ARTIFICIAL INTELLIGENCE
EMPOWERING THE
DIGITAL TRANSFORMATION

Ernesto Damiani, Antonio Manzalini

L'inizio dello sviluppo dell'IA (*Intelligenza Artificiale*) risale al 1943 quando Warren McCulloch e Walter Pitt proposero un primo modello di neurone artificiale (perceptron). L'arrivo, alcuni anni più tardi, dei primi prototipi di reti neurali determinò un crescente interesse scientifico per l'IA, anche grazie ai nuovi lavori del giovane Alan Turing, volti a capire se un computer possa comportarsi come un essere umano. Oggi, a distanza di circa settant'anni, dopo il cosidetto *inverno* dell'IA, i recenti sviluppi tecnologici dell'ICT e delle Telecomunicazioni stanno rivitalizzando l'interesse, e significativi investimenti, per l'IA, indirizzandone addirittura un ruolo chiave nella Trasformazione Digitale.

Introduction

The evolution of Telecommunications infrastructures towards Future Networks and 5G (i.e., 5th Generation of networks) is facing today three major techno-economic challenges: *simplifying* the networks architectures (e.g., delayering and decommissioning, while achieving more efficiency and cost effectiveness) to provide any sort of digital services, with shorter time to market and better quality for the Customers; *cloudifying/edgefying* the virtual network functions and services; *optimising* and automating OSS/BSS processes to mitigate an increasing complexity.

In the ongoing Digital Transformation, these challenges are expressed in the emergence of a common high-level architectural model, for the Telecommunications and ICT ecosystems: future networks and services platforms will become software environments, almost fully decoupled from an underneath physical infrastructure; these platforms will be capable of hooking together processing, memory/storage, networking virtual resources, as well as network functions and services (e.g., *service chains* executed in network slices).

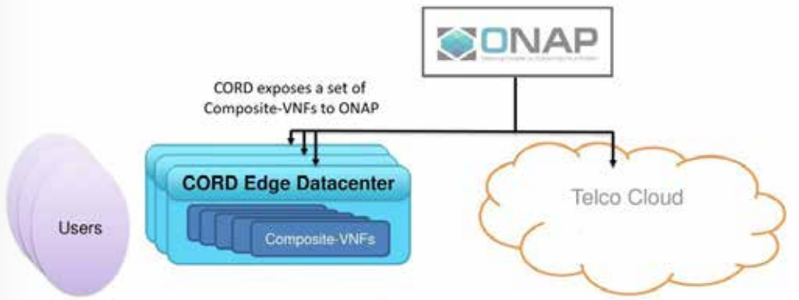
As a matter of fact, since a few years we have been witnessing some key drivers which are paving the way to this transformation. Among these drivers there are: the diffusion of ultra-broadband connectivity, the increasing of performances of IT systems as well as the down-spiralling costs of hardware, the emergence of innovative technological paradigms such as SDN (*Software-Defined Networks*) and NFV (*Network Function Virtualization*), the growing availability of open source software and also the impressive advances of AI/ML (*Artificial Intelligence/Machine Learning*).

It has been the growing complexity of this transformation, both in term of technologies and business implications, which has determined a revamped interest on AI/ML, which is manifested in growing investments in related systems, methods and tools for Future Networks and 5G. In fact, SDN and NFV paradigms will allow decoupling hardware and software architectures: if on one side this will offer an improved flexibility, on the other side a new complexity will emerge. The infrastructure physical layer will includes processing, memory/storage and network resources (up to the edge), while a virtualization layer (i.e., software) will provide, through Application Programmable Interfaces (APIs), different levels of abstractions all of resources, functions and services (e.g., from middle-boxes to applications).

In this scenario, for example, VNF (*Virtualized Network Function*) and services would be dynamically combined and orchestrated to create specific end-to-end service chains serving applications. At the same time, seen from the slicing viewpoint, the infrastructure would provide *slices*, as *isolated* pool of resources, where to execute multiple chains to serve applications (following specific QoS requirements of the Verticals). This is like saying that the Central Offices will become Data Centres, but with another element of innovation: the Cloud Computing will be complemented by Multi-Access Edge Computing (*Figure 1*). This means that small-medium Data Centres at the edge of the infrastructure can host smaller Central Offices (*edgefication of the network*).

In this high level architectural model, the OS (*Operating System*), as shown in *Figure 2*, will play the role of a software platform enabling management, control and orchestration capabilities (e.g., OSS/BSS processes) and services, through a secure and controlled access to all the abstractions, in order to serve any vertical applications. This same model will allow also Third Parties to access the infrastructure service planes, augmenting the Operator's role into Service Enabler [Manzalini 2014].

This technological evolution will dramatically increase the flexibility of network and service platforms while ensuring the levels of programmability, reliance and performance required by future 5G scenarios and applications (e.g., Internet of Things, Tactile Internet, Immersive Communications, Automotive, Industry 4.0, Smart Agriculture, Omics and E-Health, etc). On the other hand



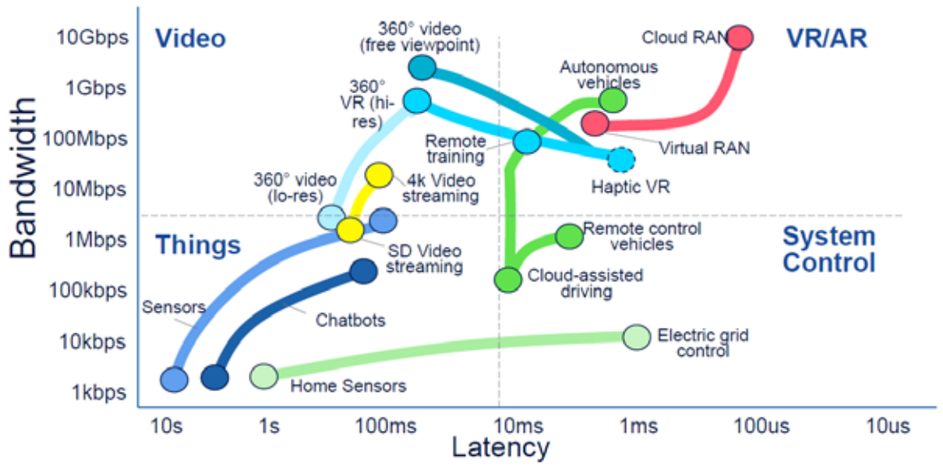
1
Example of integrated orchestration (ONAP) of Cloud and Edge Central Offices *(Source: ONF)*

the management complexity of such future infrastructures (e.g., for FCAPS and orchestration of virtual resources and services) will overwhelm human-made operations, thus posing urgent needs of designing and deploying OS with AI/ML features. The use of the huge *data lake* generated by the infrastructure

2
Emergence of a common reference model for Future Networks and 5G

will allow automating processes by introducing cognitive capabilities at various levels. Example of cognitive capabilities include: understanding application needs and automating the dynamic provisioning of services; monitoring and maintaining network state; dynamically allocating virtual network resources and services; ensuring network reliability and enforcing security policies.
Moreover, although we cannot yet fully grasp how pervasive AI/ML will be, it is likely that it will also enable innovative features when provisioning future digital cognitive services for homes, businesses, transportation, manufacturing, and other industry verticals, included the smart cities. For example, the strict requirements of future 5G services (*Figure 3*) will require automatic intent frameworks capable of: compiling intents for OS processing; applying policies and verifying properties; submitting for resource allocation; verifying feasibility, configuring the network resources.
It's then reasonable to argue that AI/ML potential will be so valuable for the business sustainability of Telecommunications and ICT, that not only will help extracting the required level *simplicity* for operating future infrastructures, but also it will potentially enable new service paradigms and business roles.

3
Bandwidth vs Latency requirements of applications and services



Potential Impacts of AI on 5G and Future Networks

What are the potential game-changing consequences of AI/ML in the Digital Transformation of Telecommunications? In this section we are briefly reporting some areas of innovation which are already ripe for AI/ML exploitation in network and services infrastructures.

Automated Operations and Network Intelligence

Today, many Telecom Operators are still relying on manual management processes: on the other hand there is a clear awareness of the potential for using AI-powered solutions for automation thus reducing costs, increasing productivity, and driving more value. The rationale is to use of AI for automating the operations processes based on collection and elaboration in (almost) real time of data about states and levels of performances of nodes/systems and logical/virtualised resources etc. For example, AI can automate the management, control and orchestration (e.g. MCO) processes of physical pieces of equipment, which today are mostly carried out by humans, introducing control loops acting on virtual/logical entities (e.g., Virtual Machines, Containers, appliances etc). In this direction, AI promises to deliver scalable OSS/BSS functions based on ML models capable of seeing and interpreting the state of millions of network entities via the analysis of huge data streams. Moreover, network and service computational intelligence (e.g., in the Radio Access Networks and in the Core), based on data about Customers' service patterns and traffic, would allow improving the quality of the customer experience whilst optimizing the use of resources.

Cybersecurity

Future Networks and 5G will have to face all the security challenges typical of today's Telecommunication infrastructures, but with a new and IT-oriented perspective brought by SDN and NFV. Nevertheless, these same enabling technologies, integrated with AI/ML will provide new instruments to mitigate such risks. To mention some examples: AI/ML will allow inferring proactive actions (even based on early-warning signals of attacks); the adoption of flexible and automatic features for fast traffic steering (e.g., quarantine, honey pots, slicing segregations); the automatic configuration of security virtual appliances to be added into the service chains.

Smart Capex and Opex

Concerning the Smart Capex and Opex, AI/ML would enable the adoption of "QoE" models and indicators to support investment and design processes based on a data-drive approach (e.g., selection of deployment regions, strategic priorities, etc). Moreover, regarding Opex optimization, it is well known that energy consumption is one of the major cost items for Network Operators: AI/ML methods would allow using the *data lake* for implementing performance analysis and optimization methods for energy consumption vs quality of service.

Improving agility and enabling new services

Operators and Service Providers are very aware that they are competing with agile players, such as Google or Amazon, and are looking to build an infrastructure where AI/ML will make the service platforms agile enough to bring up new services in minutes or hours versus weeks or months. Moreover, AI/ML is expected to enable the development new applications in key domains like Internet of Things, Tactile Internet, Immersive Communications, Assisted or Autonomous Automotive, Industry 4.0, Smart Agriculture, Genomics/Omics and E-Health, and others. By doing so, AI promises to deliver a key contribution to the infrastructure's business sustainability. In fact, increasing competitive pressure in the Telecommunications market is forcing Network Operators and Service Providers to look for novel solutions for reducing/optimising operations costs to compensate the cases where revenues are declining. By promising to decrease the management costs (e.g., with automated management, control and orchestration) and to boost revenues (e.g., by enabling innovative applications), AI adoption is increasingly perceived as a key competitive advantage.

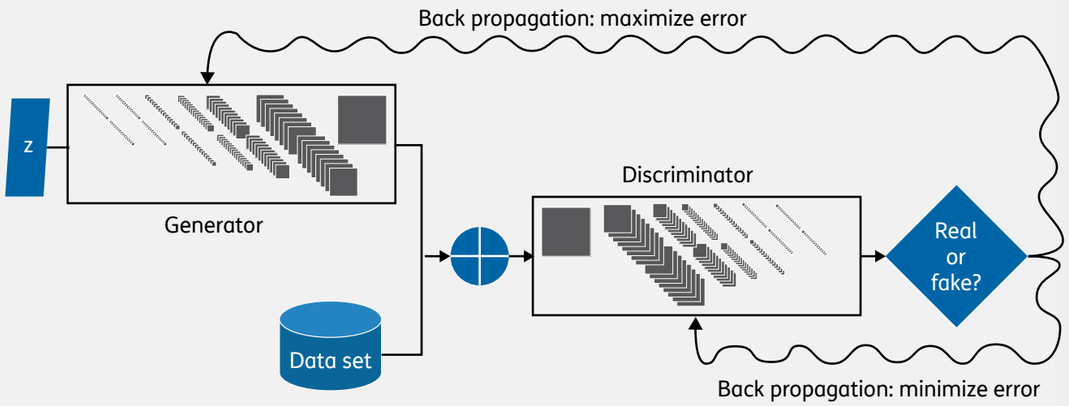
Data Multi-dimensionality and Deep Learning

After the so-called AI *winter*, in the last few years, AI has delivered many applications showcasing impressive performance and potential of practical exploitations. A convergence of technology trends and drivers, social transformations, and new economic needs is now paving the way for a wide adoption of AI in several businesses and industries. For example, the availability of software frameworks for handling big data revived AI innovation of ML and DL (*Deep Learning*).
Multi-dimensionality of data is one of the major challenges. Multi-dimensionality is a property of data

GANS
5G GENERATIVE ADVERSARIAL NETWORKS

GAN is a relatively new Machine Learning architecture for neural networks: it was first introduced in 2014 by University of Montreal (***see this paper note 1***). In order to better capture the value of GANs, one has to consider the difference between Supervised and Unsupervised learning. Supervised neural machineries are trained and tested based on large quantities of "labeled" samples. For example, a supervised image classifier engine would require a set of images with correct labels (e.g. cats, dogs, birds,...). Unsupervised neural machineries learn on the job from mistakes and try avoiding errors in the future. One can view a GAN as a new architecture for an unsupervised neural network able to achieve far better performance compared to traditional ones.

A
Architecture of the Generative Adversarial Networks



Main idea of GAN is to let two neural networks competing in a zero-sum game framework. A first network takes noise as input and generates samples (generator). The second one (discriminator) receives samples from both the generator and the training data, and has to be able to distinguish between the two sources. The two networks play a game, where the generator is learning to produce more and more realistic samples, and the discriminator is learning to get better and better at distinguishing generated data from real data. These two networks are trained simultaneously, in order to drive the generated samples to be indistinguishable from real data.

GANs will allow training a discriminator as an unsupervised "density estimator", i.e. a contrast function that gives us a low value for data and higher output for everything else: discriminator has to develop a good internal representation of the data to solve this problem properly. More details [***note 2***].

GANs were previously thought to be unstable. FAIR (*Facebook AI Research*) published a set of papers on stabilizing adversarial networks, starting with image generators using LAPGAN (*Laplacian Adversarial Networks*) and DCGAN (*Deep Convolutional Generative Adversarial Networks*), and continuing into the more complex endeavour of video generation using AGDL (*Adversarial Gradient Difference Loss Predictors*).

As claimed [***note 3***], it seems that GANs can provide a strong algorithmic framework for building unsupervised learning models that incorporate properties such as common sense ■

where each record includes hundreds or thousands of dimensions (also called *features*). The human brain itself is a deep neural network where many layers of neurons act as feature detectors, detecting more and more abstract features. A Telecommunications network, which actually operates over multiple protocol layers at the same time, is a nice example of a source of highly dimensional data streams.

In this direction, many recent researches are focusing on DL techniques based on the notion that the analysis (e.g., for classification or anomaly detection) of data points characterized by many features needs to follow a hierarchical abstraction process, which cannot easily be modeled by classic NNs (*Neural Networks*).

While the potential of DL-NN (*Deep Learning Neural Networks*) for analysing highly dimensional data streams was recognized already in the Nineties, DL was impaired for a long time by the fact that traditional NN training techniques like *gradient descent* (following the negative gradient of the network error function) turn out to be very slow to converge (if they converge at all) when the gradient must be spread across multiple layers. This is known as the *vanishing gradient problem*: the distance between the desired and the actual values of the function computed by a NN (the error) shrinks exponentially with the number of layers, making the error gradient difficult to identify and follow on the

[1] https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
[2] https://ishmaelbelghazi.github.io/ALI/
[3] https://www.linkedin.com/pulse/gans-one-hottest-topics-machine-learning-al-gharakhanian/

part of traditional training methods. In other words, traditional NN training makes poor use of additional layers.

Figure 4 shows the DL principle and the approach to this problem: rather than training the network's intermediate layers based on the final classification or anomaly detection analytics to be performed, DL techniques train each internal layer pair to translate the original feature space into another one. This translation preserves the *probability density* of the input data space: input data points are mapped into (a usually smaller) number of outputs, each of which has many occurrences equal to the sum of the occurrences of the inputs mapped into it. The important notion here is that training a layer pair to preserve the input space probability density does not involve minimizing the global error function of the desired analytics, but a local, independent error function. Therefore, training of internal layer pairs can be performed independently (e.g. part on the terminal devices, part at the network edge/core) and quickly. Only the final layer pair ("ML Model" in the right of *Figure 4*) will undergo the traditional NN gradient-based training, minimizing the error on the desired analytics function.

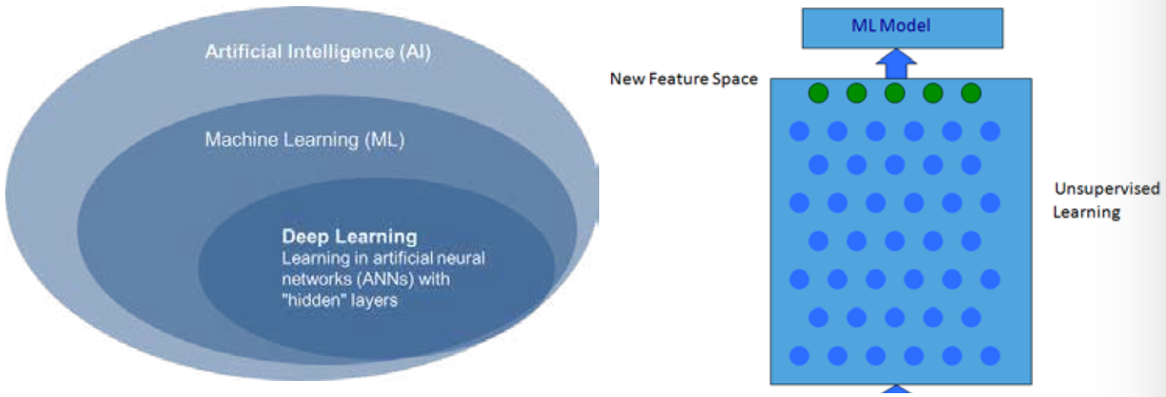
DL-style training has delivered a series of impressive results from 2011 onwards (including IBM Watson's success against a human champion at "Jeopardy!", or the recent defeat of a human player of "Go" at the hands of a DL model).

However, it is important to remark that the last few years of research have proven that DL analytics can achieve satisfactory performance only provided that two key conditions are met:

a. input data is an accurate statistical representation of the physical environment, suitable for the chosen ML model;
b. distributed training and execution of the chosen ML model can meet the deadlines given the application's latency and resource constraints.

These two key conditions are not always met when considering the available *data lake* of Telecommunications infrastructures. As a matter of fact, network data are made available by the network entities through an implicit transformation process that includes a diverse panoply of preparation and pre-processing activities. This transformation is independently performed devices at the periphery, as well as by edge processing, and is far from an ideal statistical measurement process (e.g. the classic one mapping a point value into a normally distributed measurement). Input data latency, availability and veracity, as well as the corresponding computational load, may widely vary, depending on the conditions in the field.

4
Deep Learning principle



Example of applications and challenges

In future scenarios, as shown in *Figure 2*, where the network is more and more pervasive and dynamic, like a sort of *nervous system* of the Society, the *data lake* is made by are huge amout of highly heterogeneous which are coming from a sheer number of diverse sources (e.g., network nodes, IT systems, terminals, devices, smart things, etc). These data include states, configurations, performance data, alarms, logs, etc. and they will show different feature subsets: for example, in 5G it is expected that each data item may easily have hundreds or even thousands of dimensions. This is a big challenge for AI systems.

Let us consider the example of using DL to quickly identify cyber-threats in a 5G network. In principle, the 5G allows extracting highly dimensional, multi-layer data points describing complex entities like "connections" from network flows. DL learning techniques can support adapting the configuration of the cyber-defence architecture according to fluctuation of traffic gathered from 5G subscribers' UE (*User Equipment*) in real-time (e.g. the abstraction of a "too long" connection may depend on context-related, slow-varying features such as the data traffic cost per Gigabit for each customer segment). DLs distributed training can transparently fine tune the behaviour and performance of the network's own analysis and detection processes.

In this further example [Veeramachaneni 2016] an end-to-end system combines analyst intelligence with state-of-the-art machine learning techniques to detect new attacks and reduce the time elapsed between attack detection and successful prevention. The system presents four key features: a big data behavioural analytics platform; an ensemble of outlier detection methods; a mechanism to obtain feedback from security analysts; a supervised learning module.

As another example, let us consider future applications such as autonomous vehicles for SDT (*Self-Driven Transport*) scenarios. The latencies requirements for these applications are so strict (e.g., order of ms) which is not possible to close the loop executing the applications in the Cloud Computing. For example if the distance of the Cloud is one hundreds of km, only the round trip-and-back transmission latency is the order to tenth of ms which should be added to the processing latency of the end-to-end service chain. This is not compatible with the reaction times required by SDT. The deployment of local processing and MEC solutions can help mitigating this problem but it requires an OS (i.e., management/control and orchestration capabilities) capable integrating on-device, edge-based and cloud-based training of DL models. In this context, local resources (e.g., in the vehicle, or mobile terminals) could be very promising for hosting the training of the intermediate layers of DL-NN models; in this case the MEC could efficiently aggregate abstractions while the Cloud could deliver fast final layer training and model updates. In this direction, Technology Providers have been gradually introducing high cognitive capability in mobile terminals (for instance, Apple has introduced the framework for Deep Learning Inference BNNS – Basic Neural Network Subroutine): this can potentially pave the way even to AI-supported smartphone-to-smartphone interaction for several applications.

In synthesis, in order to face Future Networks and 5G applications, AI should aim at cognitive solutions with more human-like characteristic s-such as an intuitive understanding of the real world and more efficient ways of learning. This may require a renewed exchange of ideas between AI and neuroscience can create a 'virtuous circle' advancing the objectives of both fields [Hassabis 2017].

Trends and perspectives

The effective applications of AI in Future Networks and 5G scenarios are likely to require multi-domain orchestration of distributed processing in the terminals/devices (e.g., Fog Computing), at the edge (e.g., MEC) and in the Cloud Computing facilities. Moreover, in

multi-Operators scenarios, it is expected that the composition of AI virtual pipelines (i.e., acquisition, preparation, pre-processing, and analytics) will look like, or better will become, part of a chain of network services [Damiani2017], where some of them can belong to different Operators and may pursue different and non-perfectly aligned goals.

In this direction, the end-to-end interoperability is a must and it requires standardization more efforts and further achievements. First of all, it is necessary to consider the impact of current, and future, AI systems and methods in the functional architecture of Future Networks and 5G. This means understanding which and how the architectural functional blocks will be impacted, and what will be the related standardised interfaces. In this direction a global effort is still required from both hardware and software vendors to participate in standardization bodies. Moreover the design and development activities of Future Networks and 5G system will have include collaborations between industry standardization forums and Open Source communities (e.g., ETSI, ITU-T, Linux Foundation, ONF, OCP-TIP, IETF, IEEE).

Overall some key technical challenges have to be solved for a mature application of AI in the Telecommunications domains. One challenge, for example, is the loss of precision due to missing data, alignment of different dimensions and artificial autocorrelation in the data flows, and assure the desired non-functional properties of AI model computation in terms of privacy, data integrity and protection. In the medium term, moreover, AI will have to rely on two pillars: structural awareness, making the learning process reminiscent of the multiple sources contributing to multi-dimensional data, and adversarial composition, modelling the data preparation/gathering as a source of perturbation/noise. This paradigm would take as parameters the pertinent uncertainty models and the related uncertainty principles. It will also be important to address ethics and legal concerns that may depend on the cultural and regulatory environment where the pipeline is deployed.

Moreover, it should be considered that feature data sets collected by mobile terminals and devices at the network periphery have natively faceted structures that can be exploited in the learning strategy. For example, recent ML analytics called multi-view learning can treat input data facets (called views) differently, e.g. using multiple learning classification models or coordinating training of multiple models (co-training). Notably, multiple learning algorithms exploit learners that naturally correspond to different views and combine them linearly or non-linearly to improve learning performance. In turn co-training algorithms pursue agreement between models trained on distinct views, and subspace learning algorithms try to identify a latent subspace shared by multiple views by assuming that the input views are generated from it.

Learning model composition are also expected to incorporate adversarial learning, which deals with highly dimensional data where features may have diverse veracity, due to the presence of hostile, un-trusted or semi-trusted components along the model training chain. The adversarial paradigm (see box for details) considers the data preparation/gathering as inherently including a source of perturbation/noise and train DL models considering the uncertainty type and the corresponding uncertainty principles.

In the technological trend towards exploiting more and more general-purpose systems and hardware, AI could play the crucial role to make the systems able to automatically customized/adapt to the specific context, data and access patterns of services requests.

So, if it's true, according to Darwin's Origin of Species, that it is not the strongest the species that survives, but the one that is able best to adapt and adjust to the changing environment in which it finds itself, Network Operators which will be able to integrate AI in 5G are likely to gain a new competitive advantage, even beyond adaptability: their infrastructures will improve with age [MIT] ■

References

[Manzalini 2014] A. Manzalini, D. Soldani, "5G: The Nervous System of the True Digital Society", E-Letter for Multimedia Communications Technical Committee IEEE Communications Society, Septenber 2014, *http://www.comsoc.org/~mmc*

[Veeramachaneni 2016] K. Veeramachaneni, et al. "AI2 : Training a big data machine to defend" available at *https://people.csail.mit.edu/kalyan/AI2_Paper.pdf*

[Hassabis 2017] D. Hassabis, et al., "Neuroscience-Inspired Artificial Intelligence", Neuron , Volume 95 , Issue 2 , 245 - 258 available at *http://www.cell.com/neuron/fulltext/S0896-6273(17)30509-3*

[Damiani 2017] E. Damiani, et al. "Toward Model-Based Big Data-as-a-Service: the TOREADOR Approach", ADBIS 2017: 3-9

[MIT2018] Will Knight, "Your next computer could improve with age", MIT Technical Review available at *https://www.technologyreview.com/s/610453/your-next-computer-could-improve-with-age/?utm_source=newsletters&utm_medium=email&utm_content=2018-03-13&utm_campaign=the_download*

Ernesto Damiani *ernesto.damiani@kustar.ac.ae*

Ernesto Damiani is a Full Professor at the Università degli Studi di Milano, where he leads the SESAR research lab, and he is the leader of the Information Security Center at the EBTIC/Khalifa University in Abu Dhabi, UAE. From 2017, he is the Principal Investigator of the H2020 TOREADOR project, the biggest project on Big Data Analytics led by an Italian group. He was a recipient of the Chester-Sall Award from the IEEE IES Society (2007) and received the Stephen S. Yau Services Computing Award (2016). In 2017, he received a Laurea Honoris Causa from the Institut National des Sciences Appliquées of Lyon, France. He is the authors of more than 500 papers and his work has been cited more than 5000 times ■

Antonio Manzalini *antonio.manzalini@telecomitalia.it*

ingegnere elettronico, Ph.D è entrato in Telecom Italia nel 1990 ed ha partecipato a diversi progetti di ricerca internazionali riguardanti reti di trasporto SDH ed ottico (WDM), occupando varie posizioni di responsabilità. Ha inoltre partecipato a molte attività di standardizzazione, guidando alcuni gruppi di lavoro in ITU-T ed IEEE. Attualmente si occupa di tecnologie ed architetture di reti evolutive in ottica 5G, basate sull'integrazione di SDN, NFV con Cloud-Edge Computing e sistemi di Intelligenza Artificiale. È autore un centinaio di pubblicazioni internazionali e di sei brevetti ■