IEEE Communications Magazine, Volume 58, Issue 6, June 2020, pp. 39-45
DOI:10.1109/MCOM.001.2000050

Explainable Artificial Intelligence (XAI) for 6G: Improving Trust between Human and Machine

Weisi Guo *IEEE Senior Member, RSS Fellow*

***Abstract*—As the 5th Generation (5G) mobile networks are bringing about global societal benefits, the design phase for the 6th Generation (6G) has started. Evolved 5G and 6G will need sophisticated AI to automate information delivery simultaneously for mass autonomy, human machine interfacing, and targeted healthcare. Trust will become increasingly critical for 6G as it manages a wide range of mission critical services.**

As we migrate from traditional mathematical model-dependent optimisation to data-dependent deep learning, the insight and trust we have in our optimisation modules decrease. This loss of model explainability means we are vulnerable to: malicious data, poor neural network design, and the loss of trust from stakeholders and the general public; all with a range of legal implications. In this review, we outline the core methods of Explainable Artificial Intelligence (XAI) in a wireless network setting, including: public and legal motivations, definitions of explainability, performance vs. explainability trade-offs, and XAI algorithms. Our review is grounded in cases studies for both wireless PHY and MAC layer optimisation and provide the community with an important research area to embark upon.

***Index Terms*—machine learning; deep learning; deep reinforcement learning; XAI; 5G; 6G;**

I. Introduction

An essential fabric of modern civilization is the digital economy, which is underpinned by wireless networking. We are on the cusp of entering a new era of mass digital connectivity enabled autonomy. An increasing number of people, machines, and things are being connected to automate and digitise traditional services. Wireless networking has transitioned from its traditional role as an information channel (1G to 3G) to a critical lever in the new industrial revolution of automation (5G and beyond to 6G [1]). It is envisaged that by 2030, 6G services require 1000× data rate and manage diverse service requirements such as massive ultra-reliable low latency communication (M-URLLC) for control of autonomous entities across transport to precision manufacturing.

Orchestrating co-existence via spectrum aggregation between different radio access technologies (RATs) is essential to meeting this demand. As such, real-time complex radio resource management (RRM) is critically important with strict guarantees. However, this has become too complex for conventional optimisation. As such, there is a global push for Artificial Intelligence (AI) driven information ecosystems [2] to support more fine-grained user-centric service provision (see 3GPP Release 16 TR37.816). Recent research on the application of AI in 5G PHY and MAC layers can be found in IEEE ComSoc Best Readings in Machine Learning in Communications.

A. AI and Trust

As communication systems increase complexity, Deep Learning (DL) in the popular form of Deep Neural Networks (DNNs) is set to transform both PHY layer (e.g. blind signal detection in nonlinear channels) and MAC layer (e.g. rapid power control for massive MIMO) modules. In this new era of complexity explosion, previous model-based optimisation lack either explicit mathematical models or do not have the processing time to calculate heuristic solutions. DNN presents an excellent opportunity to transform complex data-rich problems into solutions.

An open challenge with DNN is the lack of transparency and trust compared to traditional mathematical model-based optimisation. Neural networks (NN) with multiple layers cannot explain the essential features that influence actions, nor the impact of data bias on the uncertainty of outputs. Beyond supervised learning for PHY layer signal detection, DNN is especially opaque when coupled with reinforcement learning (RL) [3], where the Markov Decision Process (MDP) is integrated with hidden layer dynamics. As such, there is the need to develop explainable algorithms that can quantify uncertainty, especially mapping data inputs, algorithm design, to the projected wireless key performance indicators (KPI). A trustworthy AI should be able to explain its decisions in some way that human experts can understand (e.g. the underlying data evidence and causal logic). Understanding both our opportunity and vulnerability to deep learning is essential to the success of future wireless services.

B. Novelty & Organisation

In this review, we outline the core concepts of Explainable Artificial Intelligence (XAI) for future wireless systems:

1) Section II-A: Public and legal motivations for improving the transparency and trust in AI algorithms;
2) Section II-B: Definitions of explainability from specific quantitative indicators to general qualitative outputs;
3) Section III: Review of current deep learning techniques in PHY and MAC layer and their level of performance vs. explainability trade-off;
4) Section IV: Technical methods to improve explainability in deep learning;
5) Section V: Summary of open challenges;

Weisi Guo is with the Alan Turing Institute, British Library, 96 Euston Rd, London, NW1 2DB, United Kingdom; and Cranfield University, College Road, Bedford, MK43 0AL, United Kingdom. *Corresponding Author: wguo@turing.ac.uk. The author wishes to acknowledge EC H2020 grant 778305: DAWN4IoE - Data Aware Wireless Network for Internet-of-Everything, and The Alan Turing Institute under the EPSRC grant EP/N510129/1.

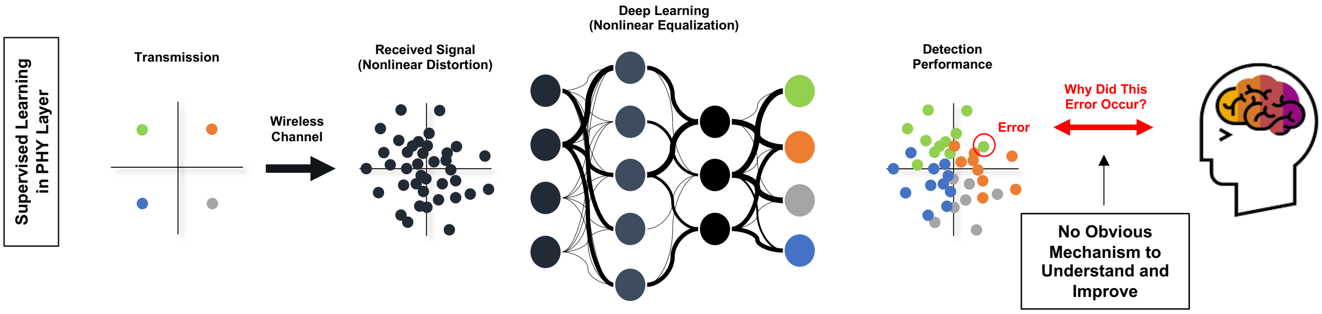

Supervised Learning in PHY Layer
Transmission
Wireless Channel
Received Signal (Nonlinear Distortion)
Deep Learning (Nonlinear Equalization)
Detection Performance
Error
Why Did This Error Occur?
No Obvious Mechanism to Understand and Improve

Fig. 1. Example of deep learning applications in supervised learning of equalisation with nonlinear symbol distortion.

Our review is grounded in cases studies for both PHY and MAC layer optimisation, including examples of explainability in existing algorithms.

II. Motivation and Definitions of XAI

A. *Public Trust & Legal Frameworks*

At the heart of our need to add explainability / interpretability to DNNs is the need to build trust in a quantifiable way. Traditional mathematical model-based algorithms have reasonably high clarity in how a *mathematical model* and the *input data* leads to *output decisions*. For example, water-filling (WF) power allocation shows clearly how the Lagrangian multiplier transforms input channel gains to output power allocation solution. Whilst DNNs can accelerate the optimisation time and often the accuracy, they remain opaque and doesn't tell us the impact of input data and bias on decisions, the reasoning for decisions, and how the DNN logic can reverse teach human experts.

Beyond the technical requirements, the *legal framework* for AI is still in its infancy, and there are several explicit requirements for XAI in different regions, such as EU GDPR requires machine learning algorithms to be able to explain their decisions (see Recital 71), or that the French Digital Republic Act requires transparency in the degree and mode of algorithms that contribute to decisions, the data used and its provenance, the weight of different data features, and the resulting actions. The key is that rightly or wrongly, humans can attempt to explain if prompted to, and we need machines to have that equal capability in order to ensure trust and a legal pathway towards improving safety and reliability.

B. *Definitions and Modes of Explainability*

In classic wireless systems, explicit models seek to map inputs to outputs and when models are well known, Bayesian inference outperforms deep learning (DL). In absence of models, DL attempts to automatically construct high-dimensional non-linear models based on data. Whilst some DL models can be interpretable (e.g. deep random forests and decision trees), the most scalable deep learning algorithms (DNNs) lack explainability.

An intuitive and good starting point for *explainability* is for it to meet two conditions:

1) Prediction is correct
2) Prediction is based on the correct data features and logic

The latter is much harder to define numerically, let alone implement alongside a DNN framework. This is particularly challenging when we are dealing with DRL, large input data sets, and multiple hidden layers – we will discuss these aspects later in the paper. For now, we discuss the different modes of explainability, with an illustration in Fig. 3.

1) Visualisation: The simplest form are visual outputs from the DL algorithm highlighting data features that causally lead to the output choice (e.g. DeepLIFT [4]). This may or may not map to the human perceptions of key features which also contribute to our cognitive reasoning. When combined with well known case studies, whereby the input and output mapping is established, we can both satisfy that predictions are correct and it is likely the human operator can easily accept or reject the key visual features.

2) Hypothesis Testing: A more rigorous form of the aforementioned is hypothesis testing, whereby a well formulated argument is tested based on the input data and output decision. Here, we can test if: i) certain key features are important in the mapping, ii) the mapping function behaves as we expect (monotonic, nonlinear, ...etc.), and iii) we can accept or reject the hypothesis.

3) Didactic Statements & Symbolic Representation: Perhaps the ultimate form of explainability would use natural language or mathematical models to communicate to the human operator, explaining what data features and reasoning led to a decision/output. The metrics considered in natural language processing (NLP) would range from the repressiveness and accuracy of a n-gram linguistic output, the brevity penalty of short communications, and many of the metrics are universalised under the Bilingual evaluation understudy (BLEU) framework. Mathematical algebraic expressions of the NN's actions will require flexible functions to explain the NN's mapping, such as hyper-geometric functions. Both will require intuitive machine-human interfaces to explain the learning and decision process.

C. *Metrics of Explainability*

There are several metrics that can be used to quantify the accuracy of explainable models: i) the accuracy or representativeness of the local model (e.g. polynomial fit or sensitivity analysis at a neuron in DNN) or global model (e.g. generalised

TABLE I
AI EXAMPLES IN WIRELESS COMMUNICATION

Problem Domain	Representative Paper	Classic Approach	ML or DL Approach	Improvement at BER	Explainability
Signal Detection	Ye18 (WCL)	DFT with LS or MMSE	DNN with 3 hidden	>15dB at 10^{-1}	Low
Channel with Memory	Farsad18 (TSP)	Viterbi Detector (VD)	SBRNN with 1 hidden	20 VD mem. at 10^{-1}	Low
Decoding of LDPC	Nachmani18 (JSTSP)	Belief Propagation (BP)	RNN with 5 hidden	1dB at 10^{-3}	V. Low
Channel Estimation	Neumann18 (TSP)	Orth. Matching Pursuit	CNN with 1 hidden	2dB at 10^{-1}	Low
NOMA SCMA Detection	Kim18 (CL)	Message Passing	DNN with 4 hidden	2dB at 10^{-3}	V. Low
Channel Est. mm-M-MIMO	He18 (WCL)	Support Detection	CNN and 3 layers	17dB at 5dB SNR	V. Low
Cognitive Radio	Tsakmalis18 (JSTSP)	Expectation Prop.	Bayesian MCMC	25 flops at 10^{-1} error	Medium
Power Allocation	Nasir19 (JSAC)	Frac. Prog. & WMMSE	DQN‘with 3 hidden	1bps/Hz	None
Cross RAT Channel Access	Yu19 (JSAC)	RL	DQN with 6 hidden	5% rate	None
Interf. Align with Cache	He17 (TVT)	RL	DQN with 4 hidden	20% rate	None
Antenna Sel.	Joung16 (CL)	MaxMinNorm	SVM	5% at 10^{-1}	Low
WSN Diagnostics	Liu10 (TON)	Clustering	Bayesian Belief Net.	5%	Medium
User Behaviour Recog.	Wang10 (TMC)	SVM	Random Forest	2-6%	Low
QoE of Multimedia	Hameed16 (TM)	Fixed	Decision Tree	50% overhead	High

TABLE II
METHODS AND METRICS FOR XAI APPROACHES

XAI Approach	Method	Relevant Measures	Application Areas
Feature Sensitivity	DeepLIFT feature analysis [4]	Variogram (VARS)	RRM: impact of input states on action
Accept or Reject Null Hypothesis	Bayesian or Frequentist	p-value	Cell Planning: inclusion of social factors
Local Fitted Model	Local Linear Model (LIME) [5]	Coeff. of Determination	Optimisation: discover input interactions
Global Fitted Model	Meijer G [6], B-spline	Coeff. of Determination	Optimisation: model discovery
Physics Informed Model	Surrogate Twin (PhyML) [7]	Loss, Confusion Matrix	Channel non-linear equalisation
Reduced MDP Model	State Reduction [8]	computational complexity	RRM & Optimisation
Reduced Neural Network Model	Pruning [9]	Loss, Confusion Matrix	RRM on Mobile devices
Didactic Statements	Natural Language Processing	n-gram: precision, brevity, BLEU	AI to engineer interface

hyper-geometric function fit across whole DNN), ii) the performance of an explainable physics informed DNN, and iii) the computational complexity cost of the explainable models. The approaches which we explain above are summarized with their metrics and KPIs in Table II below.

III. DEEP LEARNING IN WIRELESS: EXPLAINABILITY VS. PERFORMANCE

A. Review of Deep Learning & Wireless Applications

1) PHY Layer: Supervised DL has a wide range of applications in the PHY layer. In signal detection, it can equalise non-linear distortions by feeding the received signals corresponding to transmit data and pilots [10], outperforming classic MMSE approaches - see example in Fig. 1. When channels have memory, a bidirectional recurrent neural network (RNN) is more suitable and does not require channel state information (CSI), out performing Viterbi detection [11]. Similar approaches for mm-Wave Massive MIMO, and end-to-end channel estimation have also been performed – a summary of their performances is given in Table I, along with their reported performances and potential level of explainability.

2) MAC Layer: In MAC layer RRM, classic reinforcement learning based solutions do not rely on accurate mathematical models. Whilst this overcame the challenges faced by traditional model dependent optimisation, the Q-table used in RL cannot scale to more complex problem sets such as coordinated multi-BS offloading to heterogeneous devices, and will lead to non-convergence and a high computational delay. Deep RL (DRL) relies on the powerful function approximation and representation learning properties of DNN to empower traditional RL with robust and high efficiency learning. In Fig. 3, we demonstrate an example of offloading user traffic based on observed state, and reward inputs. This in turn is translated into a reward distribution over possible actions and an action is selected. In the next time iteration, the consequence of those actions are observed.

A summary of DRL performance gains is given in Table I, along with their reported performances and potential level of explainability. Currently, most existing DRL solutions applied in RRM use off-the-shelf algorithms with little consideration on the RRM feature set and DRL design. This means that the resulting benefit and penalties incurred (e.g. latency and energy consumption) cannot be understood by the radio engineers monitoring and configuring the network. In order to achieve a trusted autonomy, the DRL agents have to be able to explain its actions for transparent human-machine interrogation.

B. Trade-off and Interpretation Bias

In Fig. 2 we show a illustrative mapping of AI algorithms reviewed in Table I. There is an intuitive trade-off between explainability and performance when the **mathematical model is not known**. In the case of known or good models, DL cannot outperform classic statistical / signal processing methods - this is a mute point. When it is not known, as is the case for many complex systems, DL improves the performance at the cost of explainability. Whilst the DNN's

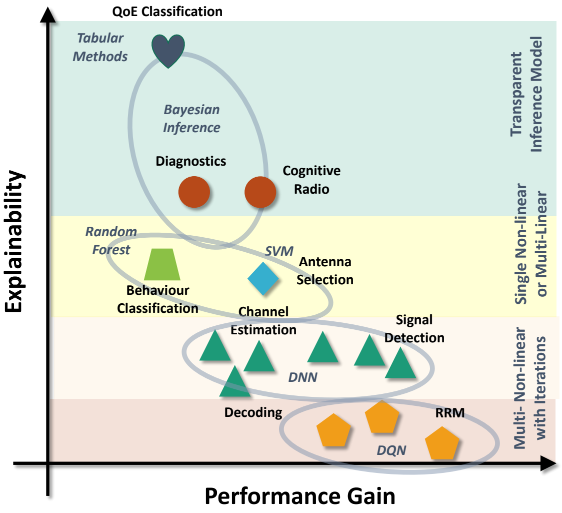


Fig. 2. Trade-off between AI performance gain and explainability with a variety of PHY and MAC layer examples. Trade-off exists when there are no or poor explicit models.

performance in complex model-free problems is superior to the aforementioned Bayesian and classic non-linear techniques, its bias to data input bias is well documented but not well understood. First, it maybe intuitive to think that the weights connecting units may reveal insight (partial explainability) to its high performance. However, DNNs learn mapping in a discontinuous way. As such, adding purposefully designed input data noise (with no explainable features) into a well established classifier can lead to severe mis-classification. This area of adversarial deep learning remains an open challenge which we discuss more at the end of the paper.

1) Bayesian Methods: Here, we can see that Bayesian techniques (of which tabular and decision trees methods can also fit into) have a *high degree of explainability*, transparently mapping data evidence (marginal) to model parameter estimation to output confidence distribution (posterior). Even when Bayesian inference is problematic, we tend to understand why [12], e.g. when:

1) the number of expected outcomes is large, e.g. too many power control levels or input modulation possibilities (e.g. 256 QAM)
2) a large number of marginals of the data-generating distribution are unknown (e.g. unknown mobility speed distribution amongst a range of vehicles)

We also know how this affects outputs: (i) two sets of data from the same situation may appear completely different and lead to different decisions, or (ii) small changes in the model parameters or data (its prior) can cause a different posterior conclusion. We detail more on data and algorithm bias below.

Decision trees and random forests also have good explainability, even deep ones (deep random forests and deep decision trees), the reasoning behind how the tree is formed is less clear compared to Bayesian model based methods. Furthermore, RF finds the optimal decision tree, but is often vulnerable to random permutation in out-of-bag (OOB) samples, otherwise known as Mean Decreased in Accuracy (MDA).

2) Non-Linear Methods: As we move away from the Bayesian framework, non-linear classification techniques such as Support Vector Machine (SVM) quickly lose explainability and there is no clear reason why data leads to one type of classification nor do we understand how over-fitted it is. DQNs stack several layers of non-linear activation functions and the explainability of a DQN from either a model transparency or a reasoning transparency perspective is not clear. Furthermore, the problem of sample bias and overfitting is further exasperated when we use DQNs to resolve a wide range of signal detection and channel estimation problems. The explainability is further reduced in DRL, whereby we further complicate the explainability surrogate model, reaching almost zero explainability in the DRL naive form.

IV. Methods to Improve Explainability

Here, we give a review of recent attempts to improve explainability in DNNs. To motivate the reader, we given an example of RRM in a 5G UAV setting [13]. Whilst UAVs are already helping to improve 5G networks, building explainable trust between the coordination modules and human operators is critical in 6G. As shown in Fig. 4, a UAV small-cell can fly between different service regions as well as recharge. At each service region it performs power allocation over a large number of parallel OFDM channels. To achieve real-time optimisation, a DNN is used to approximate the classic iterative Water-Filling (WF) power allocation solution, whilst a Double Dueling Deep Q-Network (DDDQN) is used to approximate the MDP for the UAV's flight actions. We map the previous and following XAI methods to the aforementioned wireless communication context. A summary of the methods listed below, their metrics for performance and potential applications in wireless communications is given in Table II.

A. Symbolic Representation

A mathematically rigorous form would be to find the most likely or the precise form of mapping performed by DNN, as a function of the NN's weights and activation functions. There are a number of approaches, including using the generic Meijer G-function [6], or Fox H-function. Meijer G-function is a general hyper-geometric function intended to include most known special functions and classes. As such, it provides a flexible framework to discover the mapping between input variables and output solution. In the example in Fig. 4, the neural network $f(\cdot) \in F$ can rapidly map input channel gains to output power allocation without the iterative search of classic WF. In order to verify that the solution mapping doesn't yield unexpected results, we map $f(\cdot) \rightarrow g(\cdot)$, where $g(\cdot) \in G$ is a hyper-geometric function. The end result is that a strong match is achieved and an analytic equivalent solution to iterative WF is found.

B. Feature Visualisation Techniques

At the perhaps most intuitive level of explainability, one can post-hoc visualise the features that are important based on their

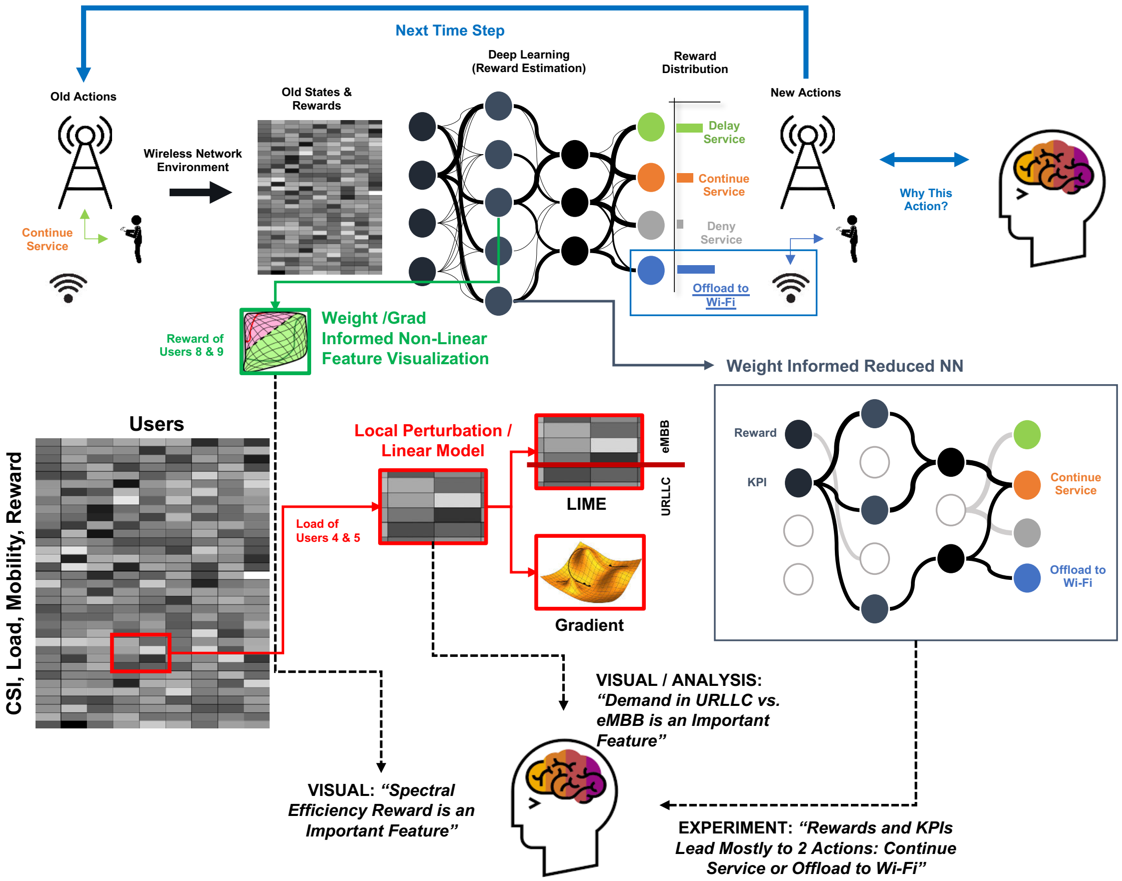

Next Time Step
Deep Learning (Reward Estimation)
Reward Distribution
Old Actions
Old States & Rewards
New Actions
Wireless Network Environment
Continue Service
Delay Service
Continue Service
Deny Service
Offload to Wi-Fi
Why This Action?
Reward of Users 8 & 9
Weight /Grad Informed Non-Linear Feature Visualization
Weight Informed Reduced NN
Users
CSI, Load, Mobility, Reward
Local Perturbation / Linear Model
Load of Users 4 & 5
eMBB
URLLC
LIME
Gradient
Reward
KPI
Continue Service
Offload to Wi-Fi
VISUAL / ANALYSIS: "Demand in URLLC vs. eMBB is an Important Feature"
VISUAL: "Spectral Efficiency Reward is an Important Feature"
EXPERIMENT: "Rewards and KPIs Lead Mostly to 2 Actions: Continue Service or Offload to Wi-Fi"

Fig. 3. Explainability examples in DRL: DRL without explainability coupled with surrogate models with a range of explainability options using data features and compressed neural network (NN).

weights or gradients of local nodes in the NN after training. In a gradient based approach, we calculate the gradient of each input feature with respect to an output, where a small change in the input data feature leads to the level of outcome change can be visualised. Using the example in Fig. 4, we implement a DDDQN reinforcement learning model and highlight the impact of different state features on the resulting UAV actions. The weights for different state values highlight that certain features such as battery power b and load satisfied l are more important than other factors. One challenge is that local features in hidden layers are non-linear and therefore the interpretation maybe not trivial. This explainability process can be further enhanced by yielding didactic statement explanations by layer-wise relevance propagation (reversing the NN by weight importance).

C. Local Machine Learning Model Reduction

Instead of reducing the global DL machine learning model, we can also create simpler surrogate models of selected partial data. For example, we can select only the load demand data (see states in Fig. 3 to see how this input feature affects the output. In general, one attempts to identify one or a set of interpretable model (such as the interpretable linear models, decision trees, rule tables discussed previously) that is locally faithful to the classifier in question [5].

We can also create local explainable surrogate models to understand better what DL is doing. In Fig. 3, we can see that the load of users 4 & 5 influence action choice and can be local linearly divided between the URLLC and eMBB load demand - and this output can be either visual or quantitative analysis. One popular approach based on the above logic is called Local interpretable model-agnostic explanations (LIME) [14]. LIME introduces a measure of complexity such that one attempts to find the most faithful local explainable model with the smallest complexity. As such, in our case in Fig. 3, LIME has quantified that the linear model that divides URLLC and eMBB demand is more explainable and less complex than a higher order polynomial model.

D. Physics Informed Design

Designing DL algorithms that are physics based can negate many of the concerns, as they have direct explainability. For example, equalising the nonlinear channel loss (e.g. a multitude of dispersion and phase noise in NLSE channels) is traditionally achieved via digital back propagation methods such as Split-Step Fourier Method (SSFM). Designing DNN that approximates this process in the form of a Learned Digital Back Propagation (LDBP) is achieved by unrolling the SSFM iterations and approximating each span inversion with 2 layers [7]. However, in many cases, this is not possible because we

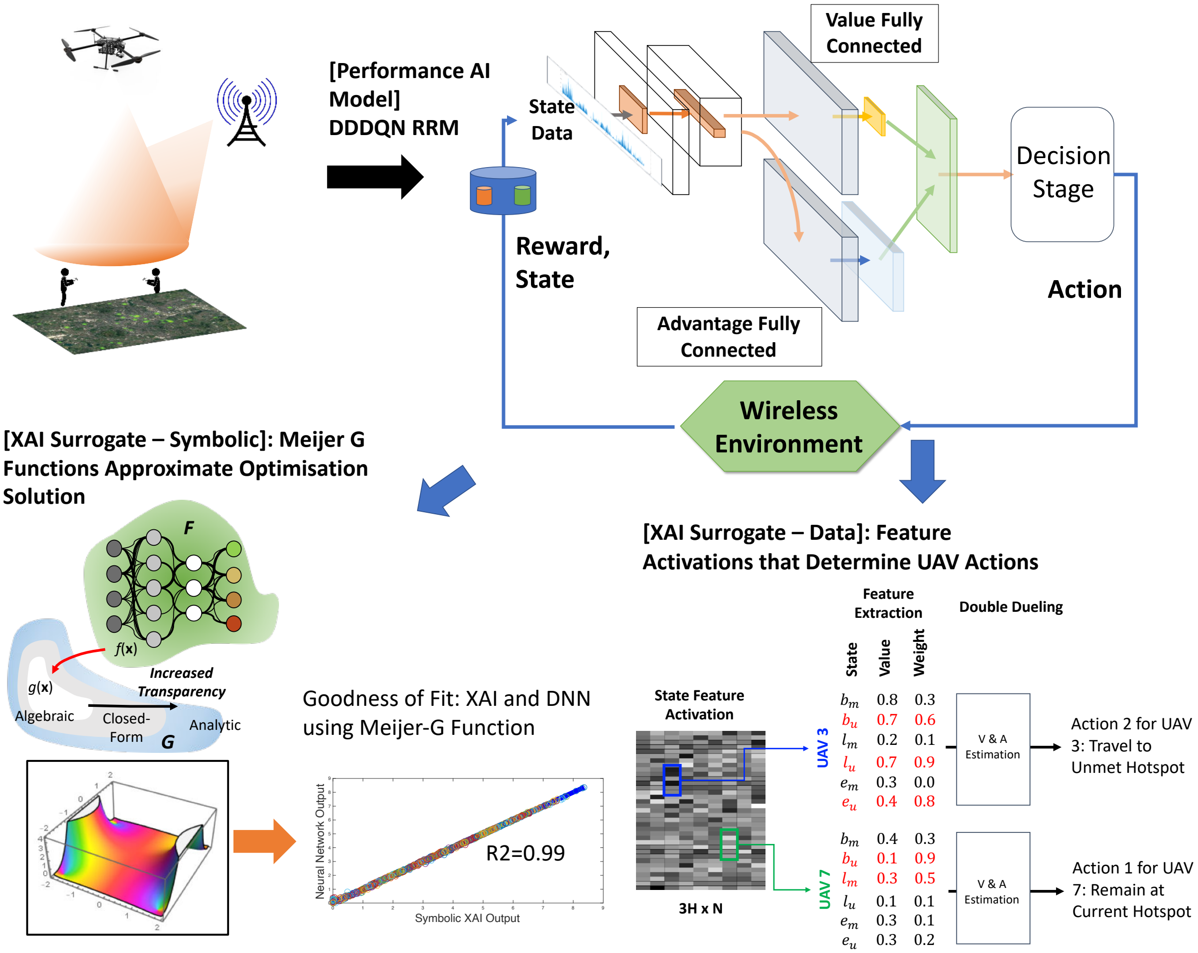


Fig. 4. XAI Integration demonstration with UAV enabled coverage - offloading data demand from BS, Performance AI: Double Dueling Deep Q Network (DDDQN) implementation, and XAI surrogates: (bottom left) global symbolic model mapping using hyper-geometric functions that map input state values algebraically to the output actions; (bottom right) data driven feature activation map of how state weights cause actions.

lack a workable traditional mathematical model or that it has unsatisfactory performance.

E. Global Machine Learning Model Reduction

Since we know that simpler machine learning models are more likely to be explainable, e.g. fewer parts to link mathematically, more likely to be in a form we recognise, ...etc., and as such model reduction makes sense. There are a multitude of ways in which this can be achieved with varying results and we detail some, but not all approaches below. Reduced models are particularly useful for reducing the long term energy expenditure of DNN algorithms, which is of benefit to mobile devices.

1) Problem Reduction: In reinforcement learning, the framework is often formulated from a Markov Decision Process (MDP). The size of MDP is directly determined by the state and action spaces, which grow super-polynomially with the number of variables that characterise the domain. To support fine-grained RRM, we have to adopt high-resolution communication context to accommodate context-aware optimization, which often results in a large-scale Partially Observable MDP (POMDP). The worst-case complexity is determined by the model, ranging from POMDP with PSPACE-complete (polynomial to input) to PO Stochastic Games with NEXP-complete (non-deterministic Turing machine) complexity. In general, one can compress MDP model in two stages:

- MDP model construction: one can appropriately choose the definitions of state and/or action to adjust their resolution. For example, when the transmit power constitutes the action space, we could use a limited number of discretised levels to approximate their dynamic range with controlled performance loss. Example: hierarchical action space methods can be used to approximate the POMDP problem, achieving a scalable compression.
- During learning: the size of MDP model can be further reduced by aggregating identical or similar states, allowing us to reduce learning complexity with a bounded loss of optimality [8]. The similarity of states can be measured in terms of optimal Q function, reward and state transitions, Boltzmann distributions on Q values, *etc.*.

2) Neural Network Reduction: Previous studies have revealed that NNs are typically over parameterised [9], and one can achieve similar function approximation by removing components (e.g. pruning the network as shown in Fig. 3

and only retaining useful parts with greatly reduced model size. There are several typical ways on compressing DNN by exploiting sparsity in NN:

- Reducing the number of parameters: removing the number of connections/weights, or pruning filters.
- Architectural reform: replacing fully-connected layers with more compact convolutional layers.
- Weight quantization: reduce the bit width integer.

In general, selecting appropriate local data or reducing the global model also gives extra explainability power by developing experiential and example-based explanations.

V. Challenges and Conclusions

In the context of Beyond 5G and 6G, the main areas that require improved trust are mainly in automation for transport, precision manufacturing, healthcare, and human machine brain interface. I believe there are three main multi-disciplinary areas for 6G. (1) **Human Machine (Brain) Interface:** developing rational and intuitive interfaces that communicate (e.g. didactic statements, interactive visual) to users and engineers. The recent advances in 6G human-brain interfacing [1] for tactile control and shared intelligence presents a futuristic framework for XAI. **XAI Twin:** develop an explainable twin AI system to work in parallel to the DL systems that are designed for optimisation performance. Recent work to develop a Neuro-Symbolic Concept Learner (NS-CL) agent that mimics human concept learning, able to translate back to the language description of the features [15]. **Adversarial AI:** Develop defence mechanisms that can recognise targeted attacks against DL and XAI engines.

As 6G will need to enable greater levels of safety-critical autonomy across a wide range of industries, building and quantifying trust between human end-users and the enabling AI algorithms is legally imperative. At the moment, we simply don't understand a wide range of deep learning (DL) modules that contribute to PHY and MAC layer roles. In this review, we outlined the core concepts of Explainable Artificial Intelligence (XAI) for 6G, including: public and legal motivations, definitions of explainability, performance vs. explainability trade-offs, methods to improve explainability, and proposed a framework to incorporate XAI into future wireless systems. Our review has been grounded in case studies for both PHY and MAC layer optimisation and provide the community with an important research area to embark upon.

References

[1] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, 2020.

[2] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," *IEEE Wireless Communications*, vol. 24, pp. 175–183, Oct. 2017.

[3] N. Jiang, Y. Deng, A. Nallanathan, and J. Chambers, "Deep reinforcement learning for real-time optimization in nb-iot networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1424–1440, 2019.

[4] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3145–3153.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[6] A. Alaa and M. Schaar, "Demystifying black-box models with symbolic metamodels," in *NIPS*, 2019.

[7] C. Häger and H. D. Pfister, "Nonlinear interference mitigation via deep neural networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[8] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 48, Jun 2016, pp. 2915–2923.

[9] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *NIPS*, Dec 2016, pp. 2082–2090.

[10] H. Ye, G. Y. Li, and B. Juang, "Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, Feb 2018.

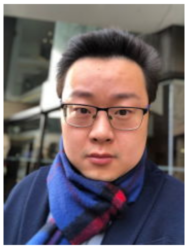
[11] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5663–5678, Nov 2018.

[12] H. Owhadi, C. Scovel, and T. Sullivan, "On the brittleness of bayesian inference," *SIAM Review*, vol. 57, no. 4, p. 566–582, April 2015.

[13] U. Challita, A. Ferdowsi, M. Chen, and W. Saad, "Machine Learning for Wireless Connectivity and Security of Cellular-Connected UAVs," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 28–35, February 2019.

[14] A. Arrieta, N. Rodriguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82 – 115, 2020.

[15] J. Mao, C. Gan, P. Kohli, J. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *International Conference on Learning Representations (ICLR)*, 2019, pp. 1–10.



Weisi Guo (S07, M11, SM17) received his MEng, MA, and Ph.D. degrees from the University of Cambridge. He is Chair Professor of Human Machine Intelligence at Cranfield University. He has published over 150 papers and is PI on over 12 research projects from EPSRC, Royal Society, EC H2020, and InnovateUK. His research has won him several international awards (IET Innovation 15, Bell Labs Prize Finalist 14 and Semi-Finalist 16 and 19). He is a Turing Fellow at the Alan Turing Institute and Fellow of Royal Statistical Society.