The Pathway to Adaptive Lightweight AI Transceivers (*Vision Paper*)

Nimrod Glazer
School of ECE
Ben-Gurion University
Be'er-Sheva, Israel
nimrodgl1000@gmail.com

Nir Shlezinger
School of ECE
Ben-Gurion University
Be'er-Sheva, Israel
nirshl@bgu.ac.il

Tirza Routtenberg
School of ECE
Ben-Gurion University
Be'er-Sheva, Israel
tirzar@bgu.ac.il

Abstract

The growing complexity of modern wireless communication systems poses significant challenges to traditional receiver designs. Artificial intelligence (AI), and in particular deep learning, offers the potential to address these challenges by learning to infer from data without relying on explicit channel models. However, direct application of conventional AI techniques is ill-suited to the real-time, data-scarce, and hardware-constrained nature of wireless environments. This vision paper advocates shifting towards a *communication-oriented AI framework* as the foundation for autonomous and lightweight deep transceivers, explicitly designed for the sub-millisecond adaptation, minimal energy budgets, and high-reliability constraints of next-generation networks. Our approach highlights the importance of modular and Bayesian neural architectures, combined with rapid training techniques based on continual learning, asynchronous adaptation, and communication-aware data acquisition. Together, these elements pave the way toward future AI-empowered physical-layer systems that can operate efficiently, reliably, and autonomously in real time for next-generation wireless networks.

1 Introduction

The constant growth in data traffic places ever-increasing demands on wireless communication systems. Meeting these demands requires a paradigm shift toward the integration of multiple diverse technologies [1], including high-frequency communications [2], massive and holographic multiple-input multiple-output (MIMO) [3], novel antenna technologies [4], and reconfigurable surfaces [5]. However, these technologies also greatly complicate physical layer signal processing tasks, rendering traditional model-based algorithms increasingly challenging and, in cases, inadequate.

A promising direction to address these challenges is through integrating AI, and particularly deep learning tools, into the physical layer [6, 7]. The use of deep neural networks (DNNs) for wireless transceivers, referred to as *deep transceivers*, allows coping with settings where: (*i*) the lack of a reliable, tractable statistical model [8], and (*ii*) the absence of efficient algorithmic tools capable of handling complex signal propagation environments [9].

The Need for Adaptive Deep Transceivers Despite their potential, the integration of AI in the physical layer faces unique challenges not encountered in traditional AI domains such as computer vision or natural language processing [7, 10]. First, wireless channels are inherently dynamic, changing rapidly over time, often within milliseconds. This temporal variability requires continual adaptation to a shifting distribution. Second, processing must be performed under stringent latency and power constraints, on hardware-limited mobile devices. These two constraints fundamentally limit

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

the direct applicability of conventional deep learning strategies, which typically assume stationary data distributions and access to extensive computational resources.

The common practice in deep learning uses pre-trained DNNs. To handle dynamic channels, one can train a DNN across various channel conditions [11], yielding a non-coherent receiver, that sacrifices performance for generality. Performance can be improved by training multiple DNNs for different channels [12], estimating channel parameters as an additional input [13, 14], or using hypernetworks [15, 16]. Still, pre-trained DNNs struggle to generalize to unseen channel distributions without adaptation, and to achieve good performance, one typically needs complex architectures that can be unsuitable for embedded hardware within the required time frames. It is thus desirable to have *adaptive* AI, that re-tunes itself to the instantaneous conditions, with *lightweight* complexity.

Adaptive Lightweight AI Vision The prevailing paradigm of large DNNs pre-trained with vast datasets stands in stark contrast to real-world wireless environments. Achieving fast, reliable adaptation with compact models trained on sparse, noisy, and rapidly acquired online data requires a fundamental rethinking of how learning is performed. This involves redesigning not only the model architecture, but also the training procedures and the way data is collected and utilized. In this paper, we detail the ongoing pathway towards a novel AI framework for the design of *flexible, lightweight, and adaptive deep transceivers*, illustrated in Fig. 1. As opposed to importing architectures and training techniques from traditional AI domains, we propose a communication-oriented approach that addresses the unique constraints and structure of wireless systems. Our vision is outlined as follows: In Section 2, we advocate *modular* and *Bayesian* architectures [17] that are based on model-based processing chains [18]. Such architecture can be both expressive and compatible with hardware-efficient inference and rapid adaptation. In Section 3, we discuss training methods that leverage modular and Bayesian architectures to (i) allow *asynchronous learning*, i.e., identify when and which module needs adaptation [19]; and (ii) replace classical stochastic gradient descent (SGD)-based learning with Bayesian filtering [20], enabling rapid adaptation via continual updates [21]. In Section 4, we detail how training data can be accumulated in a communication-structure-aware manner [22–24] to support learning with reduced overhead. We conclude and discuss the road ahead in Section 5.

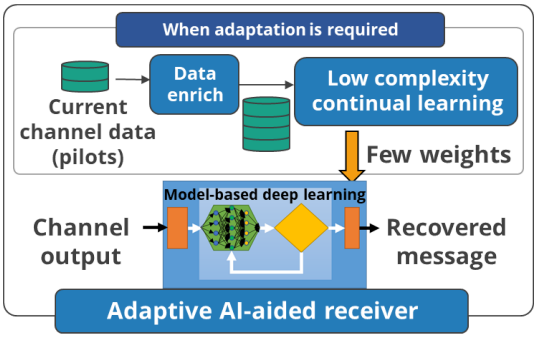

Figure 1: Flexible, Lightweight, AI Framework

2 AI Architectures for Deep Transceivers

Deep learning offers a powerful abstraction that can bypass explicit model assumptions and instead learn from data. However, realizing the full potential of deep learning for receivers necessitates addressing a critical tradeoff: while expressive models are needed to capture non-trivial dependencies, they must remain lightweight and amenable to rapid online adaptation. To this end, we advocate for a design approach based on three complementary principles:

A1 **Lightweight structures** that ensure compatibility with real-time, hardware-limited operation.

A2 **Modular architectures** that facilitate interpretability, scalability, and partial reuse

A3 **Bayesian modeling** that enhances calibration and improves learning from limited data.

To design architectures that are lightweight (A1) and modular (A2), while enabling to learn abstract mappings, we seek architectures that are based on classic processing via model-based deep learning methodologies [25]. We particularly focus on design approaches that augment traditional processing with deep learning tools, casting it as a trainable machine learning system [26]. We next exemplify this design approach through a running example of an uplink MIMO receiver.

Running Example: Bayesian DeepSIC Deep Soft Interference Cancellation (DeepSIC) is a modular DNN architecture for uplink MIMO systems [27]. It is based on the iterative soft interference

cancellation algorithm [28], which decodes each symbol by repeatedly subtracting the soft estimates of interfering symbols, while augmenting each decoding and interference-cancellation module with a compact DNN. The overall receiver unrolls the algorithm into Q iterations (viewed as layers), where in each layer q, the soft estimate $\hat{p}_k^{(q)}$ of user k is refined using a neural subnetwork. The resulting architecture can learn to operate in complex channel models, and supports *elastic inference* using only a subset of the modules, leading to faster processing when full model capacity is not required.

Despite its appearance as a collection of interconnected black-box modules, DeepSIC inherits its key benefits from its model-based roots: it is interpretable and scalable to varying numbers of users. Its combination with Bayesian DNNs [29], depicted in Fig. 2, brings several gains [17]: (i) while Bayesian learning rarely enhances performance compared to standard (frequentist) modeling, in wireless transceivers propagating distributions rather than point estimates results in better calibrated log-likelihood ratios, improving downstream channel decoding; (ii) combining modular and Bayesian DNNs facilitates rapid online training, as discussed in the sequel.

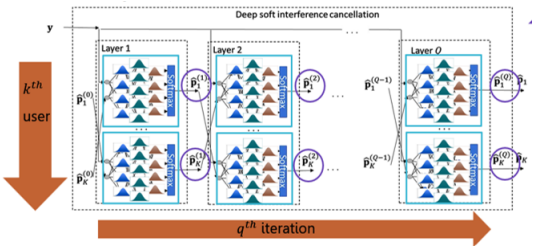

Figure 2: Bayeisan DeepSIC illustration, highlighting the output of each DNN module as a soft estimate

3 Training Methods for Deep Transceivers

A central challenge in realizing adaptive deep transceivers lies in enabling self-adaptation in real-time, and specifically on the order of less than one millisecond. As such, training must rely on limited data (e.g., pilots) and be executable on hardware-limited edge devices. To address this challenge, we envision a holistic learning framework composed of two mechanisms:

T1 Most importantly, **ultra-fast continual online learning** via **Bayesian tracking**.

T2 **Asynchronous online learning**, namely, train only when needed.

Mechanism T1 enabling real-time adaptation by departing from how DNNs are traditionally trained, i.e., using numerous lengthy gradient adaptations. Instead, it views of online learning as *state tracking*, such that learning can be done via one-shot Bayesian filters [30]. Specifically, it models the variations in the *desired* DNN parameters as a dynamic systems, such that training can be carried out using Bayesian tracking algorithms, such as the extended Kalman filter (EKF) [21]. Mechanism T2 is based on the observation that not all changes in the wireless channel necessitate retraining. In an asynchronous setting, dedicated drift detectors identify which modules need to be updated [19].Both T1 and T2 are most efficient when combined with A1-A3, as we show for our running example.

Running Example: Continual Online Learning of Bayesian DeepSIC Consider a DeepSIC receiver with $\boldsymbol{\theta}^{(k,q)}$ denoting the parameters of the (k,q) DNN module. Using T1, their evolution assumes to hold a Markov model $\boldsymbol{\theta}_t^{(k,q)} = \gamma \boldsymbol{\theta}_{t-1}^{(k,q)} + \mathbf{w}_t^{(k,q)}$ with $\mathbf{w}_t^{(k,q)}$ being Gaussian noise. The observations, derived from the module output from the received $\boldsymbol{r}_t$ and kth user pilots $\mathbf{p}_t^k$, yield a nonlinear state-space model $\mathbf{p}_t^{(k)} = \hat{\mathbf{p}}(\mathbf{r}_t; \boldsymbol{\theta}_t^{(k,q)}) + \mathbf{v}_t^{(k,q)}$, where $\hat{\mathbf{p}}(\cdot)$ denotes the (probabilistic) DNN output and $\mathbf{v}_t^{(k,q)}$ models noise. This formulation enables leveraging tools such as the EKF for online learning, updating the weight distribution in a single gradient and forward pass per block.

This powerful mechanism is made feasible by A1-A3: Due to *modularity* (A1), each subnetwork can be treated as an independent dynamic system, enabling per-module EKF updates that scale with user count and allow parallel or pipelined execution; The *lightweight* nature of the modules (A2) ensures that EKF-based updates are computationally tractable and hardware-compatible; and the *Bayesian modeling* (A3) is what enables EKF-based adaptation, which relies on Bayesian modeling of the tracked state. The modular Bayesian structure also greatly benefits T2, as it enables per-module monitoring and targeted adaptation. The Bayesian modeling allows drift detection without labeled data by examining the epistemic uncertainty (i.e., the variance of the predicted probability

distributions) [31]. When the epistemic uncertainty of a module increases beyond a threshold, it signals a potential mismatch with the current environment, prompting adaptation.

4 Data for Online Training of Deep Transceivers

The available data for training DNNs in the physical layer online primarily includes: (i) **Pilots**, which are known a priori and sparsely inserted in time and frequency. These allow for direct supervised training, but are limited in number and diversity due to spectral efficiency constraints. (ii) **Information payload**, which, while abundant, is not inherently labeled and thus require indirect strategies for use in training. While inherently constrained, these forms of data can be transformed into rich, task-relevant datasets suitable for online adaptation, with potential tools including

D1 **Self-supervision**, that uses information symbols for learning online;

D2 **Data augmentation**, enhancing an available data set,

D3 **Digital twins**, generating additional synthetic data for a characterized wireless settings.

Self-supervision (D1 leverages digital communication structures to obtain pseudo-labels, extending the training set beyond pilots. These can be done on the *codeword-level*, where the decoder's output is used for training if it passes error correction check [24]; or in the *symbol-level*, where individual symbols are used based on a confidence metric [32]. Data augmentation (D2) exploits the structure of modulation and the channel to enhance a small labeled data set, e.g., a set of received pilots, into a large data set by generating new synthetic samples [22, 23]. Digital twins (D3) are virtual replicas emulating propagation and hardware effects, which can be used to provide some level of real-time refinement (thought possibly to slow variations), by generating synthetic labeled samples adapted to current channel conditions [33]. While D1-D3 expand the volume and diversity of training data, they may also result in large datasets that increase training complexity. This can be alleviated by *active selection* of informative samples [34]. The overall envisioned data acquisition pipeline is illustrated in Fig. 3.

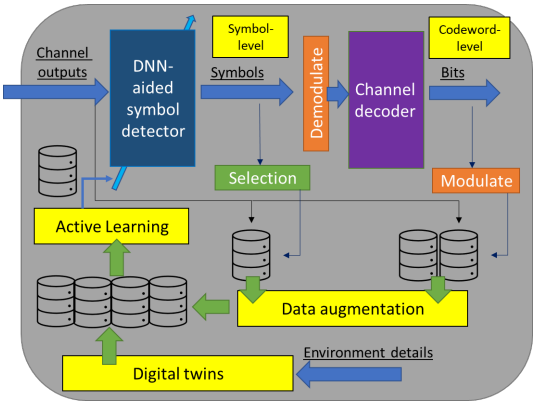

Figure 3: Data acquisition pipeline

5 Conclusions

We presented a holistic vision for *flexible*, *lightweight*, and *adaptive* AI transceivers. We argued for the use of modular and Bayesian architectures, combined with novel training methods using asynchronous adaptation, continual Bayesian learning formulated as tracking, and data acquisition pipelines. Our approach advocates a design paradigm shift that reflects on various aspects of wireless systems. For once, future wireless transceiver hardware need not rely on over-parameterized DNNs or high-end compute resources. Instead, the key enablers are (i) lightweight modules that can be updated locally; (ii) parallelism-aware architectures that allow pipelined or distributed processing; and (iii) hardware-level support for uncertainty-aware, online Bayesian learning. Such hardware need not be overly complex, as it needs to execute module-specific adaptation within short time frames, leveraging even limited parallel computation. However, there are still several key challenges in system-level integration of the components described herein into an operational system. These include devising efficient schedulers for adaptation under real-time constraints and controlling the interplay between inference and learning within tight latency budgets. Still, we conjecture that by reducing this gap sufficiently, one can realize fully autonomous AI-aided physical layer operation with targeted hardware co-design, realizing a new form of AI-native, real-time communication systems.

References

[1] M. Shahjalal, W. Kim, W. Khalid, S. Moon, M. Khan, S. Liu, S. Lim, E. Kim, D.-W. Yun, J. Lee *et al.*, "Enabling technologies for AI empowered 6G massive radio access networks," *ICT Express*, vol. 9, no. 3, pp. 341–355, 2023.

[2] W. Jiang, Q. Zhou, J. He, M. A. Habibi, S. Melnyk, M. El-Absi, B. Han, M. Di Renzo, H. D. Schotten, F.-L. Luo, T. S. El-Bawab, M. Juntti, M. Debbah, and V. C. M. Leung, "Terahertz communications and sensing for 6G and beyond: A comprehensive review," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2326–2381, 2024.

[3] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. Di Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Commun. Mag.*, vol. 27, no. 5, pp. 118–125, 2020.

[4] M. Ikram, K. Sultan, M. F. Lateef, and A. S. Alqadami, "A road towards 6G communication—a review of 5G antennas, arrays, and wearable devices," *Electronics*, vol. 11, no. 1, p. 169, 2022.

[5] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, 2021.

[6] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep learning for wireless communications: An emerging interdisciplinary paradigm," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 133–139, 2020.

[7] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 140–145, 2022.

[8] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, 2017.

[9] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, 2019.

[10] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, "Adaptive and flexible model-based AI for deep receivers in dynamic channels," *IEEE Wireless Commun.*, vol. 31, no. 4, pp. 163–169, 2024.

[11] J. Xia, D. Deng, and D. Fan, "A note on implementation methodologies of deep learning-based signal detection for conventional MIMO transmitters," *IEEE Trans. Broadcast.*, vol. 66, no. 3, pp. 744–745, 2020.

[12] T. Raviv, A. Goldman, O. Vayner, Y. Be'ery, and N. Shlezinger, "CRC-aided learned ensembles of belief-propagation polar decoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8856–8860.

[13] M. Honkala, D. Korpi, and J. M. Huttunen, "DeepRx: Fully convolutional deep learning receiver," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3925–3940, 2021.

[14] M. Goutay, F. A. Aoudia, J. Hoydis, and J.-M. Gorce, "Machine learning for MU-MIMO receive processing in OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2318–2332, 2021.

[15] G. Liu, Z. Hu, L. Wang, H. Zhang, J. Xue, and M. Matthaiou, "A hypernetwork based framework for non-stationary channel prediction," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8338–8351, 2024.

[16] T. Raviv and N. Shlezinger, "Modular hypernetworks for scalable and adaptive deep MIMO receivers," *IEEE Open Journal of Signal Processing*, vol. 6, pp. 256–265, 2025.

[17] T. Raviv, S. Park, O. Simeone, and N. Shlezinger, "Uncertainty-aware and reliable neural MIMO receivers via modular Bayesian deep learning," *IEEE Trans. Veh. Technol.*, 2025.

[18] N. Shlezinger and Y. C. Eldar, "Model-based deep learning," *Foundations and Trends® in Signal Processing*, vol. 17, no. 4, pp. 291–416, 2023.

[19] N. Uzlaner, T. Raviv, N. Shlezinger, and K. Todros, "Asynchronous online adaptation via modular drift detection for deep receivers," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 4454–4468, 2025.

[20] M. Jones, P. Chang, and K. Murphy, "Bayesian online natural gradient (bong)," *Advances in Neural Information Processing Systems*, vol. 37, pp. 131 104–131 153, 2024.

[21] Y. Gusakov, O. Simeone, T. Routtenberg, and N. Shlezinger, "Rapid online Bayesian learning for deep receivers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[22] L. Huang, W. Pan, Y. Zhang, L. Qian, N. Gao, and Y. Wu, "Data augmentation for deep learning-based radio modulation classification," *IEEE Access*, vol. 8, pp. 1498–1506, 2019.

[23] T. Raviv and N. Shlezinger, "Data augmentation for deep receivers," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8259–8274, 2023.

[24] M. B. Fischer, S. Dörner, S. Cammerer, T. Shimizu, H. Lu, and S. Ten Brink, "Adaptive neural network-based OFDM receivers," in *IEEE Signal Processing Advances in Wireless Communications (SPAWC)*, 2022.

[25] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, "Model-based deep learning: On the intersection of deep learning and optimization," *IEEE Access*, vol. 10, pp. 115 384–115 398, 2022.

[26] N. Shlezinger and T. Routtenberg, "Discriminative and generative learning for linear estimation of random signals [lecture notes]," *IEEE Signal Process. Mag.*, vol. 40, no. 6, pp. 75–82, 2023.

[27] N. Shlezinger, R. Fu, and Y. C. Eldar, "DeepSIC: Deep soft interference cancellation for multiuser MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1349–1362, 2021.

[28] W.-J. Choi, K.-W. Cheong, and J. M. Cioffi, "Iterative soft interference cancellation for multiple antenna systems," in *Proc. IEEE WCNC*, 2000.

[29] V. Fortuin, "Priors in Bayesian deep learning: A review," *International Statistical Review*, vol. 90, no. 3, pp. 563–591, 2022.

[30] S. Farquhar and Y. Gal, "A unifying Bayesian view of continual learning," *arXiv preprint arXiv:1902.06494*, 2019.

[31] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[32] R. Finish, Y. Cohen, T. Raviv, and N. Shlezinger, "Symbol-level online channel tracking for deep receivers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8897–8901.

[33] J. Hoydis, F. Aït Aoudia, S. Cammerer, F. Euchner, M. Nimier-David, S. Ten Brink, and A. Keller, "Learning radio environments by differentiable ray tracing," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 1527–1539, 2024.

[34] I. Be'Ery, N. Raviv, T. Raviv, and Y. Be'Ery, "Active deep decoding of linear codes," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 728–736, 2019.