{t+h} = \hat{\lambda}_{t+h} \times \widehat{E[S_{t+h}]}$$

Where $E[S_{t+h}]$ is the predicted average handling time. This separation is significant since a lower number of cases per day could result in heavy workloads if the cases are highly complex, escalated or assigned to specialized agents.

Value added in the density forecasting also exists because managers require uncertainty intervals as well as a point forecast. Taylor [12] demonstrated that density forecasts for intraday arrivals have the potential of helping risk based staffing by informing

about the possible range of actual demand. This is directly applicable to chance-constrained staffing where staffing decisions are required to meet service-level targets in the presence of uncertainty about the service demand [13].

4.2 Machine Learning and Deep Learning Forecasting

Machine learning techniques extend contact-center forecasting and are able to detect non-linear relationships between workload and contextual factors. The general forecasting literature indicates that statistical techniques continue to be competitive but machine learning can achieve a good performance in the presence of complex predictors and nonlinear temporal interactions [19], [20]. These are factors that can be used to predict a customer like customer type, product type, and channel, customers exposed to the campaign, chatbot containment percentage, available agents, service disruptions, sentiment, and historical queue congestion in contact centers.

The multi-horizon and high-frequency workload forecasting is particularly relevant for deep learning methods. Recurrent neural networks can capture the temporal relationship between sequential data, and probabilistic recurrent networks can produce uncertainty-aware forecasts of sequential data [21, 22, 23]. N-BEATS (Neural-Based Expansion and Decomposition) is an interpretable algorithm for time-series forecasting based on basis expansion and decomposition [24]. These can be extended to encompass attention-based architectures and transformers which are able to determine which previous time steps, covariates, or seasonal factors are more relevant to future demand [25]–[27]. Temporal Fusion Transformers (TFTs) are an interesting family of models for interpretable multi-horizon forecasting, as they incorporate recurrent units, attention, static input covariates and variable-selection [25]. Long-sequence forecasting can be problematic for models due to poor performance of transformers; models that work with work data of high frequency and long-sequence are called informer type models [27].

Table 3. Forecasting Approaches for Contact-Centre Workload

Method Type	Typical Use	Strengths	Limitations	Relevant References
Exponential smoothing / classical time series	Intraday and daily arrivals	Simple, fast, operationally interpretable	Limited nonlinear modelling capacity	[8], [12], [19]
Arrival-process models	Daily and interval arrival estimation	Strong fit to queueing-based staffing	Often assumes stable historical patterns	[7], [11]
Intraday updating models	Real-time reforecasting	Adapts to same-day demand deviations	Requires timely data pipelines	[9], [12]
Probabilistic forecasting	Staffing under uncertainty	Produces prediction intervals and risk estimates	More complex to calibrate	[12], [13], [23]
Machine learning models	Demand with external drivers	Handles nonlinear covariates	May lack transparency	[20], [21]
Deep learning models	High-frequency, multi-horizon forecasting	Captures complex temporal dependencies	Requires data volume and monitoring	[21]–[24]
Attention / transformer models	Multivariate and long-horizon forecasting	Captures long-range dependencies and variable importance	Computationally intensive; risk of overfitting	[25]–[27]

The best forecasting method for AI-driven workload management is not necessarily the most complicated but the method that also has the most positive effect on the operation. A very accurate black-box prediction might not be as valuable as a less accurate prediction that includes some lines of explanation on why it is believed that demand will increase, is accompanied with calibrated uncertainty bounds, and is somewhat transparent to other engineering choices for staffing. Therefore, evaluation of the forecasts should involve statistical and/or operational indicators:

$$\text{Operational Forecast Value} = g(\text{MAPE}, \text{RMSE}, \text{Prediction Interval Coverage}, \text{Service-Level Impact}, \text{Staffing Cost})$$

This more comprehensive assessment approach is consistent with a shift in assessment approach from forecasting accuracy to decision aware analytics. Forecast error should not only be measured by the traditional error tools [13], [17], [18] in the context of a contact center but by its influence on the waiting time, abandonments, occupations and stability in the schedule.

Overall, the literature indicates that contact-center workload forecasting is making the transition from volume-based forecasting to adaptive forecasting, probabilistic forecast, multivariate and AI-assisted forecast. While statistical approaches can continue to be useful benchmarks, deep learning and transformer based models can be more effective in more complex, multi-modal set up.

The next hurdle is translating these projections into staffing and scheduling decisions, planning allocations, and real-time control decisions – a unified AI-powered workload management framework will be necessary to do that.

5. AI-Driven Predictive Workload Management Framework

Predictive workload management should be based on an integrated forecast-to-decision system motivated by AI, not on an AI forecast model. The traditional approach to workforce management is based on queueing theory and managers' rules of thumb to deduce required staffing needs from estimated arrival rates. Historical data, real time signals, forecasting models, optimization engines, metrics for automation, and human supervisory judgement are all integrated into the process to better inform strategic decisions and intraday control decisions in an AI-driven system [13]–[18],[28].

The proposed framework composes of 5 related layers where: the data acquisition concerns the acquisition of data; the predictive modelling is assigned to building predictive models; the prescriptive optimization will involve the building of prescriptive optimizations; the real-time control will be dealt with in real-time control; and the last one is the governance feedback, which consists of governance feedback. The data layer aggregates the historical data of contact volumes, AHT, service levels, abandonment, channel mix, agent schedules, shrinkage, customer segments, campaign events, chatbot containment rate, escalation rate, and external calendar/metric effects. The forecasting layer relies on these inputs to forecast the workload at multiple time scales, from long-term capacity planning, through weekly scheduling, and down to daily staffing and intraday reforecasting [7]– [12] and [27]– [19].

The prescriptive layer transforms predicted workload to decisions related to staffing and operations. This is where the difference between AI-based workload management and traditional predictive analytics lies. Forecasts need to be translated into decisions regarding staffing, skills allocation, breaks, overtime, voluntary time off, routing rules, rules for chatbot escalation and cross channel load balancing. The shift from prediction to prescription is vital since operationally, value lies in the improvement of forecasts for their impact on decisions [28]. In Contact Centers, this conversion needs to be done in combination with forecasting models, queueing logic, simulation and optimization techniques [13] – [16].

To communicate an objective that broadly identifies what one want to achieve:

$$\min_{x_t} \sum_{t=1}^T (C_s x_t + C_w E[Wq_t] + C_a A_t + C_o O_t)$$

subject to:

$$P(SL_t \geq SL^{target}) \geq 1 - \alpha$$

Where, x_t is Staffing level in interval t ; C_s is its cost; C_w is waiting time cost; C_a is abandonment cost; C_o is overtime or occupancy related cost; Wq_t is the expected waiting time; a_t is the expected abandonment and; SL_t is the service level. The “chance constraint” concept is related to the logic underlying uncertainty-aware staffing [13] and guarantees high probability of meeting service level goals when demand is uncertain.

Figure 1. Conceptual Architecture of AI-Driven Predictive Workload Management

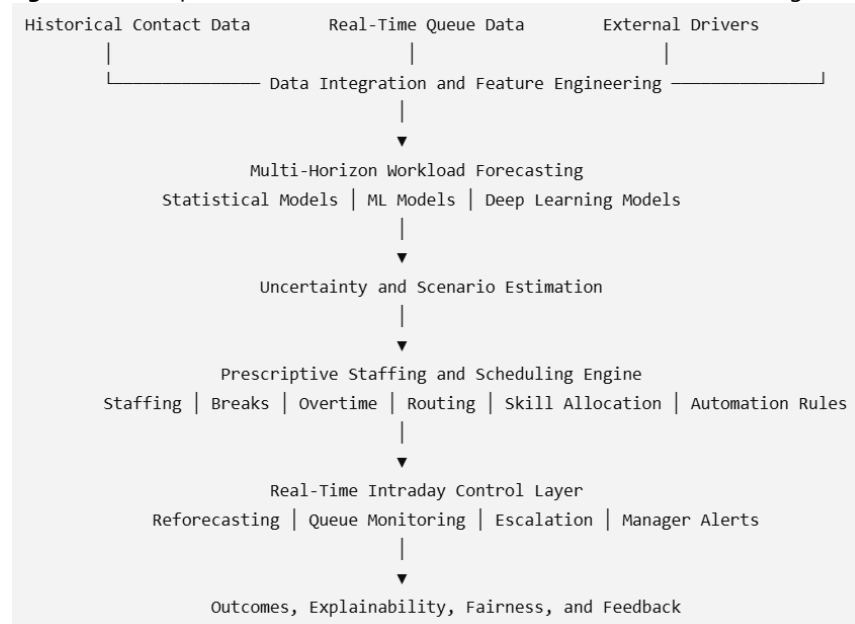


Table 4. Functional Layers of the Proposed AI-Driven Workload Management Framework

Framework Layer	Main Function	Typical Inputs	Expected Output	Key References
Data acquisition	Collect operational and contextual signals	Contact volume, AHT, abandonment, schedules, channel mix, chatbot data	Unified workload dataset	[7]–[12]
Forecasting layer	Predict future workload across horizons	Historical demand, seasonality, external drivers, real-time deviations	Point and probabilistic forecasts	[19]–[27]
Queueing layer	Translate workload into service performance	Arrival rate, handling time, patience, staffing level	Waiting time, abandonment, service level	[1]–[6]
Prescriptive layer	Recommend staffing and routing actions	Forecasts, constraints, costs, service targets	Staffing, scheduling, routing decisions	[13]–[18], [28]
Real-time control	Adjust operations during the day	Live queues, forecast error, agent availability	Intraday alerts and re-optimization	[9], [12], [17]
Governance layer	Ensure trust, fairness, and accountability	Model explanations, audit logs, privacy rules	Transparent and responsible decisions	[31]–[35]

In modern contact centers, the real-time control layer plays a crucial role. Static schedules are provided days or weeks ahead, but actual work orders may vary significantly based on things such as outages, payment cycles, product failures, marketing campaigns or viral events on social media channels. Deferred work, re-allocating agents from one queue to another, modifying the rules used for chatbot escalation, extending hours of overtime or adjusting the staffing level prior to impacting the service level are some of the possible scenarios where intraday reforecasting can be helpful [9] [12] [17]. Therefore, the system should continually compare predicted workload or observed workload:

$$FE_t = W_t - \hat{W}_t$$

Shown as $FE_t = (W_t - \hat{W}_t)$ where FE_t is the forecast error, W_t is the observed workload and $\hat{W}_t$ is the predicted workload. Constant error can be a sign of concept drift, the inappropriate representation of the features, unusual events or changes in customer behavior.

Additionally, workload management is a socio-technical process, which is the case that needs to be understood at a high level. Agents are influenced by AI-generated recommendations in their schedule, workload, breaks, overtime, and performance-related stress. Hence, explainability and human oversight should be an inherent part of a framework, rather than an addition to it after deployment. Explainable AI methods have the potential to demonstrate if the rise in forecasts is countered seasonally, as a result of campaign effects, Chatbot failing, migrating channels, or atypical handling times [31, 32]. In some cases, the use of sensitive customer and employee data on distributed service platforms can be used to mitigate risk, especially in the context of privacy preserving learning [33]. In addition, there are standards of fairness and human-AI interaction to account for undesirable shifts, overload or performance pressure being systematically allocated to certain groups of the agents being optimized [34, 35].

The proposed framework presents AI-enabled predictive workload management as an adaptive operational intelligence system, as a whole. It is not just about predicting the need for capacity; it is about bringing together prediction, uncertainty, staffing, routing, automation, and governance as a closed loop decision-making system. The need arises in the context of more complex, AI-assisted customer interactions, reducing labor inefficiency, supporting agents, and enhancing service reliability, which are all critical concerns for contact centers in today's landscape.

6. Automation, Agent Assist, and Demand Deflection

AI-powered workload management shouldn't be about forecasting workloads in the contact center, it's about changing your workload. Chatbots, voice bots, interactive self-service, automated triage, agent-assist systems, call summarization, knowledge recommendation, sentiment detection, and generative AI tools can improve handling time and/or decrease the number of contacts for human agents, or increase the amount of first-contact resolution. They could also prioritize the remaining tasks in more advanced, emotional or escalated interactions. Hence, automation should be considered as a variable in the workload workflow, and not as a non-signed bendable thick layer of technology outside the workflow [30, 35].

It can be represented by a simple workload-equation suitable for simple automation:

$$W_{t,c}^{AI} = \lambda_{t,c}(1 - d_{t,c}) \times E[S_{t,c}^H] + \lambda_{t,c}e_{t,c} \times E[S_{t,c}^E]$$

where $W_{t,c}^{AI}$ is AI-adjusted workload for channel c in interval t , $\lambda_{t,c}$ is contact arrival volume, $d_{t,c}$ is the successful automation deflection rate, $E[S_{t,c}^H]$ is expected human handling time for non-deflected contacts, $e_{t,c}$ is the escalation rate from AI to human agents, and $E[S_{t,c}^E]$ is expected handling time for escalated cases. As the equation above suggests, a high deflection rate may not be the answer to all managerial problems because escalated contacts may become longer, more complicated and/or may require specialist agents.

Table 5. Automation Effects on Contact-Centre Workload

AI Capability	Expected Workload Benefit	Potential Operational Risk	Workload Metric Affected
Chatbots and virtual assistants	Deflect routine contacts and reduce queue pressure	Failed containment may increase repeat contacts	Deflection rate, repeat contact rate
Voice bots and IVR automation	Automate identification, routing, and simple requests	Poor intent recognition may increase abandonment	Containment, abandonment, transfer rate
Agent-assist tools	Reduce search time and improve response consistency	Overreliance may reduce agent judgement	Average handling time, quality score
Generative AI summarization	Reduce after-call work and documentation burden	Inaccurate summaries may require correction	After-call work, compliance errors
Sentiment and intent analytics	Improve routing and escalation prioritization	Misclassification may create unfair service outcomes	Escalation rate, service level
Predictive routing	Match customers with suitable agents	May overload high-performing agents	Occupancy, workload balance

Existing models heavily depend on the historical data about time of arrival and handling time [7] – [12]. However, historical workload may not equal future workload when the percentage of "containment" of the chatbot increases or productivity of agents change with a generative AI assistant or a product outage. This will give rise to concept drift wherein demanded drivers change over time, compared with the operational load. However, such patterns and their changes can be identified using deep learning and attention-based forecasting models, which needs to be carefully monitored and linked to the operational feedback loops [21]–[27].

But from a prescriptive analytics point of view, the use of automation should be assessed as a part of staffing and routing rather than on its own. Some of the suggestions include the expansion of live-chat capabilities, reallocation of qualified agents, prioritization of non-urgent back-office responsibilities, or adjustment of the bot escalation rules, among others, when the rate of chatbot escalation reaches a predetermined threshold. This mirrors the broader shift to the more recent approach of prescriptive decision support, involving the use of AI generated forecast information to make specific operational decisions [28] [29].

But there are issues about governance with automation-driven workload management, too. Agent-assist systems impact employee performance, management judgement of productivity and customer prioritization. These systems, if not well-designed, may increase workload, decrease autonomy or surface challenging contacts to certain agents. Human-AI interaction principles are therefore essential to ensure that AI is used to aid the agent, rather than to monitor or increase the monitoring and performance pressure of the agent. Also when the data of automation are used to inform decisions regarding workforce size, scheduling, or staff performance, explainability, fairness and privacy safeguards are essential [31]–[34].

When considering automation, it's important to view it as a workload reducing, as well as a workload redistributing, process. What the best "AI-powered" contact-center solutions will do is not merely quantify how many contacts are being "deflected"; they will also reduce complexity of escalation, adjust handling times, balance workloads, enhance customer satisfaction and impact fairness.

7. Explainability, Fairness, Privacy, and Human-AI Governance

Workload management systems using AI have implications on how business is conducted that impact both customers and employees. A workload forecast can forecast how many agents are needed, which agents to place in problematic queues, whether overtime is necessary, how the overtime breaks are split up and the speed at which the customer can be served. Hence,

one should not consider such systems merely in terms of forecasting or in cost-saving with respect to labor. These need to be evaluated for explainability, fairness, privacy, accountability, and human oversight as well [31]–[35].

Explainability is key, as when an AI system is predicting a surge in workload or suggesting a change in staffing, contact-center managers need to know the reason behind these predictions. While black-box models can make very good predictions, sometimes it may be hard to believe these predictions, especially when they relate to staffing, routing, or service-level commitments. These techniques (SHAP and LIME) are useful to explain the inputs and outputs of AI based forecast engines and determine if the predictive model was attributed to other factors including seasonality, campaign activity, unpredictable handling time etc. [31, 32]. These explanations are useful in the field to enable managers to differentiate between "normal demand fluctuations" and "abnormal demand spikes that trigger interventions.

A "simplified" explaining function is given by:

$$\hat{W}_{t+h} = f(X) \quad ; \quad E_i = \phi_i(X)$$

Where, W_{t+h} is the estimated workload, X is the set of features, and E_i or $\phi_i(X)$ is the value of a particular feature to the prediction. When implemented practically, this indicates that workload will grow by 18% and it should detail that this increase is linked to billing-cycle timing, reduced spending of time in a chatbot, and extended average handling time.

Fairness matters too as optimization may generate unfair burdens on the agents or groups of customers. For instance, routing systems can have high-performing workers receive the same contacts over and over again, which can lead to burnout. It can be the case that employee groups receive more shifts with poor attributes than others. Also if service patterns have historically been biased, there might be unequal service times for different customers using customer prioritization models. The fairness-aware workload management therefore should track customer-facing and employee-facing effects [34].

Table 6. Responsible AI Risks and Mitigation Strategies in Predictive Workload Management

Governance Area	Risk in Contact-Centre Workload Management	Mitigation Strategy	Key References
Explainability	Managers cannot understand why workload or staffing recommendations change	Use SHAP/LIME explanations and feature-level forecast drivers	[31], [32]
Fairness to agents	Difficult contacts, overtime, or undesirable shifts are unevenly allocated	Monitor workload distribution by team, skill, tenure, and protected attributes where lawful	[34], [35]
Fairness to customers	Certain customer groups may experience longer waits or poorer routing	Audit waiting time, abandonment, and resolution outcomes across customer segments	[34]
Privacy	Sensitive customer and employee data may be over-collected or centralized	Apply data minimization, access controls, anonymization, and federated learning	[33]
Accountability	Managers may over-rely on AI recommendations	Require human approval for high-impact staffing and routing decisions	[35]
Model drift	AI recommendations degrade as demand patterns or automation behaviour changes	Use continuous monitoring, forecast-error alerts, and periodic retraining	[21], [25], [27]

Maintaining customer privacy is also an essential need as predictive workload systems tend to involve a lot of sensitive data, such as transactions with customers, calls is transcribed into, sentiment labels attached to them, agent performance data, adherence to schedule, behavior indicators, and more. It may be more advantageous to centralize such data, which could raise questions of surveillance and cyber security. To prepare models in a distributed environment without transferring all raw data ultimately to a single repository, organizations could adopt federated learning privacy-preserving approaches [33]. This is especially important in that of multinational or outsourced contact-center networks where data protection regulations and contractual obligations and geographical requirements might differ.

The governance of human-AI should build on and apply the judgment of managers and agents, and should not supplant it. The system should suggest, in addition to the recommendations, the levels of confidence in its recommendations, and then explain them; and should be able to override recommendations at the last moment when contextual information is stronger than any model output. One of the key guidelines regarding human-AI interaction is visibility, controllability, feedback, and appropriate reliance all of which are key in high-stakes clinical settings [35]. This is the case for contact centers where supervisors need to understand how the system is recommending overtime, the underlying assumption that governs the overtime, the possible alternatives and the service-level risk associated each of them. There is a disjointed governance workflow that can be expressed as:

AI forecast → Explanation → Risk score → Human review → Operational decision → Outcome audit → Model update

This workflow will also help with accountability as it will generate an algorithmic recommendation log, as well as a human decision log. It also mitigates a typical pitfall that many managers fall into: blindly following the AI-generated output or perhaps disregarding it completely. The objective is calibrated trust: Where AI is dependable, users should trust it, but where uncertainty exists, users should question it, and (when appropriate and warranted by context analysis) contradict it.

There is no way for a responsible AI to not be integral to the deployment quality of predictive workload management. Any non-stable socio-technological system that makes the forecasts more accurate but leads to the unfair distribution of the workload, to an increase of privacy risks, to less transparency in management is not successful. From a combined perspective of AI performance and prediction, value for operation, explainability, fairness, privacy, and human-centered governance, AI workload management should be evaluated at the high level in the research and implementation phase.

8. Research Gaps and Future Agenda

Yet, although significant strides have been made in forecasting, optimizing and automating service via AI technologies in contact centers, the field of study has been divided across several different domains such as operations management, time-series forecasting, machine learning and responsible AI. The earliest research has developed models for the call-center arrival, abandonment, staffing and service level [1]–[6], and subsequent models further developed intraday forecasting and density forecasting and uncertainty-aware staffing [7]–[18]. But this new generation of predictive workload management needs a more comprehensive research program that ties the forecasting accuracy, prescriptive decision-making, automation impact, and agent well-being to AI governance.

There is a big missing link when it comes to multi and cross channel workload forecasting. Many of the classical literature was written about the arrival of telephone calls [1], [7]–[12]. But today's contact center can function via voice, live chat, email, messaging apps, social networks, self-service portals and AI chatbots. These channels are not distinct. Low bot engagement rates can drive up the demand for live-chat; late email replies can lead customers to voice; and high messaging volume in a queue can swell the volume in another queue. To this end, future studies need to focus on multivariate and hierarchical forecasting models, which take into account the relationship between channel substitution, escalation and journey of customers.

A second shortcoming is the weak connection between Prediction and Prescription. Model evaluation is commonly performed with statistical error measures in forecasting studies, and in staffing studies, constraints important for services and the operational feasibility are the foci [12], [13], [17] and [18]. But, in reality, the benefit of a forecast is based on the benefit to staffing, routing, break planning, making overtime decisions and service outcomes. It is recommended that AI models be evaluated with future studies on decision-based metrics to assess service-level improvements, reduction in abandonment, staff cost savings, schedule stability, and schedule fairness [13], [28], [29].

The third gap is in relation to the redistribution of workload associated with automation. Chatbots and Agent assist can help in lowering general demand but may also lead to increased complexity of human interactions [30], [35]. There is a lack of adoption of chatbot containment, escalation difficulty in the existing workload models, customer frustration after interacting with a bot and reduction in the after-call-work generated by AI. Future models should no longer consider automation as an external workload type, but rather as an endogenous workload type.

Table 7. Research Gaps and Future Research Directions

Research Gap	Why It Matters	Future Direction	Relevant References
Multichannel workload modelling	Voice-only assumptions no longer reflect modern contact centres	Develop cross-channel, hierarchical, and customer-journey-based forecasting models	[1], [7]–[12], [21], [25], [27]
Forecast-to-decision integration	Accurate forecasts may not improve operations if not linked to staffing decisions	Use decision-aware evaluation and prescriptive analytics	[13]–[18], [28], [29]
Automation-driven workload shifts	Chatbots may deflect simple contacts but increase escalation complexity	Model containment, escalation, repeat contacts, and human workload intensity	[30], [35]
Concept drift and demand shocks	Historical patterns may fail during outages, crises, campaigns, or AI deployment changes	Apply adaptive retraining, drift detection, and real-time reforecasting	[9], [12], [21], [25], [27]
Explainability and trust	Managers may reject or misuse opaque AI recommendations	Integrate SHAP, LIME, forecast-driver explanations, and human review	[31], [32], [35]
Fairness and agent well-being	Optimization may unevenly allocate undesirable shifts or difficult contacts	Audit workload allocation, schedule equity, occupancy, and burnout indicators	[34], [35]
Privacy-preserving learning	Contact-centre data include sensitive customer and employee information	Explore federated learning and data-minimization frameworks	[33]

Another significant direction is the concept drift. Seasonality, product issues, marketing activity, regulatory shifts or changing customer preferences, and new automation software can all bring in changes in workload patterns. Thus, proposed drift monitoring systems should contain drift monitoring:

$$D_t = |P_t(W, X) - P_{t-k}(W, X)|$$

Where D_t represents changes in the distribution of workloads' features to current workload patterns and past workloads' features. If the drift is above an operation threshold, then there should be a recalibration of the model or a review of the staffing by human.

Last but not least, research into predictive workload management should regard it as a human-oriented operational intelligence system. Cost and automation should not be the end goal. Conversely, the future frameworks should be designed to maximize the benefits of service quality, staff efficiency, fairness for consumers, satisfaction of service providers, transparency of operations and consumers as well as privacy of consumers [31]–[35]. Such interdisciplinary research is needed to bring together the fields of operations management, AI forecasting, human factors, labor analytics, and responsible AI governance.

9. Conclusion

Predictive workload management has been revolutionized by AI and it is taking contact-center operations from the world of static, rule-based staffing to that of intelligent, data-driven and human-based operational intelligence. From literature review, it is found that queueing theory is important to understand the service level, abandonment and the staffing requirement, while statistical and probabilistic forecasting methods are helpful to predict workload variation in a day and also intraday workload variation [1]–[6]. Recent machine learning, deep learning, and transformer based forecasting models bring this capability a step further by modelling the workload patterns in nonlinear, multivariate, and longtime horizons [19]–[27]. However this is not enough in forecasting accuracy. This is where the predictive capabilities out value itself when combined with prescriptive staffing, scheduling, routing and automation control and feedback systems and real-time mechanisms [13]–[18], [28], [29].

It also notes that automation remodels the type of work it aims to alleviate. While chatbots and agent assist systems can help offload simple requests, they can also add complexity to the escalation path and bring about changes in the workload intensity of agents as a result of the automated nature of the assist which may alter their types of interaction [30] [35]. Thus, deployment needs to be responsible and could be made easy thanks to explainability, fairness auditing, privacy-preserving learning, and human oversight [31]–[35]. Going forward, researchers are required to develop frameworks that integrate customer experience, staffing efficiency, employee wellness, transparency and ethics for robust customer support services that also deliver efficient and sustainable processes for contact center agents.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID iD:

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- [1] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79–141. <https://doi.org/10.1287/msom.5.2.79.16071>
- [2] Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6), 665–688. <https://doi.org/10.1111/j.1937-5956.2007.tb00288.x>
- [3] Koole, G., & Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113, 41–59. <https://doi.org/10.1023/A:1020949626017>
- [4] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469), 36–50. <https://doi.org/10.1198/016214504000001808>
- [5] Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3), 208–227. <https://doi.org/10.1287/msom.4.3.208.7753>
- [6] Mandelbaum, A., & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5), 1189–1205. <https://doi.org/10.1287/opre.1080.0651>
- [7] Avramidis, A. N., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7), 896–908. <https://doi.org/10.1287/mnsc.1040.0263>
- [8] Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2), 253–265. <https://doi.org/10.1287/mnsc.1070.0786>
- [9] Shen, H., & Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3), 391–410. <https://doi.org/10.1287/msom.1070.0199>
- [10] Aldor-Noiman, S., Feigin, P. D., & Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, 3(4), 1403–1447. <https://doi.org/10.1214/09-AOAS255>
- [11] Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3), 865–874. <https://doi.org/10.1016/j.ijforecast.2015.11.012>
- [12] Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3), 534–549. <https://doi.org/10.1287/mnsc.1110.1434>
- [13] Gurvich, I., Luedtke, J., & Tezcan, T. (2010). Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7), 1093–1115. <https://doi.org/10.1287/mnsc.1100.1173>
- [14] Atlason, J., Epelman, M. A., & Henderson, S. G. (2004). Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127, 333–358.
- [15] Cezik, M. T., & L'Ecuyer, P. (2008). Staffing multiskill call centers via linear programming and simulation. *Management Science*, 54(2), 310–323. <https://doi.org/10.1287/mnsc.1070.0824>
- [16] Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., & Wu, X. (2007). A survey and experimental comparison of service-level-approximation methods for nonstationary M(t)/M/s(t) queueing systems. *INFORMS Journal on Computing*, 19(2), 201–214.
- [17] Koole, G., & Li, S. (2023). A practice-oriented overview of call center workforce planning. *Stochastic Systems*. <https://doi.org/10.1287/stsy.2021.0008>
- [18] Ding, S., & Koole, G. (2022). Optimal call center forecasting and staffing. *Probability in the Engineering and Informational Sciences*. <https://www.cambridge.org/core/journals/probability-in-the-engineering-and-informational-sciences/article/optimal-call-center-forecasting-and-staffing/52796A0DC4D999FDAF019E03DAD100BD>
- [19] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>
- [20] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889. <https://doi.org/10.1371/journal.pone.0194889>
- [21] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209. <https://doi.org/10.1098/rsta.2020.0209>
- [22] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388–427. <https://doi.org/10.1016/j.ijforecast.2020.06.008>
- [23] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181–1191. <https://doi.org/10.1016/j.ijforecast.2019.07.001>
- [24] Oreshkin, B. N., Carpo, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations*. <https://openreview.net/forum?id=r1ecqn4YwB>

- [25] Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- [26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [27] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12), 11106–11115. <https://doi.org/10.1609/aaai.v35i12.17325>
- [28] Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044. <https://doi.org/10.1287/mnsc.2018.3253>
- [29] Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1), 69–88. <https://doi.org/10.1287/msom.2015.0561>
- [30] Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, 100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- [31] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [32] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [33] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
- [34] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. <https://fairmlbook.org/>
- [35] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>