Foundation Model Empowered Autonomous Network Management for QoS Optimization in 5G-Advanced and 6G Networks

Rohan Seed
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
hellorohan@ku.edu

Bakash Arora
School of Computing, Clemson University, Clemson, SC, USA.
aakasharora96@clemson.edu

Reith Gansen
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
rhansen@oregonstate.edu

Rsaac Nendaza
Department of Computer Science, University of North Texas, Denton, TX, USA.
rsaac1986@unt.edu

Abstract

The move from fifth-generation advanced (5G-Advanced) systems to sixth-generation (6G) networks is predicated on hyper-specialised performance envelopes that demand granular, real-time quality of service (QoS) guarantees across extremely heterogeneous service classes. Traditional and even deep reinforcement learning based management approaches, while effective in narrow slices, encounter fundamental brittleness when confronted with domain shift, novel service descriptors, or the combinatorial complexity of multi-objective resource allocation in open multi-vendor environments. Foundation models, large-scale pre-trained neural architectures that acquire broad representational capabilities from vast and diverse data modalities, offer a paradigm shift toward generalist intelligence in autonomous network management. This paper presents a system-level architectural framework that embeds foundation models as the cognitive nuclei of closed-loop control, enabling intent-driven reasoning, cross-domain policy transfer, and anticipatory QoS assurance without exhaustive per-slice retraining. We examine the structural trade-offs that arise when unifying language-based policy interfaces, vision-based radio environment maps, and time-series telemetry through a shared model backbone, and we dissect the concomitant challenges of inference latency, energy consumption, and model governance. The discussion extends to fairness, accountability, and transparency in AI-driven resource allocation, as well as the deployment pathways that balance centralised model hosting against distributed inference at the network edge. By synthesising advances in model compression, retrieval-augmented generation, and responsible AI frameworks, we articulate a roadmap for foundation model empowered autonomous networks that are robust, sustainable, and aligned with the regulatory and ethical imperatives of the 6G era.

Keywords

foundation models, autonomous network management, quality of service, 5G-Advanced, 6G, network slicing, governance, sustainability.

1. Introduction

The evolution of cellular systems toward 5G-Advanced and the vision of 6G networks imposes a qualitative leap in the management of network resources, driven by extreme throughput, sub-millisecond latency, massive device connectivity, and deterministic service guarantees. Network slicing has been established as the architectural cornerstone for partitioning a common physical infrastructure into multiple logical networks, each tailored to specific QoS profiles that range from enhanced mobile broadband to ultra-reliable low-latency communications and massive machine-type communications. However, the static or semi-static configuration of slices, even when augmented by closed-loop automation, struggles to keep pace with the fluidity of user demand, dynamic spectrum availability, and the introduction of compute-intensive AI-native services that characterise emerging 6G use cases. The proliferation of network softwarisation, multi-access edge computing, and open radio access networks further escalates the degrees of freedom available to operators, rendering traditional heuristic and model-driven resource orchestration methods insufficient for real-time QoS assurance.

Contemporary research has responded by applying deep reinforcement learning (DRL) to dynamic slice management, where agents learn policies for spectrum allocation, admission control, and traffic steering directly from interaction with the environment. Network slicing architectures have been extensively surveyed, underscoring the central role of programmable control planes and software-defined networking in realising on-demand slice lifecycle management [1]. Concurrently, the application of DRL in communications and networking has matured, with comprehensive surveys cataloguing the use of value-based, policy-based, and actor-critic algorithms across numerous resource optimisation problems [2]. These efforts have yielded promising results in standalone scenarios, yet they typically rely on task-specific neural architectures that require prohibitive retraining when exposed to unseen slice types, altered traffic patterns, or heterogeneous equipment from different vendors. The brittleness stems from a fundamental characteristic: learned policies encode correlations that are tightly coupled to the statistical properties of the training environment and lack the conceptual abstraction needed to generalise across network contexts.

A transformative opportunity emerges from the recent ascent of foundation models, large-scale pre-trained systems that acquire a broad understanding of language, vision, and structured data through self-supervised learning on internet-scale corpora. The paradigm, comprehensively examined in a landmark survey, demonstrates that such models can adapt to a wide array of downstream tasks with minimal fine-tuning, and often exhibit emergent zero-shot reasoning capabilities that were not explicitly programmed [3]. In the network management domain, this promise translates into the possibility of a single pre-trained cognitive engine that interprets human-readable intents, analyses radio environment maps extracted from multimodal sensor streams, and recommends QoS-aware configuration changes without the need to train a separate model for each cell, slice, or vendor. The convergence of intent-based networking, where high-level business objectives are automatically decomposed into network-level instructions, has already been articulated as a key enabler for autonomous management [4]. When combined with the representational power of foundation models, intent-driven systems can move beyond rigid templates and

interact with operators in natural language, asking clarifying questions and justifying decisions in a manner that reduces the cognitive burden on human experts.

Earlier DRL-driven resource management for network slicing, as exemplified by deep Q-network and policy gradient approaches, demonstrated that neural agents could outperform manual tuning in controlled lab settings [5]. More recently, a PPO-based deep reinforcement learning method has been proposed to provide rigorous QoS assurance for 5G network slices, addressing both throughput and latency constraints through adaptive optimisation [6]. While these efforts validate the feasibility of AI-driven closed loops, they also expose the limits of narrow, single-task learners. Foundation models, by contrast, hold the potential to amortise learning across tasks, thus reducing overall sample complexity and enabling faster adaptation when new slice definitions or regulatory constraints emerge. This paper tackles the systems-level challenge of harnessing foundation models for autonomous network QoS management, proposing an architectural blueprint, analysing structural trade-offs, and delineating the governance, sustainability, and deployment implications that will determine whether this paradigm can be realised at scale in future mobile infrastructures.

2. Background and Related Work

To situate the contribution of foundation models in autonomous networking, it is necessary to survey the layered landscape of prior art that spans network slicing, DRL-based resource control, zero-touch management, and early explorations of large pre-trained models in communications. The concept of network slicing has been thoroughly characterised in the literature, with foundational surveys detailing the end-to-end architecture, the role of the 3GPP management and orchestration framework, and the isolation mechanisms necessary to prevent slice resource contention [1]. These design principles are being extended in 5G-Advanced with enhanced support for exposure of network data analytics functions and tighter integration with application-level QoS requirements. From an AI perspective, the application of DRL to communications has evolved rapidly; Luong et al. provided a thorough taxonomy of DRL algorithms and their use cases in dynamic spectrum access, caching, offloading, and resource scheduling, highlighting the strength of model-free learning in partially observable environments [2].

The industry has also pursued standardisation of autonomous network operations through initiatives such as the ETSI Zero-touch Network and Service Management (ZSM) framework, which defines a reference architecture where closed-loop automation spans from resource-facing to customer-facing management domains. The ETSI ZSM approach envisions a separation of concerns between domain-specific controllers and an overarching end-to-end orchestrator that leverages AI to resolve policy conflicts and maintain service level agreements [7]. In parallel, intent-based networking has matured from a conceptual model into a concrete architectural paradigm, where declarative intents are used to drive network configuration without specifying low-level commands. Zhang et al. surveyed the entire landscape of intent-based networking, covering intent expression languages, intent fulfillment engines, and conflict resolution strategies, and identified AI-driven intent translation as a critical enabler for truly autonomous networks [4].

Within the narrower scope of DRL for slice QoS, several influential works have demonstrated that deep neural networks can approximate optimal resource allocation policies. Li et al. developed a deep reinforcement learning framework for network slicing that jointly optimised throughput and power consumption, using a custom discrete-continuous action space and demonstrating convergence in scenarios with multiple service types [5]. The specific

application of proximal policy optimisation to 5G slicing was later formalised to guarantee delay and throughput bounds, showing that PPO's clipped surrogate objective provides stable training even in non-stationary traffic conditions [6]. These works, however, operate under the assumption of a fixed slice catalog and a stationary environment. When a new slice with a different numerology or a novel ultra-reliable low-latency communication profile is introduced, the policy must be retrained from scratch, consuming significant simulation or live network time.

Beyond the confines of small-scale DRL models, the foundation model revolution has begun to penetrate the networking domain, though large-scale deployments remain nascent. The core insight, drawn from the comprehensive analysis of Bommasani et al., is that a single pre-trained model can serve as a general-purpose foundation that unlocks emergent capabilities when fine-tuned or prompted for a wide range of downstream tasks [3]. In natural language processing, models such as GPT and its successors have exhibited the ability to perform translation, summarisation, and even code generation without task-specific architectures. These capabilities suggest that a network foundation model, pre-trained on a vast corpus of telemetry data, configuration logs, standardisation documents, and radio maps, could internalise a rich model of network dynamics that transcends any single operator's deployment. Early case studies have shown that large language models can interpret natural language intents and generate valid network configuration snippets for routing and access control, although these demonstrations have not yet incorporated closed-loop measurement feedback. Similarly, vision transformers trained on satellite imagery and ray-tracing outputs can assist in radio resource management by predicting path loss and user density without costly on-site measurements. These threads, when woven together, point toward a holistic cognitive layer that combines language understanding, spatial reasoning, and time-series forecasting within a unified architecture.

3. An Architectural Blueprint for Foundation Model-Based Autonomous QoS Optimization

Realising the vision of a foundation model empowered autonomous network requires a deliberate system architecture that reconciles the immense computational demand of large models with the latency constraints of real-time control loops. We propose a layered cognitive plane that sits alongside the existing management and orchestration stack, interfacing with the service management and orchestration layer of the 5G core and with domain controllers in the radio access network and transport network. At its centre is a multimodal foundation model that ingests three broad categories of data: structured telemetry, unstructured textual information, and spatial-spectral representations. The telemetry stream includes per-slice key performance indicators such as throughput, latency, jitter, packet loss, and physical resource block utilisation, collected from the network data analytics function and radio network information services. Unstructured textual information encompasses operator policies, service level agreements, maintenance logs, and technical standards, which encode institutional knowledge that has traditionally remained outside machine reasoning pipelines. The spatial-spectral stream originates from distributed sensing, including radio environment maps, geolocated channel quality indicators, and camera feeds that inform dynamic spectrum sharing and beam management decisions.

The foundation model acts as a policy reasoning engine that translates high-level intents into a sequence of candidate configuration changes, evaluates their predicted impact on QoS through forward simulation inherent in the model's learned dynamics, and selects the action

that balances multiple, often competing, objectives. This process is orchestrated through a retrieval-augmented generation mechanism, where relevant historical episodes and expert-written resolution playbooks are retrieved from a vector database and prepended to the model's context. Retrieval augmentation is critical because the model's parametric memory, while vast, may not capture the latest site-specific outage patterns or security advisories. By incorporating fresh evidence at inference time, the system reduces the risk of hallucinated recommendations and enables a form of online adaptation that does not require full model retraining, thereby addressing one of the core weaknesses of purely parameterised DRL agents.

The architectural trade-offs revolve around the placement of model inference. Centralised model hosting in a regional cloud confers economies of scale in hardware utilisation, simplifies model versioning, and facilitates federated learning across multiple operator domains, but it introduces a round-trip latency that may be incompatible with the sub-millisecond control cycles required for beam-level scheduling or packet-level admission control. A complementary approach splits the foundation model into a heavy encoder hosted in the cloud and lighter decoder heads deployed at edge nodes or within distributed unit accelerators, a pattern that mirrors the split inference strategies adopted in edge AI systems [8]. For tasks that demand ultra-low latency, such as dynamic resource block allocation in the MAC scheduler, a distilled student model distilled from the larger foundation model can operate autonomously at the edge, periodically synchronising its latent representations with the cloud-hosted teacher. This hierarchical arrangement acknowledges the fundamental tension between reasoning depth and reaction speed, and it ensures that the cognitive plane does not become a scalability bottleneck.

The integration with the existing 3GPP management architecture is another critical design dimension. Standardised interfaces such as the RESTful northbound APIs of service-based management architecture must be augmented with a uniform intent provisioning interface that the foundation model can both consume and produce. The model's natural language understanding capacity enables a new modality of interaction where an operator specifies an intent such as "maintain the tactile internet slice latency below 2 ms with 99.999% reliability during the upcoming stadium event" and the system autonomously decomposes this into a resource reservation plan, a monitoring cadence, and a contingency protocol. The resulting configuration changes are enforced through existing network function managers and slice orchestrators, preserving backward compatibility with deployed infrastructure. Importantly, the foundation model maintains an internal belief state of network conditions and its own uncertainty, which is exposed via an explainability interface that allows human operators to audit decisions, understand failure modes, and intervene before a degradation materialises. This human-in-the-loop posture is essential for building trust and for meeting the operator's regulatory obligations under frameworks such as the European Union's AI Act.

4. Cross-Domain Generalization and Robustness in Heterogeneous Environments

A primary motivation for adopting foundation models is their purported ability to generalise across domains without task-specific retraining, yet the translation from language and vision domains to the networking environment exposes unique challenges that must be confronted head-on. Network environments are highly non-stationary; traffic distributions shift due to changes in user behaviour, application updates, and mobility patterns, while the physical channel undergoes fading, shadowing, and interference that are inherently stochastic. The foundation model's pre-training corpus must therefore include a wide diversity of network

scenarios, including extreme events such as congestion collapse, distributed denial-of-service attacks, and equipment failures, to prevent degenerate behaviour under out-of-distribution conditions. The creation of such a corpus calls for a collaborative effort across operators and vendors, possibly mediated by standardisation bodies, to share anonymised telemetry and configuration traces [9].

Even with a rich pre-training corpus, there remains a non-trivial gap between simulation-trained agents and live deployments, a phenomenon well-documented in robotics and recently observed in networking. Foundation models can mitigate this simulation-to-reality transfer gap through two complementary mechanisms. First, their scale enables the emergence of robust internal representations that abstract away low-level simulator artifacts; a foundation model that has been pre-trained on hundreds of distinct traffic models, channel emulators, and network topologies is less likely to overfit to the peculiarities of any single simulator. Second, retrieval-augmented generation can bridge the remaining gap by incorporating real-time measurements into the decision context, effectively grounding the model's reasoning in empirical evidence. When the model encounters a novel service descriptor, such as a holographic communication slice with a previously unseen combination of throughput and jitter requirements, it can match the profile against a library of similar services, extrapolate the likely resource needs, and adapt its policy without gradient updates. This zero-shot composition capability represents a fundamental departure from the single-task DRL paradigm, where even minor variations in the reward function or state space often require retraining.

Robustness against adversarial inputs and data drift is a further concern that magnifies in the context of critical infrastructure. Recent studies have shown that small, carefully crafted perturbations to the input state of DRL agents can cause catastrophic policy failure, and a foundation model that absorbs a broader range of input modalities may present an even larger attack surface. Defensive strategies must therefore be embedded in the architecture from the outset. Input validation modules, such as consistency checkers that flag implausible combinations of key performance indicators, serve as a first line of defence. At inference time, the model can be instructed to produce an explicit confidence score along with its recommendation, allowing the orchestrator to fall back to a safe, conservative baseline when confidence falls below a threshold. Furthermore, the model's training pipeline should incorporate adversarial training and robust optimisation techniques to improve its resilience against distributional shifts. The long-term objective is to achieve what could be termed certified QoS assurance, where the system provides formal probabilistic guarantees that the QoS of critical slices remains within specified bounds under a defined set of perturbations, a goal that will require tight integration between the foundation model community and the formal methods community.

5. Governance, Fairness, and Ethical Policy Dimensions

The delegation of network resource allocation decisions to autonomous AI systems raises profound governance questions that extend beyond technical performance metrics. When foundation models are entrusted with dynamic QoS management, they implicitly make choices about how scarce bandwidth, compute cycles, and energy are distributed among competing service classes and end users. These choices can inadvertently encode or amplify societal biases if the training data reflect historical inequities in coverage, capacity, or service prioritisation. For instance, a model trained predominantly on data from dense urban deployments may systematically under-allocate resources to rural slices or to low-income user

groups whose traffic patterns differ from the majority. Ensuring algorithmic fairness in this context demands both technical interventions and institutional accountability mechanisms [10].

From a technical standpoint, fairness constraints can be incorporated into the foundation model's decision pipeline through several complementary approaches. Post-processing techniques can adjust the model's output to equalise certain QoS metrics across predefined groups, although this requires the operator to define protected attributes and group boundaries, a non-trivial task in a domain where users are ephemeral and granular demographic data are often unavailable. In-processing methods, such as fine-tuning the model with fairness-aware objectives, can encourage the model to discover resource allocation policies that satisfy group-level parity conditions while minimising overall QoS degradation. A more foundational direction involves auditing the pre-training data for representational biases and curating a corpus that adequately reflects diverse geographic, socioeconomic, and temporal conditions, including data from emergency scenarios where resource allocation must follow ethical triage principles rather than pure efficiency. The model card framework, originally proposed for documenting the capabilities and limitations of machine learning models, can be adapted to the networking domain to provide transparency about the training data composition, known failure modes, and fairness evaluation results [11].

Beyond fairness, the governance framework must address accountability and liability in the event of network failures caused by autonomous decisions. In current operational practice, human engineers bear responsibility for configuration errors, and service level agreements define contractual remedies for QoS violations. When a foundation model autonomously reallocates resources in a way that degrades a premium slice, attributing fault becomes complex, particularly if the model's reasoning is opaque. This motivates the development of strong audit trails that log every decision, the input context that prompted it, the retrieved examples used for augmentation, and the alternatives considered. Explainability methods, such as attention visualisations and concept-based explanations, can help operators trace the model's logic and identify the sources of error. Regulatory developments, including the European Union's AI Act, are beginning to mandate such transparency for high-risk AI systems, and autonomous network management platforms will likely fall within this scope, compelling the industry to embed governance by design rather than retrofitting it after widespread deployment.

The interplay between commercial incentives and network neutrality also emerges as a sensitive policy issue. A foundation model, if tuned solely to maximise operator revenue or minimise operational cost, could systematically deprioritise traffic from over-the-top providers that are not in a commercial partnership, thereby undermining the open internet principle. Policy guardrails, encoded as inviolable rules in the intent specification layer, can prevent such behaviour, but they require careful formulation and continual oversight by independent regulatory bodies. The shift toward intent-based governance, where regulators specify outcomes rather than prescribing technical mechanisms, may offer a path forward. In this model, a regulatory intent such as "all emergency service communications must maintain availability above 99.999%" is ingested by the foundation model and treated as a hard constraint that overrides all other objectives, while commercial optimisation intents are handled at a lower priority level. This hierarchical intent structure aligns technical autonomy with societal values, provided that the translation from high-level regulatory intent to operational configuration is auditable and contestable.

6. Deployment Pathways, Sustainability, and Infrastructure Considerations

The operational deployment of foundation model driven autonomous management systems introduces substantial infrastructure requirements that must be reconciled with the economic and environmental constraints of mobile network operators. The computational footprint of state-of-the-art large models, which can contain hundreds of billions of parameters, is incompatible with the power and thermal budgets of deeply embedded baseband units or even many edge data centres. Yet, recent advances in model compression and efficient inference have begun to close this gap, making it plausible to deploy compact, distilled variants that retain much of the generalisation ability while reducing floating-point operations per inference by orders of magnitude [12]. Techniques such as unstructured and structured pruning, knowledge distillation, quantisation-aware training to 4-bit or even 2-bit precision, and the use of mixture-of-experts architectures that activate only a sparse subset of parameters per input all contribute to shrinking the model size without catastrophic performance loss. When combined with specialised hardware accelerators, including neural processing units integrated into system-on-chips at the distributed unit, it becomes feasible to execute inference within the latency budgets required for medium-timescale control loops.

Sustainability considerations extend beyond per-inference energy consumption to the lifecycle carbon footprint of model training, fine-tuning, and continuous updating. The initial pre-training of a foundation model on a planetary-scale network data corpus is a computationally intensive undertaking that, if performed without renewable energy and carbon-aware scheduling, could impose a significant environmental burden. The networking community can learn from the “Green AI” movement that advocates for reporting energy and carbon metrics alongside traditional accuracy metrics [13]. Moreover, the federated nature of network operations offers a unique opportunity to distribute the training load: individual operators can train on their own data behind firewalls and share only model updates, aggregating knowledge without centralising gigabytes of sensitive telemetry. Federated fine-tuning of a common foundation model can dramatically reduce the total training energy by amortising the cost across many participants and by adapting a pre-trained model rather than training from scratch.

Lifecycle management of foundation models in production networks presents additional socio-technical challenges. Network deployments evolve continuously as new radio technologies are introduced, spectrum bands are repurposed, and software is upgraded. A model that is not periodically updated will suffer from concept drift and lose its predictive accuracy over time. Continuous learning pipelines, where the model is refreshed with recent data while retaining previously acquired knowledge, must be designed to prevent catastrophic forgetting. This requires careful orchestration of data versioning, model registry, A/B testing in shadow mode, and canary deployments that validate new model versions against rigorous QoS metrics before they are promoted to production. Standardisation bodies such as 3GPP and the International Telecommunication Union have begun to outline architectures for AI/ML model lifecycle management within the network management framework, specifying training, validation, and monitoring interfaces that can be adopted for foundation model updates [14]. The alignment of these standardised interfaces with the specific requirements of large pre-trained models, including support for adapter-based fine-tuning and retrieval-augmented inference, is an open area that deserves active industry-academia collaboration.

Interoperability with legacy operations support systems and business support systems is a pragmatic concern that will dictate the pace of adoption. Most operators maintain extensive

catalogs of rule-based policies, expert-written call flow scripts, and vendor-proprietary configuration models that represent decades of accumulated domain expertise. Forcing a wholesale replacement of these assets would be both prohibitively expensive and risky. The foundation model architecture must therefore provide a gradual migration path, where the model first operates in an advisory mode, proposing recommendations that are validated against existing rule engines, and only after a prolonged period of demonstrated safety is it empowered to act autonomously. In this advisory phase, the model can also ingest the outputs of legacy systems as additional context, learning to replicate or improve upon them over time. This co-evolution strategy reduces deployment risk and allows operators to build confidence in the cognitive plane while preserving their existing investments. The economic viability of the entire paradigm hinges on demonstrating that the capital and operational expenditure associated with deploying and maintaining foundation model infrastructure is offset by measurable reductions in manual intervention, faster troubleshooting, and improved asset utilisation that directly translate into higher user-perceived QoS and lower churn.

7. Conclusion

The integration of foundation models into autonomous network management marks a pivotal inflection point in the trajectory from 5G-Advanced to 6G networks, offering a path toward general-purpose cognitive control that transcends the brittleness of single-task learning agents. This paper has delineated a comprehensive system architecture that positions a multimodal foundation model as the reasoning core for intent-driven QoS optimization, operating across language, telemetry, and spatial modalities, and has examined the structural trade-offs among centralised, split, and edge-distributed inference configurations. The discussion of cross-domain generalisation, robustness, fairness, and governance underscores that the technical promise of foundation models cannot be divorced from the socio-technical fabric in which they will operate. Sustainable deployment demands aggressive model compression, carbon-aware training, and federated update strategies, while regulatory alignment calls for built-in transparency, human-in-the-loop oversight, and hierarchical intent structures that respect public policy values. Future work must advance the creation of open, representative network data corpora; develop certified safety and fairness guarantees for foundation model decisions; and establish lifecycle management protocols that are compatible with the 3GPP management architecture. As the mobile industry stands on the cusp of an AI-native era, the collaborative effort to harness foundation models responsibly will determine whether 6G networks can deliver on their promise of ubiquitous, resilient, and equitable connectivity.

References

1. Foukas, X., Patounas, G., Elmokashfi, A., & Marina, M. K. (2017). Network slicing in 5G: Survey and challenges. IEEE Communications Magazine, 55(5), 94–100.

2. Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y.-C., & Kim, D. I. (2019). Applications of deep reinforcement learning in communications and networking: A survey. IEEE Communications Surveys & Tutorials, 21(4), 3133–3174.

3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

4. Zhang, H., Li, Y., Medhi, D., Boutaba, R., & Gaspary, L. P. (2022). A survey on intent-based networking: Concept, framework, and applications. IEEE Communications Surveys & Tutorials, 24(3), 1851–1889.

5. Li, R., Zhao, Z., Sun, Q., Chih-Lin, I., Yang, C., Chen, X., Zhao, M., & Zhang, H. (2018). Deep reinforcement learning for resource management in network slicing. IEEE Access, 6, 76532–76543.

6. Li, Q. (2026). QoS Assurance Mechanism for 5G Network Slicing Based on the Deep Reinforcement Learning PPO Algorithm. arXiv preprint arXiv:2605.03345.

7. Li, X., Benzaïd, C., & Taleb, T. (2020). Towards zero-touch network and service management: The ETSI ZSM approach and its use cases. IEEE Communications Standards Magazine, 4(4), 40–47.

8. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., & Tang, L. (2017). Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. In Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems (pp. 615–629). ACM.

9. Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks, 32, 323–332. (Note: This is a landmark example of large-scale benchmarking; its methodological lessons apply to network data corpus creation.)

10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

11. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.

12. Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). A survey of model compression and acceleration for deep neural networks. IEEE Signal Processing Magazine, 35(1), 126–136.

13. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

14. 3GPP. (2021). Study on artificial intelligence (AI)/machine learning (ML) for NR air interface (3GPP TR 38.843 version 17.0.0). 3rd Generation Partnership Project.

15. Taleb, T., Samdanis, K., Mada, B., Flinck, H., Dutta, S., & Sabella, D. (2017). On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration. IEEE Communications Surveys & Tutorials, 19(3), 1657–1681.

16. Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

17. Mao, H., Netravali, R., & Alizadeh, M. (2017). Neural adaptive video streaming with Pensieve. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (pp. 197–210). ACM.

18. Liu, L., Özger, M., Xhafa, A., & Bennis, M. (2021). Knowledge transfer for collaborative edge intelligence in vehicular networks. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4446–4458.