STRATEGIC MANAGEMENT OF NOVEL DISTRIBUTED AIML TECHNIQUES FOR NEXT GENERATION TECHNOLOGY AND BUSINESS - A HOLISTIC STUDY FOCUSING ON FRAMEWORKS FOR BUSINESS GROWTH AND OPTIMIZATION

by

Kausik Chakrabarti, M.Sc Engineering, MBA

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

February, 2025

<STRATEGIC MANAGEMENT OF NOVEL DISTRIBUTED AIML TECHNIQUES FOR NEXT GENERATION TECHNOLOGY AND BUSINESS - A HOLISTIC STUDY FOCUSING ON FRAMEWORKS FOR BUSINESS GROWTH AND OPTIMIZATION>

by

Kausik Chakrabarti

APPROVED BY

Apostolos Dasilas

, Chair

SSBM Representative

Renee Goldstein Osmic

Dedication

This research is dedicated to my grandfather Dr. Kumud Bandhu Chakrabarti, whose indomitable spirit inspired and influenced me in all my academic endeavors.

ABSTRACT

<STRATEGIC MANAGEMENT OF NOVEL DISTRIBUTED AIML TECHNIQUES ON NEXT GENERATION TECHNOLOGY AND BUSINESS - A HOLISTIC STUDY FOCUSING ON FRAMEWORKS FOR BUSINESS GROWTH AND OPTIMIZATION>

Kausik Chakrabarti
2024

Infusion of novel Artificial Intelligence and Machine Learning (AIML) algorithms and tools in technology and business platforms herald a transformative evolution in how businesses reinvent themselves to a more cognitive, autonomous structure - toppling the evolutionary process of technology and process change management through siloed views by incumbent technology leaders and shifting the complete ecosystem to a marketplace economy, with alliance and ecosystem formation at the center to derive significant values for each other. This thesis thus aims to deep-dive into some novel AIML technologies and key real-time platform technologies to research the value propositions through adoption of strategic business frameworks aiding the decision-making process for technology and business transformation through the formation of a converged yet dynamic partnership ecosystem. The results and outcomes of the research study and learnings are expected to be strategically and tactically helpful for businesses adopting AIML in their core businesses, helping to adapt and construct effective

evaluation methodology to navigate through the highly dynamic technology and business landscape that need strategic partnerships and alliances to derive meaningful values.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:
INTRODUCTION

1.1 Introduction

Technology evolution with cloud-based technologies is at a crossroad (Chakraborty et al., 2017). The new vision for Business and Technology growth opportunities in Cloud platforms are being guided by leveraging novel techniques of Artificial Intelligence and Machine Learning (AIML).

Market dynamics have recently shown a high flux and a natural shift towards adoption of a more open market economy and technologies that these new innovations demand and value creation (WEF and Kearney. A. T, 2017). Also, the rise of social media platforms as a front-runner in the heralding of technology revolution - toppling the evolution process of incumbent technology leaders and shifting the complete ecosystem to a marketplace economy is the new reality. Technical reasons aside, from the perspectives of Market economy, the adoption changes ring the bell for the following necessary shifts, using AIML techniques:

- A more competitive shared eco-system for business and technology development - with partners driving the cooperation mode in the marketplace.
- Adoption of a more dynamic pricing strategy based on new cloud-based business models that will help long-term partnership development based on trust and scale of adoption.
- In the new normal, based on cloud technology platforms, a more adaptive approach towards performance measurements will be adopted, based on service requirements and cloud KPIs and metrics.
- High-tech mergers and acquisitions will be normalized and balanced by a more democratic alliance between industry players.

In this research, the aim will be to deep-dive into some novel techniques of Artificial Intelligence and Machine Learning technologies (AIML) that are expected to be significant and strategic contributors in realizing the above business transformation in the future.

1.2 Research Problem

In Strategic management of Wireless Technology-related businesses, strong partnership and alliances lead to an ecosystem formation that then drives the engine of growth and consumer satisfaction - a concept pioneered by the very successful android ecosystem formed by Google, as per Fautrero and Gueguen (2013). However, forming such an ecosystem by next-gen businesses, needs careful consideration of strengths and weaknesses in adoption of AIML as the fundamental core enabler for each of the partners. failing which the ecosystem is doomed to its eventual collapse. There is thus an urgent strategic business need to seek the rules of engagement and adoption of working frameworks and value-based methodology that helps to form a winning ecosystem - based on its choice of adoption of short-term and long-term AIML technologies, algorithms and processes.

Intelligent conversational agents based on Deep Learning and Generative AI (Adamopoulou and Moussiades, 2020) are expected to shape up various domains of future businesses. Their transformation journey will need to be researched thoroughly to assess their impact to business and technology and finally their adoption strategies and tactics.

1.3 Purpose of Research

The overall aim of the research is thus to delve into the current application of AIML technologies and processes in technology and business and seek out a framework that can dictate the trajectory of a winning ecosystem and make the best economic decision for build (based on open source) or buy (from vendors like Cloud Service Providers). For this reason, the research will not only focus on the novel algorithms, processes and techniques for AIML that are currently coming to the market but it will also address the business values and decision making for adoption based on sound economics. The following research questions that will need to be broadly addressed:

- What role do novel AIML techniques and automated machine learning methods play in fast-forwarding the efficiency and utilization of Machine Learning in Business and Technology? How do they impact the business outcomes and how the end-to-end Machine

Learning model delivery eco-system is ultimately stitched in real-time streaming platforms for efficient business outcomes.

- What is the expected depth and breadth of usage of novel Deep Reinforcement Learning and its derivatives in Future Technologies (like AI training and interference for 6G Wireless Platforms) and Businesses (especially in AIML-based business Platforms).
- How are next generation businesses and wireless technologies, (Drones, UAVs and Edge Networks) impacted by distributed AI based updates/enhancements, reinforcement Learning and Deep Reinforcement and multi-agent Q-Learning? How Deep Reinforcement Learning-based Multi-Agent Systems and game-theoretic evaluation of optimal performance in a distributed system can be exploited in both Technology and Business domains.
- From next generation platform business perspectives how real-time streaming data platforms impact the business outcomes? through the usage of real-time streaming platforms?
- Ultimately, the above exploration will be utilized to derive a holistic ecosystem partnership formation strategy based on Value Chain analysis and/or Value Train Analysis and Lean Business Models and may also use valuation of various parameters, based on the depth and breadth of AIML technology incorporation in their solutions to make an informed decision for build (in-house) OR buy (from Cloud Service Providers) strategies and decisions.

The above questions call for the answers and will be addressed in the thesis.

1.4 Significance of the Study

Little has been investigated in the area of applying existing and novel business frameworks that leads to a generalized application of novel techniques of AIML in Business and Technology. The same will be explored via technical deep dives in various AIML technologies of strategic importance, further leading to specific case studies rounded up with business strategic and tactical plays using partnership mode analysis, Platform Business Model Map and Value Train Analysis (Rogers, 2016) respectively; aiding to connect the dependencies of various companies in the partner ecosystem that derive value out of AIML infusion at the core of the business processes and functions.

1.5 Research Purpose and Questions

Based on the broad areas proposed for the research, the main aim of the research and exploration can be drilled down to some key AIML algorithms, techniques and solutions that commonly solve business and technology problems resulting in transformative value creation, yet addressing some key challenges that need to be tackled along the journey. This shall lead to the following objectives:

1. Research of novel and automated AIML algorithms, online/incremental and federated learning with streaming data for real-time platforms & processes common to key areas of next generation autonomous networks; and allied businesses and assist in synergizing their common value propositions, culminating in a business analysis of AIML impacts.
2. Through a few focused case studies, investigate the existing strategic and tactical frameworks that redefines the rules of engagement in the light of next generation novel AIML technologies and platforms, to assist in the formation of winning ecosystem partnerships for businesses adopting AIML to undertake the journey of value-based transformation.

This research thus should lead to deeper understanding of impacts of novel AIML technologies to undertake AIML-based business transformation and how best they can be leveraged to form strategic ecosystem partnerships and alliances with sound decisions on “make or buy decision” or “open-source versus proprietary” solutions to generate viable and durable business propositions.

CHAPTER II:

REVIEW OF LITERATURE

2.1 Theoretical Framework

Literature review conducted provided several directions for application of novel techniques in Technology and Business); while most of the novel techniques are leading to emergent strategies, we expect to converge on a few key few ones that promise to have wider acceptance in next generation technology and business domains to accomplish the research objectives, while continuing to study the competing technologies and processes.

AIML in Traditional wireless technology and business management

Message-Oriented Mobile Middleware and Multi-agent Systems have ruled the roost as far as intelligent application orchestration in mobile wireless systems is concerned. Synchronous and asynchronous communication interaction models, aided by traditionally accepted remote-procedure calls and Message-oriented middleware have been covered in the book (Tanenbaum and Steen, 2002), and formed the fundamental frameworks and methods for achieving distributed computing in the traditional distributed computing systems in the earlier decade. When it came to Data integration, aggregation and mining technologies for business intelligence, OLAP, OLAM laid the foundation stones, detailed in the book (Han and Kamber, 2006) aided by traditional machine learning algorithms like classification and clustering and their allied sub-categories studied in the book (Alpaydin, 2014).

The paper by H. L. Zhang and H. C Lau (2014) focused on the study undertaken, using agent-based methods to solve the complex computational problems arising in Big Data environments. The paper aggregated and covered various research work that were being conducted and the advances made in the areas of distributed problem solving, agent-based data mining and recommendation systems, working with data extracted from both physical and online environments. This body of work helps to bridge the gap between the past research and their applications in latest technologies, preparing us for further studies on the cutting-edge research work that is “in-progress” in the current times with further heavier emphasis on AIML augmentation.

Game Theory for years has served as a vital toolbox for Strategic Business Management. In a very relevant study, Oderanti (2001) in his paper, notes:

“In business games, the firm identifies the moves that the rival could make in response to each of its strategies. The firm can then plan counter-strategies (Griffiths and Wall 2000). As Doug Ivester, Coca-Cola’s president put it (Himmelweit et al. 2001):

“I look at the business like a chessboard. You always need to be seeing three, four, five moves ahead; otherwise, your first move can prove fatal…”

Oderanti (2001) further reiterates in the paper:

“Game theory helps to explore the impact of calculations about future market advantages on a firm’s current strategies.”

Progressively, the research paper by Agarwal and Jaiswal (2012) furthers the use of Machine learning in Game theory - it states its study object as follows:

“We study the problem of development of intelligent machine learning applications to exploit the problems of adaptation that arise in multi-agent systems, for “expected-long-term” profit maximization…Second, we study the same problem from the aspect of zero-sum games. We discuss how AI and Machine Learning techniques work closely to give our agent a ’mind-reading’ capability…”

With the above paper, the seed of automation of Game Theory with AIML was thus possibly sowed, germination of which is still in progress in the shape of an active research area via Multi-agent Reinforcement Learning.

In the next section, we move on to capture and further elaborate the current state of affairs in each of the above presented topics in technology and business that showed promises of value enhancement by the application of AIML.

Nextgen technology and business management - impact of AIML Algorithms

As already discussed, Kubernetes-based cloud-native platform, AIML and agent-based distributed monitoring form the basis for O-RAN, a successor of RANaaS - a cloud-based framework conceived for future generation Radio Applications as suggested by Nikaein, et al. (2017). The O-RAN Alliance (2018), proposes in their whitepaper that their core objective, other than to provide an open multi-vendor compatible vibrant and competitive ecosystem, is to incorporate intelligence to manage increasingly complex network configurations and demanding applications while reducing network operation costs with hosting newly defined RAN Intelligent Control (RIC) located between RAN and OAM in O-RAN architecture as analyzed in the paper (Lee et al., 2021). As per another whitepaper by O-RAN Alliance (2020), RIC can be used to enhance the efficiency of traditional Radio Resource Management (RRM) functions with advanced control capabilities.

The book by Russel and Norvig (2003) undertakes a thorough study of each of the use cases found to achieve sophistication only by applying Supervised Machine Learning, Unsupervised Machine Learning or Reinforcement Learning. Again, the authors in their book (Seppo, H. et al., 2011) specifically cover the area of distributed applications where there are several interacting entities in a resource constrained environment, game theoretic constructs were offered as a solution to create strategies for maximizing overall utility.

Basic machine learning algorithms (unsupervised, semi-supervised and SVM) which form a vital basis of data science are still of relevance for current and possibly future applications, and their innovative application and possible replacement with Quantum Machine Learning (QML) algorithms will continue to be researched, as suggested by V. Kulkarni, M. Kulkarni and A. Pant (2020) in their paper on Quantum Computing Methods for supervised Learning. These algorithms will be compared for similar applications in future technology businesses through the Automated Machine Learning (He, Zhao and Chu, 2021) lens, a tool that allows to automate various machine learning activities, most suited for researching machine learning algorithms from a business application view-point.

Also, contribution of AI-powered chips for training and interface at the edge of the network will be studied.

Marketing and Pricing Analytics have foundation on the competitive moves, both offensive and defensive as noted in his book by Sorger (2013). In order to understand this shift brought over by AIML thoroughly, we propose to research further and delve deeper on the process by which novel ML algorithms are absorbed into strategic business functions.

Next generation algorithms to tackle future challenges - Reinforcement Learning, Deep Reinforcement Learning interwoven with game theoretic analysis will continue to create more intelligent systems aided by the ease of real-time data availability and continuous learning. A related paper (Mao et al., 2016) focuses on the application of deep reinforcement learning in networking and workload scheduling management settings while the paper (Shuyang Li et al., 2016) researches on using reinforcement learning and game theory to develop “computation offloading technology” that:

“Can offload computing tasks to multi-access edge computing (MEC) servers, which is an appealing choice for resource-constrained end-devices to reduce their computational effort.”

While concepts and assumptions of predictive marketing analytics have been researched thoroughly by Tarka et al. (2014), the generational shift in adopting new algorithms is noted the paper by Shihab and Wei (2021) who delve into the nuances of applying cutting edge reinforcement learning algorithms for pricing and revenue analytics in a highly competitive business domain of airlines business. Our further research initiative will be to dive into the nuances and techniques that could be applied to derive benefits in communication-related business domains using AIML-driven Algorithmic Marketing Optimization techniques and Strategies - based on integrated data platforms, AIML tools and techniques.

Though the study of these new AIML algorithms aided by Game-theoretic policies does form a relevant research area, with the unification of Game Theory and Reinforcement Learning for possible business application scenarios is a topic of interest. The pioneers of Reinforcement Learning Research, Dr. Sutton and Dr. Barto note in their book (Sutton and Barto, 2018) that "Although reinforcement learning in economics developed largely independently of the early work in artificial intelligence, reinforcement learning and game theory is a topic of current interest in both fields…"

Deep Reinforcement Learning and its derivatives are extremely relevant in Future Technologies like 6G for AI training and interference for 6G Wireless (Khaled B. Letaief, 2019) and need to be researched thoroughly.

However, from the context of Technology and business, relevance application of reinforcement learning will be researched to derive a convergent view. Thus, a common theme for algorithmic frameworks leveraging AutoML for traditional ML algorithms and visualization and Reinforcement Learning do emerge as key techniques that will form the basis of AIML algorithm research for technology and business applications.

In the next section, we move on to capture and further elaborate the current state of affairs in each of the above presented topics in technology and business that showed promises of value enhancement by the application of AIML.

NextGen Technology and Strategic management - based on Real-time Data Platforms

Application and adoption of AIML in wireless technology is at its crescendo with the adoption of transformative framework of Cognitive Radio Networks and Software-defined Radios in the thesis paper (Mitola, 2000); research in Ran Intelligent Controller (RIC) in the open and distributed framework of Open RAN heavily relies on AIML as outlined in a publication by Analysys Mason, (2021), thus lending a definitive direction that is progressive and futuristic.

Critical time sensitive application like healthcare (e.g., telediagnosis, tele-surgery), industry (e.g., dangerous and difficult to-reach environments), virtual and augmented reality (e.g., a firefighter training system), road traffic (e.g., automated and cooperative driving), education and serious gaming (e.g., games for personalized cardio-training) demand extremely low latency requiring the application platforms be placed at the edge of the networks rather than in the cloud, as outlined in the recent research (Promwongsa et al., 2020). In parallel, interaction of massively large numbers of sensors and devices for “Internet of Things” applications need the related computing and analytics platforms be placed at the edge where initial intelligent aggregation and processing can be done locally, thereby reducing massive upstream data load to the central cloud as researched in the paper (Ghosh et al, 2018).

Since such platforms need to be access technology agnostic, serving 4G, 5G Wi-Fi, these are termed as Multi-Access Edge Computing Platforms. AIML forms the core component for aggregating and analyzing the diverse data at the edge and AIoT

(Artificial Intelligence of Things) is a relatively new term coined to address the hot topics of AI and IoT and the editorial of the special issue on the subject by Hindawi Publications (Sung et al., 2021), which addresses some specific applications in this area.

Edge Computing and Drones and UAVs for 6G are technologies representing paradigm shifts based on explosive computing and analytics demands at the "edge networks" and satiated only by application of optimized AIML algorithms and work-flows based on Cloud-native Container Orchestration Platform, like Kubernetes, a platform that has its origin in Borg Platform developed by Google (Borg, 2015).

In the area of Edge Networks, Multi-access edge computing (MEC) is a promising solution to avail of adequate computing and storage capabilities in close proximity to mobile users as per a technical brief by Intel and Nokia Siemens Networks (2013). Considering diverse and dynamic resource availability requirements of edge applications and the massive number of devices and data generated by the computationally intensive applications, powerful AIML tools and techniques capable of dynamically allocating both communications and computing resources to users is a necessity. Particularly, deep learning networks seem promising and fit to address the hurdle and empower intelligent resource management for efficient MEC in real-time and dynamic scenarios.

A new concept of mobile edge learning (MEL) has been recently defined in which MEC interplays with AI in the sense that the learning model, parameters, and data are distributed across multiple edge servers, and an AI model is trained from distributed data, as explained in a paper (Wang et al., 2019). Such a distributed learning model is known as federated learning, in which a node plays the role of the orchestrator that aggregates locally derived parameters and returns globally updated parameters to the servers as noted in a paper by U. Mohammad and S. Sorour, (2018). Such a mutually benefiting interaction between MEC and AI paves the way for ushering an intelligent self-driven autonomous communication network.

The Whitepaper from Akraino Alliance (2020), introduces the two of the most important current technology trends at the edge networks today - i) the extension of cloud computing and ii) 5G technology.

Akraino (2023), and an alternative innovative edge computing framework (MobiledgeX, 2022), to be open-sourced soon by Google Inc., thus should act as visionary references for edge network and architecture blueprint implementor, is also a technology aggregator and ecosystem partnership enabler, working side-by-side with the specification bodies like ETSI Multi-Access Edge Computing (ETSI MEC, 202x). Akraino and MobiledgeX would thus possibly present a compelling framework for next generation ecosystem formation bringing together open-source software development communities and best-of-the-breed communication companies participating and cooperating to create business value not only for themselves but for the global communities as whole.

The principles underlying the strategies and ecosystem formation, with Akraino as an example, thus also need to be researched and generalized from business perspectives, other than, from a technology standpoint, researching on the various AIML-based edge analytics techniques that will be implemented in MEC itself.

Research on symbiotic relationship of Distributed Machine Learning, utilization of agentic framework as a foundation of an abstract policy framework for orchestration and networking of next-gen Technology and Business services and applications has the potential to be an exciting field for the researchers.

Multi-agent Systems form the foundational backbone for intelligent application orchestration in mobile wireless systems. Previously discussed synchronous and asynchronous communication interaction models, have now been replaced by Restful Application Programming Interface (Rest API, 20xx) and Kafka-based (Kafka, 201x) loosely coupled data integration models befitting the latest data analytics platforms. They will be researched as part of a real-time data platform, for next generation technology and business.

Marketing and Pricing Strategies are increasingly AIML-based, based on integrated intelligent customer data platforms backed by strong GDPR-based data security rules.

‘Multi-agent’ based orchestration and services composition will be the main-stay for the decades to come. Business Process Models (BPMs) and services orchestration tools will be enhanced with the Digital Experience Platforms (DXPs), as suggested by B. Cheung of Liferay Inc. (201x), and tools for orchestration and composition of services. As defined by Mr. Cheung, “Digital Experience Platform (DXP) is an emerging category of enterprise software seeking to meet the needs of companies undergoing digital transformation, with the ultimate goal of providing better customer experiences. DXPs can be a single product, but are often a suite of products that work together. DXPs provide an architecture for companies to digitize business operations, deliver connected customer experiences, and gather actionable customer insight.”

Thus, there is a need to track and research this progress in our work and the same will be conducted.

Now considering the shifts in strategic businesses that AIML brings in “traditional Business Intelligence”, we find evidences that the telecom operators and marketing enterprises are participating in the contextual shift of business with AIML to stay relevant in the age of digital economy - building Digital Experience Platforms (DXPs) integrating social and telecom Customer Relationship Management (CRM) harnessing Big-data in data lakes and integrating that with rich enterprise-architecture framework to provide a seamless integrated digital interaction experience as per S. K. Shivakumar (2018).

Tools that aggregate the basic AIML algorithms - Data streaming and event processing tools like Spark and Flink, as noted by Macrometa (2024) will be researched through the readily available online/incremental machine learning with streaming data using Python frameworks (Scikit-multiflow, 2019) and their next generation solutions. The python-based frameworks will form the basis of streaming platform exploration for real-time data ingestion and analytics capabilities and will form the basis for comparing other leading industry-specific solutions for IoT, 5G/6G Wireless Systems (Technology domain) and even in DXP platforms (business domain) which will be studied.

AIML process automation tools like AutoML as noted in the paper (He et al., 2021), are expected to aid the next generation Data Scientists and Engineers, enhancing the Machine Learning algorithm development process to be lean and robust. It is the need of the hour to focus on these tools to ensure that they generate value without degrading performance in a “limited human intervention” environment. In the same vein, research study on Explainable AI (Xu et al., 2019) will be continued to make sure that the relevance of AIML always stays on a solid foundation without hype and AIML process automation meets the demands of Explainable AI.

The blog by Datarevenue, (20xx) explores the new tools for orchestrating machine learning tasks and data workflows (referred to as “MLOps”). “Argo-CD”, “Kubeflow” and “MLFlow” serve the requirements as a common data workflow tool for technology and businesses using Kubernetes, catering to the niche requirements related to deploying machine learning models and tracking experiments. The application of above tools needs further exploration from the aspects of “Operationalizing” AIML, through

Automated Machine Learning frameworks (AutoML, 2018) and frameworks for Machine Learning Operations (MLOps, 201x) which will lead to studying related frameworks, that are being popularly accepted for deployments.

Here we aim to focus our penultimate survey of literature on what could be believed to be the common platforms, algorithms, tools and techniques of AIML for next-gen technology and business domains - and those will (but not limited to only those) form the basis of further research. Cloud-native Platforms - Kubernetes has been touted as the de facto cloud-native container orchestration platform by Gaur (2021). Needless to say, all major vendors now use Kubernetes in its core to develop enterprise and telecom platforms. Further deeper research is needed on the progress of Kubernetes and its allied tools and techniques for data ingestion, Platform monitoring and data aggregation and the same will be conducted.

NextGen Technology and Strategic management based on Alliance and Partnership

Valuing companies that use AIML is a culminating end to the research that will be conducted.

The paper by R. Visconti (2019), notes that incorporating valuation of Artificial Intelligence "could be similar to the use of traditional methods of valuation of intangible assets (cost, economic or market approach)" with some adaptation. We expect to progress the research already conducted in this area focusing on companies related to 5G/6G wireless communication-related business domains.

Alongside, the valuation of companies with AIML at its core will be researched using available valuation methods as noted in a paper by Hvass Laboratories (Pedersen, 2016) to generate a framework for deriving hallmarks and metrics for formulating strengths and weaknesses of a partner ecosystem in the companies adopting AIML as a core component. Thus, to reiterate, our research goal will be to weave a theme around the common yet novel algorithms, techniques and processes of AIML that are applied in next generation technology and business domains and further seek out a framework and rules that can guide the trajectory of a winning business ecosystem and alliance formation, based on novel methodologies.

2.2 Societal Impact of next-gen Artificial Intelligence

Daugherty and Wilson (2018), in their book “Human + Machine: Reimagining Work in the Age of AI”, published in the pre-era of Generative AI, present a quite positive road-map of establishing “AI” in business and technology arena. The societal impacts are treated as positive, augmenting men with machines; expecting and suggesting appropriate measures and guardrails to be implemented to make AI more responsible. However, in 2024, with disruption of workplace by machines with next-gen “Generative AI”, the large-scale “human work” focus seems to have temporarily shifted to train the machines to generate outcomes that are gradually expected to respect the benchmarks of “responsibility”.

Resonating with the above view, the grim aspect of the impact and ramification of large-scale adoption of AI has been recently published by “The Hindu” newspaper on 5th September (The Hindu, September 2014) in India, presenting their view and update based on ‘World Employment and Social Outlook: September 2024”, published by International Labour Organization, Geneva. The report headline, as presented in the paper, says “AI key reason for Labour Income dip” and goes on to say “A major reason for this fall in labour income is Artificial intelligence of AI”. Hence, we are prompted to say that the present societal outlook is a mixed one, and the impact of AI is in a state of flux.

2.3 Summary

Thus, to reiterate, we have established that the technical impacts of AIML in business have huge potential of growth impacting the present and coming generations of businesses; however, their inclusion, do impact the society in the short-run and hence the expectation is that the Global Economic Leaders may need to take a rounded view of the pros and cons of the impacts in the longer time scale, that would include the responsible societal values and not only be influenced by the over-arching profit generation perspectives of the new and established technology companies.

CHAPTER III:

METHODOLOGY

3.1 Overview

By nature, the exploratory research is flexible but often needs to fall back on exploratory and trend-based analysis approaches (Sekaran and Bougie, 2016). Guided by the characteristics of the exploratory research and in order to address the questions presented in the problem statement, an exploratory meta-analysis and trend-based analysis will be adopted for AIML impact analysis.

In order to navigate the application of AIML, we navigate the strategy landscape with the use of the following strategic frameworks, as explained briefly below.

While Gurel and Tat. (2017) detail out the strategic management process steps for traditional technology and enterprise companies using the SWOT Analysis techniques, Business Model Canvas (Osterwalder and Pigneur, 2010), provides a high-level view, aiding in connecting the dependencies of various companies in the partner ecosystem to analyze how the value is derived out of AIML infusion at the core of the business processes and functions. Further, emerging Accelerator Business Models to aid rapid innovation, are more likely to be used for next-gen AIML developments, and they are reviewed briefly.

Ecosystem formation also needs to strategize the intermediation of various partners and create values based on AIML with the partners. Strategic and tactical plays using partnership model analysis based on dyadic and multiparty relationships may be explored through case studies.

Further, Platform Business Model Map could be used to analyze the existing and future AIML platform-based partnership eco-system to generalize and form a broader perspective of future platform-based business development.

The complete view of the impact needs scenario analysis which may be conducted from viewpoints of all participant business partners, for which Value Train Analysis (Rogers, 2016) could be used.

Exploring novel AIML techniques and evaluating them on the basis of business impacts via generally accepted frameworks for multiple stakeholders in the business ecosystem, assessing their perceived values using dyadic and multi-party partnership formation constructs followed by Platform Business Model Map analysis and Value Train Analysis methodologies will be a unique research undertaking.

3.2 Methods used

In this thesis we will employ a systematic meta-analysis via exploratory research and trend analysis (based on secondary survey research report) to evaluate the impact of AI on business and technology and adoption of novel methodologies, algorithm frameworks and real-time platforms and techniques to form the fundamental core of the partnership framework research. While the research is primarily exploratory, based on systematic meta-analysis of technology and business frameworks, analysis of trends of adoption of novel AIML technologies has also been conducted.

More concretely, we will apply four different case studies that were analysed using ethnographics and case study approach where researcher has investigated in different companies the impact of AI.

Analysis via systematic meta-analysis of novel AI technologies, study of the trends of their adoption, followed by further meta-analysis of several business and technology case studies has been adopted for the research.

Thus, following a systematic exploratory approach towards novel technology deep-dives and further analysis of novel business methodologies, substantive yearly trends have been highlighted on the adoption and impacts of Novel AIML technologies, supporting as responses to the primary research question.

To conclusively answer second research question, the study further substantiated and corroborated the initial analysis with specific technical/business case studies, highlighting the usage of novel business tools in novel technology and business setting, specifically:

- Dyadic and multi-party partnership analysis for Alliance and Partnership formation in Industries that utilize AIML.
- Multi-sided Platform Business Model analysis with Platform Business Model Map (Rogers, 2016) for analyzing businesses that utilize AIML Platform Technology to deliver services in a Platform eco-system.
- Value-Train Analysis (Rogers, 2016), for conducting tactical analysis of technical and business value delivery in rapidly changing business environment.

3.3 Data Analysis

Analysis of yearly trends and corresponding trend charts of novel technology adoptions were utilized to derive conclusions on novel AIML technology impacts on businesses.

3.4 Research Design Limitations

Due to competitive nature of these novel technology creation and deployments; and high focus on fast innovation and rapid implementation for generating high future profits, has possibilities of leading to rapid creation of “half-baked” products, generating more societal harm than good. Also, they may result in high societal cost - due to less focus on adequate imposition of responsible AI (XAI) concepts and inadequacy in testing.

The nature of recent available data, as discussed above, though from credible sources, were not considered adequate for rigorous statistical analysis to draw conclusive evidences of impacts; hence the research design was limited to meta-analysis of Technology impacts of the selected novel AIML technology for deep-dive analysis, followed by analysis of trends based on the early and current reports (HAI, 2023; HAI, 2024) and lastly, through a few specific technology/business case studies of industries, which majorly impact adoption of novel AIML and Platform Technologies.

3.5 Conclusion

The novel AIML technologies like GenAI and collaborative robots, even though being early adoption phases, are already showing promises of high levels of cost reduction when adopted by industrial operations. However, risks of fast adoption also exists, thus necessitating a high oversight by the regulatory bodies to address possible adverse societal impacts (like catastrophic failures in industries and job losses).

Hence any statistical analysis was deemed highly unlikely to yield useful results promising long term-validity as the available data was based on short and near-term analysis reports. Hence the thesis chose to adopt meta-analysis, trend analysis and technology/business case studies to draw conclusions on impacts and adoption of Novel AIML technologies.

CHAPTER IV:
META-ANALYSIS OF AIML TECHNOLOGIES AND BUSINESS FRAMEWORKS

4.1 Meta-analysis of Novel AIML Techniques

IMF (2023) takes a measured view on impacts of AI and admits, in an alternative future, where benefits out-weigh the risks of Artificial Intelligence (AI), “AI leads to a higher-productivity-growth future”.

In this expected “High-productivity future”, IMF further envisages:

“AI lives up to its promise of being the most radical technological breakthrough in many decades. Moreover, it ends up complementing workers - freeing them to spend more time on nonroutine, creative, and inventive tasks rather than just replacing them. AI captures and embodies the tacit knowledge (acquired through experience but hard to articulate) of individuals and organizations by drawing on vast amounts of newly digitized data. As a result, more workers can spend more time working on novel problems, and a growing share of the labor force increasingly comes to resemble a society of research scientists and innovators.”

Echoing the energizing views of future impacts of AI in economy, with a “permanently higher growth rate”, with an AI-enabled society and AI-backed research, enabling an augmented vision of the future, enabling “radical advances in knowledge” in technology, biology and businesses, we embark on the exciting journey of discovery of AI impacts, with an eye on past developments paving way for the new.

Economics and business requirements of AIML for Digital transformation

The impact of novel AIML technologies in transforming current businesses via intelligent digital transformation route is evidenced by the “Big Data Business Model Maturity Index” framework created by Schmarzo (2020).

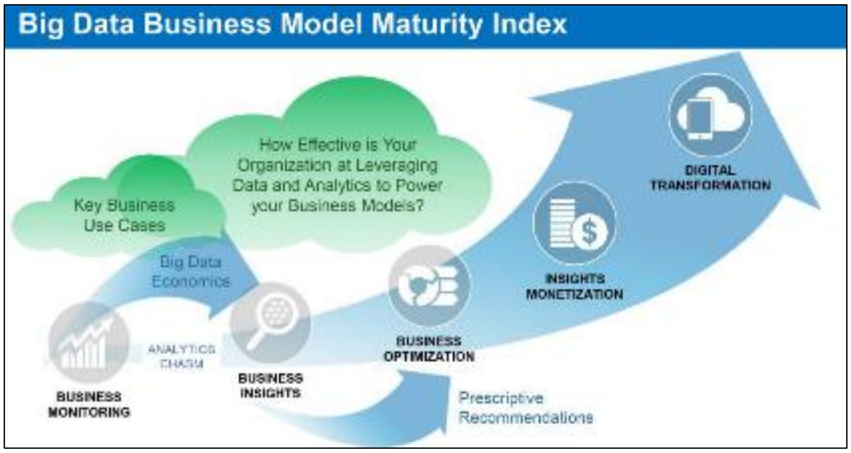
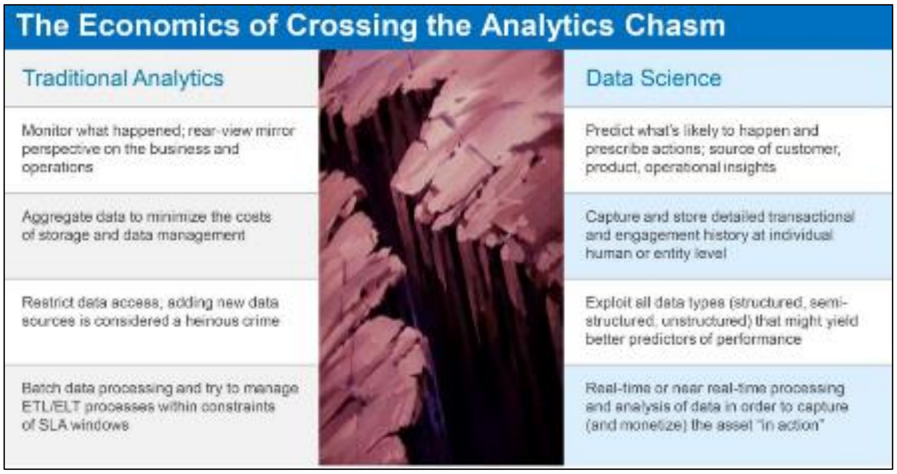

Big Data Business Model Maturity Index
How Effective is Your Organization at Leveraging Data and Analytics to Power your Business Models?
Key Business Use Cases
Big Data Economics
ANALYTICS CHASM
BUSINESS MONITORING
BUSINESS INSIGHTS
BUSINESS OPTIMIZATION
INSIGHTS MONETIZATION
DIGITAL TRANSFORMATION
Prescriptive Recommendations

Figure 4.1
Ultimate destination of big data business maturity – business transformation with AIML
Courtesy: Schmarzo (2020)

The author provides clear evidences that the ultimate aim of big data and cloud computing applications in business and technology is to transform them with the application of analytics, artificial intelligence and Machine Learning. The author further goes on to present "analytics chasm" as a phase for organizations relying on traditional rule-based predictive and prescriptive analytics to seek to move towards programmatic AIML-based models as novel models in order to optimize the operations and decision-making in businesses. He states:

"Crossing the analytics chasm is not a technical challenge; it's an economic challenge for how organizations leverage the economic value of data to derive and drive new sources of customer, product and operational value."

The author further compares and summarizes the economic value in adoption of advanced data-science-based (AIML) business compared to traditional rule-based analytics modes succinctly, as tabled below:



The Economics of Crossing the Analytics Chasm

Traditional Analytics	Data Science
Monitor what happened; rear-view mirror perspective on the business and operations	Predict what's likely to happen and prescribe actions; source of customer, product, operational insights
Aggregate data to minimize the costs of storage and data management	Capture and store detailed transactional and engagement history at individual human or entity level
Restrict data access; adding new data sources is considered a heinous crime	Exploit all data types (structured, semi-structured, unstructured) that might yield better predictors of performance
Batch data processing and try to manage ETL/ELT processes within constraints of SLA windows	Real-time or near real-time processing and analysis of data in order to capture (and monetize) the asset "in action"

Table 4.1
The Economics of Crossing the Analytics Chasm. Courtesy: Schmarzo (2020).

Schmarzo (2020) further states:

"Using Artificial Intelligence (AI), you can create assets that appreciate in value (not depreciate), the more that these assets are used."

The above comments substantiate the "economic value impact" of AI and inspirational towards further exploration urging to undertake detailed and systematic meta-analysis of technology and business frameworks and trend analysis based on early Industry Novel AIML adoption-related evidences (HAI, 2023, HAI, 2024).

Our analysis starts by conducting systematic meta-analysis of Novel AIML Technologies that are expected to revolutionize future generation businesses and some novel business frameworks to strategically and tactically analyze their adoption.

Public and Private Clouds for AIML Application development & deployment

In a public-cloud setting, the hyper-scalers provide the above easy-to-use services on a "pay-as-you-go" model (Bishai, 2018) and promises to be cost efficient thus providing a compelling reason to the business owners to build the enterprise and technical services based on these platforms rather than building in-house solutions, unless data privacy issues present a reason for preferring a hybrid-cloud solution.

In the blog article, Bigelow (2021) discusses multi-cloud vs. hybrid cloud settings for developing and deploying AIML in the transformed business settings.

The renowned machine learning blogpost from Datacamp (Cotton, 2020) compares the various traditional machine learning models, both supervised and un-supervised, recommending use in business setting.

As captured by Flach (2012) and Alpaydin (2014) in their authoritative texts on Machine Learning, the Linear models are parametric and stable while being less likely to overfit the training data than some other models, sometimes leading to underfitting (i.e., they have high bias and low variance). These models are preferred when data is limited and overfitting is to be avoided.

The above models have been treated as “error-based” learning methods in the text by Kelleher et al. (2015), bringing in a fresh perspective on classifying the ML Algorithms.

The Second class of traditional Tree-based ML Models like decision trees follow a recursive divide-and-conquer nature (Flach, 2012: Tan et al., 2012). The various types of trees and their attributed and summarized in the table in the blogpost.

Tree formation is based on classification and labelling aided by their empirical probability measured either by “Minority Class” error rate, measured as proportion of mis-classified data), or Gini Index (based on expected error, if a leaf is labelled randomly) or entropy (based on information gain on drawing a classified label randomly – higher the information gain, lesser is the predictability).

Kelleher et al. (2015) proposed to view the tree-based models and “information-based” models and thus share a novel perspective of reviewing them.

While the first and second classes are grouped under “supervised-learning” class, since they need training data to stabilize the models, the third major traditional machine learning models are normally grouped under “un-supervised” learning class as Clustering Models. To note, interestingly, this class of models have been classified under “similarity-based learning” by Kelleher et al. (2015).

Further, “Probability-based Bayesian Learning models” have been studied extensively in the standard texts (Tan et al., 2012; Kelleher et al., 2015) and their extension to traditional rule-based expert AI Systems have been studied as well (Negnevitsky, 2011).

Hastie et al. (2017) thoroughly research Support Vector Machines (SVMs) and Kernel structures in their authoritative text, as an extension to linear classification by enlarging the dimensionality of the feature space, making the classification process more “flexible”.

Traditional ML model, algorithm recommendation for Cloud-based businesses:

The paper by Botchkarev (2018) illustrates the use of Azure Machine Learning Studio, a cloud-based platform offering Machine Learning as a Service (MLaaS) for Regression analysis.

The paper declares that these MLaaS Services offered by cloud-based companies (Amazon Web Services (AWS), Microsoft Azure, Google Cloud) demonstrate the potential of expediting machine learning experiments from weeks and months to hours and days. These services generally offer a convenient integrated development environment with certain benefits, including: cloud-based machine learning offered as a service; web-based solution, with built-in ready to use regression modules, offering flexibility of using R and Python languages to code experiments; The author demonstrates the regression model evaluation to compare the trained model predictions with the actual (observed) data from the testing data set. With Azure MLS’s designated module “Evaluate Model”, comparisons can be performed with the convenience of drag and drop functionality. The author presents superior model integration to enhance the performance, which demonstrates the possibility of enhancing the tools with python or R scripts for designated high-performance application by in-house data scientists to meet the business and technology needs. The regression work-flow is captured below (Botchkarev, 2018).

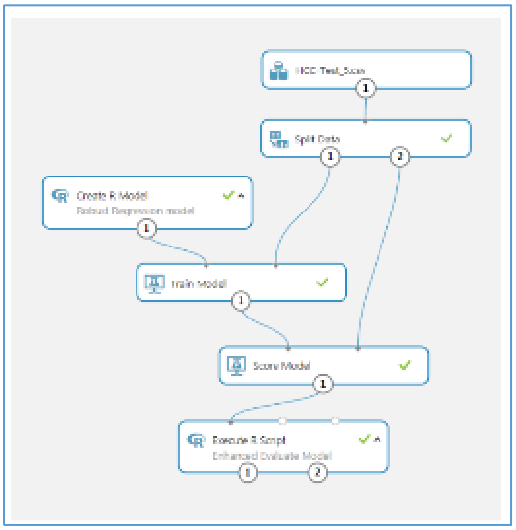


Figure 4.2
MLaaS - Regression analysis with Azure Machine Learning Studio
Courtesy: Botchkarev (2018)

Moving from traditional ML to more sophisticated Neural Network Architectures

The implementation of the Machine Learning models has moved from traditional development frameworks (e.g., object-oriented programs in Java-based language) to implementations based on shallow and deep Neural Networks frameworks for AIML productization in private or public cloud settings. The below figure (MathWorks Inc., 2019) captures the journey.

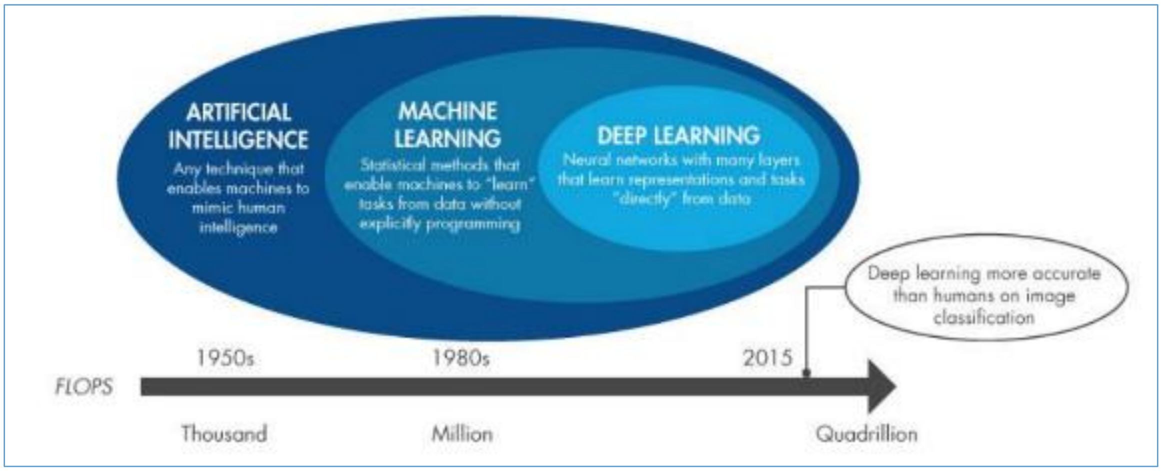

ARTIFICIAL INTELLIGENCE
Any technique that enables machines to mimic human intelligence
MACHINE LEARNING
Statistical methods that enable machines to "learn" tasks from data without explicitly programming
DEEP LEARNING
Neural networks with many layers that learn representations and tasks "directly" from data
Deep learning more accurate than humans on image classification
1950s
1980s
2015
FLOPS
Thousand
Million
Quadrillion

Figure 4.3
Machine Learning Maturity – 1950s to 2015 and beyond. Courtesy: MathWorks Inc.
Image Source: MathWorks Inc., "Deep Learning or Machine Learning," The MathWorks, Inc, 18 09 2019. [Online]. Available: https://explore.mathworks.com/machinelearning-vs-deep-learning/chapter-1-129M100NU.html. [Accessed 18 09 2019]

As exemplified in the above diagram, and explained in by Choi et al. (2020) neural networks, specifically Deep Neural Networks (DNN) represent a novel method for realizing ML and its benefits are realized maximally when several layers of it are implemented to realize training and classification of data.

Sze et al. (2017) presented a brief history of DNN with the evolution time-line, till the latest trend of DNN Accelerator research.

The superiority of Deep Learning compared to the traditional machine learning has been very precisely noted by the eminent IBM researcher, (Aggarwal 2018, p 2):

“A different view is that neural networks are built as higher-level abstractions of the classical models that are commonly used in machine learning. In fact, the most basic units of computation in the neural network are inspired by traditional machine learning algorithms like least-squares regression and logistic regression…from this point of view, a neural network can be viewed as a computational graph of elementary units in which greater power is gained by connecting them in particular ways. When a neural network is used in its most basic form, without hooking together multiple units, the learning algorithms often reduce to classical machine learning models".

The author further authoritatively claims that the deep learners tend to become more attractive than conventional methods primarily when sufficient data and computational power are available. The claim seems to be aptly supported by the massive deployments of the cloud-based AIML systems in the recent years, with massive increase in data availability and computational power, that supports the author's claim of a "Cambrian explosion" in deep learning technology, as illustrated in the below figure.

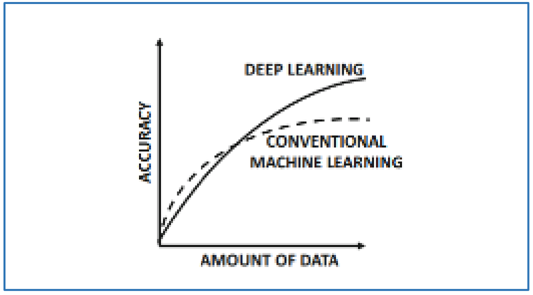


Figure 4.4
An illustrative comparison of the accuracy of a typical machine learning algorithm with that of a large neural network. Courtesy: Aggarwal (2018)

However, the performance of traditional machine learning may remain superior at times for smaller data sets because of more choices, greater ease of model interpretation, and the tendency to hand-craft interpretable features that incorporate domain-specific insights. With limited data, the best of a very wide diversity of models in machine learning is expected to perform better than a single class of models, like NNs (Mahony et al. 2019).

Thus, we gather the potential reason behind the continuing deployments of traditional ML algorithms, in the context of small data, in small-and-medium (SMB) scale businesses and local retail stores.

The above evidences and results lead us to accept neural network and its further enhancement and evolution as the bedrock of novel algorithm and processes; and their implementation techniques to form the new base-lines of adoption in the current and future industries, thus playing a lead role in revolutionizing the AIML adoption landscape.

Deep Learning Architecture

As described in the paper (Choi et al., 2020), biological neural networks helped to conceive the idea of artificial neural network (ANN). Each ANN consists of nodes that communicating with other nodes in the network via connections, which are weighted based upon their ability to provide a desired outcome. The architectural ideas are captured in the paper by Choi et al. (2020).

For most ML tasks, ANNs feed information forward (Feedforward Neural Networks), as opposed to recurrent neural networks, where information can be passed between nodes within a layer or to previous layers. The basic building block of a neural network is a single-layer “perceptron”, which acts as a logistic regression algorithm in traditional Machine Learning that takes in a series of features and their targets as input and attempts to find a line, plane, or hyperplane that separates the classes in a two-, three-, or hyper-dimensional space, respectively. Multilayer perceptron can be formed by stacking the perceptrons as a layer of input nodes, a layer of output nodes, and a number of “hidden layers” between the two, where shallow networks consist of mostly three hidden layers and the rest are considered deep neural networks, containing hundreds of hidden layers, as per need. The features, provided to the inputs layer are transformed using various activation functions like sigmoid function etc.

Training in Deep Learning Networks

Neural network processes data in two distinct phases, the training phase and the inference phase, which has been well captured in the paper by Kriegeskorte and Golan (2019).

During training of a network, the parameters (the set of weights and biases), are iteratively updated to achieve the desired output. This is commonly done using a loss function and backpropagation ultimately followed by hyper-parameter optimization (Kriegeskorte and Golan, 2019; Sze et al., 2017; Skansi, 2018; Pettersson, 2020).

Deep Neural Network Derivatives:

A summary history of Deep Learning Networks with their key descriptions is captured in the paper by Pouyanfar, Sadiq and Yan (2018). In the survey conducted by the authors, challenges and opportunities in key areas of deep learning, like parallelism, scalability, power, and optimization have been raised and explored respectively; and they propose different kinds of deep networks in different domains such as RNNs for NLP and CNNs for image processing. Further, the paper introduces and compares various popular deep learning tools including Caffe, DeepLearning4j, TensorFlow, Theano, and Torch and the optimization techniques in each deep learning tools.

Further, superiority of Deep Learning Computer vision applications has also been explored Mahony et al. (2019). Key findings of this article and future developments suggested, based on the current weaknesses of Deep Learning frameworks are captured below:

- The majority of the existing deep learning implementations are supervised algorithms, while machine learning is gradually shifting to unsupervised and semi-supervised learning to handle real-world data without manual human labels.
- Unlike human brains, deep learning needs extensive datasets (preferably labeled data) for training the machine and predicting the unseen data. With the availability of "small" datasets (e.g., healthcare data) or when real-time processing of data is required, DL has been under-performant. Recent studies on "One-shot learning" and "zero-shot learning" a have been recently studied to alleviate this problem.

Application of Deep learning on mobile devices have been extensively reviewed in the papers (Mahony et al., 2019; Deng, 2019).

Mahony et al. (2019) provides an extensive view of efficient deep learning methods, systems and applications applicable for mobile devices, introducing “popular model compression methods, including pruning, factorization, quantization as well as compact model design” and then from design cost optimization perspectives, they discus “the AutoML framework for each of them, such as neural architecture search (NAS) and automated pruning and quantization”; finally covering “efficient on-device training to enable user customization based on the local data on mobile devices”. Various nuances of exploitation of algorithm efficiencies like sparsity and temporal/token redundancy are explored and showcase “several task-specific accelerations for point cloud, video and natural language processing” and finally “introduce the efficient deep learning system design from both software and hardware perspectives”.

Deng (2019) explores various hardware architectures for mobile deep learning including FPGA, ASIC and GPUs which are then are compared for DL algorithm optimizations, such as “quantization, pruning, compression, and approximations that simplify computation while retaining performance accuracy”; further, various “resources for mobile deep learning practitioners, including tools, libraries, models, and performance benchmarks” are discussed.

Explainability in the context of deep learning has been explored in the paper by Shahroudnejad (2021), where the claim in the abstract of the paper is the following:

“…explainable artificial intelligence (XAI) domain, which aims at reasoning about the behavior and decisions of DNNs, is still in its infancy. The aim of this paper is to provide a comprehensive overview on Understanding, Visualization, and Explanation of the internal and overall behavior of DNNs.”

Further, Leventi-Peetz (2022) in a very recent paper, investigates “non-determinism of Deep Learning (DL) training algorithms and its influence on the explainability of neural network (NN) models” and further by experimenting in the context of image classification, compares the classification using two convolution neural networks to serve as “exploration of the feasibility of creating deterministic, robust DL models and deterministic explainable artificial intelligence (XAI) in practice.”

Applications of Deep Learning in Industry

Hammad, El-Sankary and Gu (2019) demonstrated the application and performance of various machine learning algorithms, including those based on Neural Networks in their paper published in IEEE journal. The illustration in the paper is relevant for fully autonomous mobile robot applications that are used in various industries like nuclear power plant, oil refineries, chemical factories, and military applications.

Natural Language Processing with deep learning algorithms have thus been widely researched and implemented for Conversation AI and chatbot Applications to comprehend intent, context and sentiment and further to simulate conversations. Typical business usage includes messaging applications, web-based marketing, and mobile apps. Use of Deep Learning algorithms like CNN, RNN and LSTM to implement such chatbots abound and have been captured in the literature (Dhankhar, 2018; Tsakiris et al., 2022).

Tsakiris et al. use CNNs as the classifier and utilizing several tokenization or “Word Embedding” techniques like AlexNet, LeNet5, ResNet and VGGNet, evaluate suitable architectures, comparing their accuracy, f1 score, training time and execution time. They report LeNet-5 to have the best accuracy compared to other architectures, the fastest training time, and the least losses on large datasets.

Implementation using “sequence-to-sequence model” consisting of two recurrent neural networks (RNNs), with the encoder processing the input and a decoder generating the output has been studied in paper by Dhankhar (2018). He explored a RNN model, with attention mechanism technique that allows decoder more direct access to hidden state output by the encoder allowing to process sequential temporal information, as in sentences in conversations more efficiently compared to CNN-based implementations. The paper highlighted the use of stacked LSTM cells, with two RNN cell layers – the encoder being the utterance by human and the decoder, the chatbot response.

From Mobile network related application perspective, stacked bi-direction and unidirectional LSTM RNN networks have been exploited (Trinh, Giupponi and Dini, 2018) for “predictive analysis on mobile network traffic”, which has assumed fundamental importance for the NG-Wireless networks to assess proactively, the user demands for optimal wireless resource allocation. The RNN-based application was based on a particular LTE BS’s mobile traffic, with “mobile traffic information gathered from the Physical Downlink Control CHannel (PDCCH) of the LTE” using passive data collection tools. The design of the prediction system used LSTM to solve the sequential data case, with the problem stated as “a supervised multivariate prediction of the mobile traffic”, with the objective function defined to minimize the prediction error given the information extracted from the PDCCH. Both single step and long-term prediction errors were deduced; different numbers for the duration of the observed values resulted in determining the memory sizes of the LSTM network which further helped to determine information storage requirement for precise traffic prediction scenarios.

For prediction of network-wide traffic speed, study was conducted in the paper (Cui et al., 2019), utilizing deep stacked bidirectional and unidirectional LSTM (SBULSTM) architecture to consider both forward and backward dependencies in time series data.

Exploitation of the full depth of the bidirectional LSTM (BDLSTM) model architecture allowed to harness the predictive power to cater to the spatio-temporal data gathered from the large prediction area, capturing spatial features and bidirectional temporal dependencies from historical data to predict traffic speed for both freeway and complex urban traffic networks has been a novel and unique endeavor. The robustness of the model was further achieved using masking mechanism to handle missing values in input data.

Thus, these wide range of diverse applications in futuristic industrial domains provide us the evidence of the importance of Deep Neural Network algorithms in engineering next generation businesses and technologies.

Novel AIML Frameworks, Architectures and Methodologies

The adoption of novel AIML algorithms and frameworks have been governed by economics of scale and changing trends in technology landscape and business environment. As discussed above cloud-based application are driving the demand of AIML technologies that can be adopted in cloud-native environments on private or public cloud. AIML technology, is rapidly progressing to help businesses generate business value and this is shown in the recent report on “Hype Cycle for Artificial Intelligence” as shown in the figure below from Gartner (Casey, 2023).

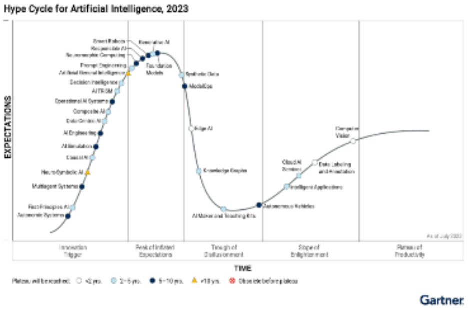


Figure 4.5
AI hype cycle – 2023. Courtesy: Gartner (Casey, 2023)

The report identified Edge Analytics, Autonomous AI systems, Generative AI and Platform-based AIML Engineering to be the prime drivers of next generation business and technology eco-system growth.

In our thesis, we identify “Distributed Agent-based edge computing and analytics Systems”, “Distributed Federated Learning frameworks for Autonomous Networks” and “Attention-based Natural Language Processing for Generative AI based on Deep Reinforcement Learning and Transfer Learning Frameworks”, all based on deployments on cloud-native Intelligent Platforms, to be the key components of these driving technologies and focus our research on them to illustrate their application in various technology and business settings.

The choice is further motivated by the below figure by RCRwireless and updated further to distinguish the impact of Novel AIL technologies, including Generative models for AI on Wireless Operators who are fast moving to encompass the enterprise business operating models (Kinney, 2023). Thus, this serves as an industry-aligned roadmap and provides impetus for researching on some of these specific novel AIML technologies.

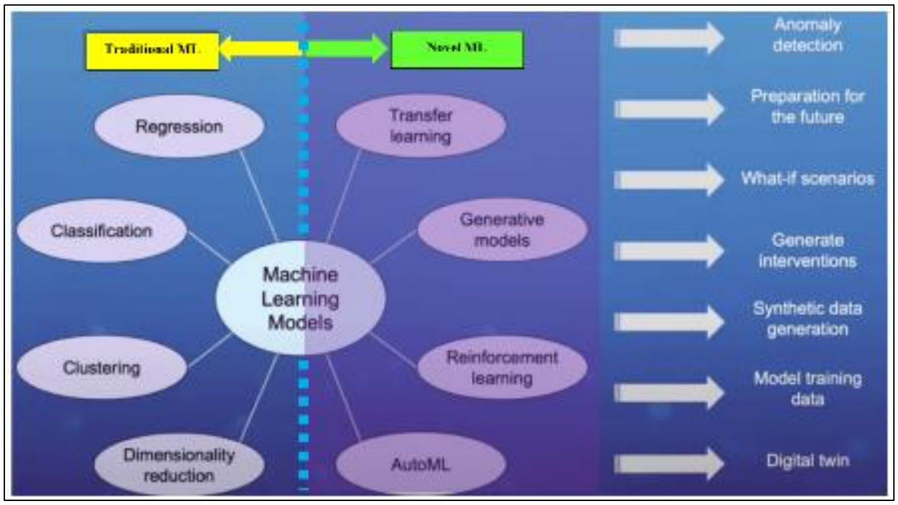


Figure 4.6
Impact of GenAI – 2023. Courtesy: Image Source: RCRWireless News (Kinney, 2023).

Deep Reinforcement Learning and Transfer Learning Frameworks

Deep Reinforcement learning originates from the rich formulation of Reinforcement Learning Framework, which indeed is novel and transformative in tackling machine Learning problems that are “markedly different from the supervised and the unsupervised varieties” and are those that can be framed “as having an agent take a sequence of actions within some environment” (Krohn, 2020, p 54).

As explained in the survey paper (Ssengonzi, Kogeda and Olwal, 2022), the policy function may map states to actions via off-policy RL, like a Q-learning algorithm (Aggarwal 2018, pp. 387 -397), evaluates and improves the update policy separately from the behavior policy, leading to the advantage of having a deterministic (e.g., greedy) policy update with the behavior policy, separately scans for all the possible actions. The policy assists in gathering and storing experiences of agent-environment interaction in a replay buffer to further update the target policy and the agent’s subsequent interaction with the environment using this new policy.

In on-policy RL, the update policy and the behavior policy essentially are same; where the collected samples of the agent-environment interactions are used to improve the same policy with the agent using for selecting actions for the subsequent interaction. Renowned on-line algorithms include SARSA, PPO, TRPO etc.

Offline or batch RL uses a phenomenon like supervised learning, where the agent uses previously collected data, without additional online data collection. There being no agent-environment interaction, and hence having no behavior policy to collect additional transition data, it resorts to utilizing a static dataset of fixed interactions with the environment, learning the best policy possible from it.

The survey paper (Ssengonzi, Kogeda and Olwal, 2022) further notes two successful algorithmic frameworks of RL – Model-based RL (MB-RL) and Model-free

RL (MF-RL), as critical components of the RL framework, as illustrated in the figure below.

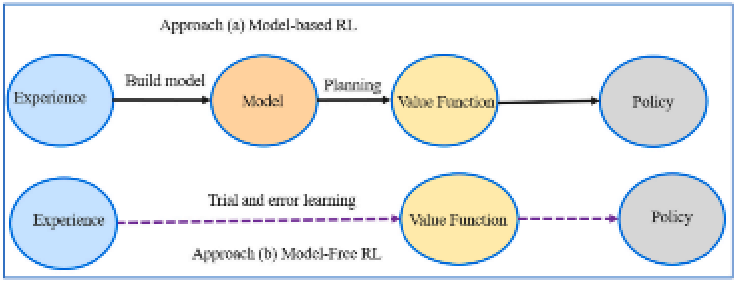


Figure 4.7
Model-based and Model-Free Reinforcement Learning. Courtesy: Ssengonzi, Kogeda and Olwal (2022)

The model of the environment in the MB-RL approach, known "state transition matrix" and "reward" prediction or distribution exist, allowing utilization of "next state" and predicted "reward" to estimate the value function (typically, Policy iteration and Value iteration) to calculate the optimal action before taking action. As noted in the paper **(**Ssengonzi, Kogeda & Olwal, 2022; Ho & Lee, 2015; Kaelbling, Littman & Moore, 1996), MB-RL approaches prove to be advantageous in real world sequential decision-making problems, by yielding good results with optimized data utilization of data leveraging domain knowledge programming to accelerate learning, assisting the learned model may assist the system when the objectives change.

However, the learning performance of the MB-RL model is limited by the accuracy of the model of the underlying learning system which may be time consuming to achieve in dynamic environments; and often with environment and resulting data drifts resulting in rules that may need re-formulation and re-optimization.

In MF-RL, the "state transition matrix" and "reward" are unknown to the agent, contrary to MB-RL, and hence the "next state" and "value" are unknow before taking the action; but based directly on interaction with the environment, MF-RL collects the set of trajectories that provides the requisite experience data for the agent to enhance its learning to estimate a value function or an optimal policy.

Lately the MF-RL algorithm, Q-learning (Sutton and Barto, 2018) which is currently finding many applications in latest 5G network optimization techniques (Mismar, Evans and Ahmed, 2020), the agent estimates the Q-value or the approximate value function for each (state, action) pair and provides the optimal policy by selecting the action that gives the highest Q-value given the state of the agent. As like all the MF-RL algorithms, Q-learning techniques require a vast wealth of experience and hence recently collected data for optimal performance.

The survey paper (Ssengonzi, Kogeda & Olwal, 2022) further emphasizes leveraging a combination of MB-RL and MF-RL techniques for a more holistic implementation to mutually counter-act on the individual model weaknesses. As per the paper, while MF-RL approach requires a complete exploration of the environment, proving to render in-efficient learning and algorithm convergence in complex application like 5G networks in limited time. Finally, the authors of the above survey paper propose a novel ML technique, Deep Q-learning, that approximate the Q-values with a DNN, to overcome the above challenge, enabling a complete exploration thus minimizing the approximation loss of the DNN.

Deep RL evolved from the use of DNNs (non-linear method) to approximate either the value function, dictating the efficacy of the states and actions; or the policy π, dictating agent’s conduct; or the model, dictating state transition function and reward function of the given environment. Deep RL uses deep learning (DL) tools to extract features from complex high dimensional data and transforms them to a low-dimensional feature space, and then uses RL to make the decisions, as illustrated in the figure below (Ssengonzi, Kogeda & Olwal, 2022).

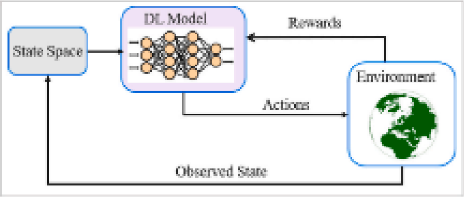


Figure 4.8
Model-based and Model-Free Reinforcement Learning. Courtesy: Ssengonzi, Kogeda and Olwal (2022)

In summary, based on the above figure, DNNs assist the agent to extract the most relevant features from the state representation. The sought parameters, to be learned for the Deep RL, are the weights in the DNN which are updated using Stochastic gradient descent. As discussed in the book by Aggarwal (2018), typical examples of DNNs employed in Deep RL include CNNs, RNNs and Deep Q-Networks. To note, the DQN applied DL to Q-learning, by approximating the Q-table using a deep neural network

The papers (Lowe et al., 2020; Sewak, 2019) discuss the Deep Q-Networks (DQN), developed at Google's 'Deep Mind' combining the Q Learning algorithm in Reinforcement Learning incorporating Deep Learning ideas to enable the concept of DQNs, which exemplify off-policy RL algorithm, the policy or its Q function is updated by training the replay buffer.

From the perspective of application to 5G, the survey paper (Ssengonzi, Kogeda & Olwal, 2022) states that the core of the study is related to Deep RL as a candidate for current and future use in heterogenous, dynamic Intelligent Wireless Networks for network slicing. The paper notes:

"…the DQL agent may function effectively and make resource allocation choices in a timely manner based on its already learnt policy. That way, complex slice orchestration and resource allocation challenges in 5G and beyond network slicing might benefit from such a strategy."

We further embark into the Agent-based frameworks to connect with the various deep learning frameworks.

Distributed Agent-based Frameworks & Computing Systems

In the paper **(**Vicente, J. and Vicente, B., 2019), the authors give vent to the trend in the use of multi agent systems (MAS) in the technology and business contexts:

"With the current advance of technology, agent-based applications are becoming a standard in a great variety of domains such as e-commerce, logistics, supply chain management, telecommunications, healthcare, and manufacturing. Another reason for the widespread interest in multi-agent systems is that these systems are seen as a technology and a tool that helps in the analysis and development of new models and theories in large-scale distributed systems or in human-centered systems…."

They also qualify MAS as being "distributed" and "open" and describe their "particularity" lying in the fact that their "components as autonomous and selfish, seeking to satisfy their own objectives, and attribute their importance to their ability to enable autonomous operations of technologies in complex and dynamic domains that use complex applications requiring distributed and parallel processing of data.

Historically, BDI Framework from the perspective of "rational agents" have been researched (Rao & Georgeff, 1995) to unify ideas of implementing scalable agent-based systems based on theoretical foundations.

Further, the papers (Guerra-Hernández et al., 2004; Silva, Meneguzzi and Logan, 2020), share the evolving perspectives of architectural and learning aspects of the BDI agents, leading to a symbolic representation of the agent-framework, originating from the Belief-Desire-Intent (BDI) agent models was built by RMIT University (Padgham, 20xx).

The BDI system presented in the RMIL University presentation slides of Padgham provided an operational view, where a "Plan", selected from the "Plan library" by the BDI execution engine, was expected to be a recipe to achieve a goal in a specific intended scenario posed by the environment. The study by RMIT University, revealed that the BDI Agent-Oriented Programming (AOP), originally proposed by Shoham (1993), provided abstraction at the level of mental attitudes to explain the operation of a system and were faster to develop, compared to object-oriented Java programming.

In the context of Telecom and other complex engineering systems, the paper (Xie and Liu, 2017) studies agent-based technology as a powerful tool for engineering applications, addressing the multi-agent systems (MASs) as a computational paradigm for these complex telecom systems, providing a good solution for distributed control schemes that are "desirable for managing and utilizing these devices, together with the large amount of data". The paper further compares selected open-source multi-agent platforms, frameworks, and simulators.

The paper by Torreno et al. (2017) emphasizes on cooperative Multi-Agent Planning (MAP) and formalized the MAP tasks declaring the necessity of undertaking the formalization process:

"From this Distributed Artificial Intelligence (DAI) standpoint, MAP is fundamentally regarded as multi-agent coordination of actions in decentralized systems."

While the MAS-based researches continued to dominate the complex engineering industries for solving specific scenarios, the efficacy of BDI-based MAS were presented articles by Georgeff and others (Georgeff, Wooldridge and Tambe, 1970; Georgeff et al., 1998) where they noted that "BDI Model in agency getting dated", a concern raised probably due to "object-orientation" of agent-based solutions, based on Java Programming, subsequently leading the technology community to undertake various attempts to find novel MAS systems.

Eventual evolution of agent based decentralized MAS solutions has been noted in the paper by Ponomarev and Voronkov (2017). The voice out the necessity of distributed agent-based interactions, and they note:

"…to solve any complex problem, as a rule, the interaction of agents is required, which is inseparable from the formation of MAS. The tasks in MAS are distributed between the agents, each of which is considered as a member of the group or organization. Distribution of tasks involves assigning roles to each of the agents, the definition of measure of its responsibility and requirements to its experience…"

Further, evolution was noted due to cloud-based "micro-services" oriented distributed agent system (Higashino, Kawato and Kawamura, 2018; Collier et al., 2019). The papers provide guidelines to refactor complex monolithic services using MAS systems, exploring the intersection between microservices and Multi-Agent Systems (MAS) and introducing the notion of a new approaches to build Multi-Agent Micro-Services (MAMS).

The evolution of the use of agent-based simulation modeling have been captured in the papers (Rand and Evanston, 2006; Sivakumar et al., 2022). While the former focused on early use of agent-based simulation for modeling, the latter proposes Agent-based simulation model for biomedical systems, integrating them with Machine Learning to derive and determine Agent rules.

The paper by Sniezynski (2008) proposed a centralized learning architecture, where an agent got percepts from an environment, and executed actions to interact with the environment, emphasizing the use of learning modules by the agent to improve the performance, defining it via four-tuple: (Learning strategy, Training data, Problem, Answer), stating that the details of the learning modules should be "domain-specific".

The above efforts of the industry further led to various learning agent architectures, especially of note, is the one proposed by Russel and Norvig (2003), applicable for use in distributed environments is noted in the paper by (Zhang et al., 2018).

From an application perspective, Zhang et al. (2018) proposed a "unified intelligence-communication (UIC)" model design of a learning agent, unifying various models for describing a single agent and any multi-agent system.

Interestingly, the authors in their paper tie the human brain model, proposed "as a collection of resources" by Marvin Minsky, as stated by the paper, with the learning agent architecture proposed by Russel and Norvig (2003).

Summarily, as in the context of the above architecture of learning agent, the paper (Girardi and Leite, 2013) notes that the percepts are expected to be used for actions as well as to prepare for improved future actions; where the learning element can at runtime, as opposed to pre-programmed basic agents, change their behavior in accordance to the changes in environment.

The behavior of the "performance" component, a basic agent that perceives and acts on the environment is progressively improved by the "learning" component using feedback from the "critic" component, based on the agent's conduct, determining modifications of goals and actions for better future performance of the performance component. The critic constantly updates the learning component about the agent's success (rewards) based on certain fixed performance criteria. The necessity of the critic arises from the fact that the percepts themselves do not provide indication of the agent's performance.

Problem generator component in the learning agent is a suggester for "exploratory actions", that are expected to lead to new and informative experiences, enabling the agent to further learn the environment beyond the one it already knows.

Further, catering to the business and technology marketing domains, Roshan Khan (Khan, 2017) has presented a contemporary architecture based on agent framework, of a “Customer Service Chatbot”; allowing to chart the steady enhancements in the development of Chatbots of the present generation, leading further to those developed by next generation Generative AI technologies.

Distributed Multiagent-based RL Frameworks

The deep connection between Game Theory multi-agent RL (MARL) have been very researched and its value demonstrated in various application. As explained in the paper by Yang and Wang (2021), in a multi-agent scenario, the agents solve the sequential decision-making problem through a trial-and-error procedure. However, evolution of the environmental state and the reward function that each agent receives then, is determined by all agents’ joint actions.

As a result, as the agents need to take into account and interact with not only the environment but also other learning agents.

The fundamental challenges involved in adjusting the single-agent framework to a multi-agent framework are discussed in the papers (Canese et al., 2021; Zhang et al., 2021) and they are typically those of non-stationarity of the environment, speed of convergence and scalability.

Zhang et al. (2021) further note that there is an increasing interest in developing planning/learning algorithms for decentralized POMDP (dec-POMDP), wherein “most of the algorithms are based on the centralized-learning-decentralized-execution scheme” where, the paper notes the algorithm working, which is summarily expressed below:

“…the decentralized problem is first reformulated as a centralized one, which can be solved at a central controller with (a simulator that generates) the observation data of all agents. The policies are then optimized/learned using data, and distributed to all agents for execution.”

To note, these algorithms require “super-exponential time” to solve in the worst case as, having no access to other agents’ observations, individual agent is incapable of maintaining a global belief state, the “sufficient statistic” for decision making in single-agent POMDPs. The state-of-the-art learning mechanism in MARL, namely centralized training decentralized execution (CTDE) has been formally defined in (Gronauer and Diepold, 2021) as follows:

“In CTDE, each agent holds an individual policy, which maps local observations to a distribution over individual action. During training, agents are endowed with additional information, which is then discarded at test time.”

The above paper also considers the reward function structure to define co-operative or competitive behaviors by motivating mutual cooperation or by motivating agents to outperform their adversary counterparts of the agents respectively.

The decision-making process that involves multiple agents, thus are usually modelled through Markov games (Littman, 1994). The salient recommendations and methods for achieving an efficient algorithm are captured below:

- Use model-free RL algorithms as efficient alternative to using dynamic programming.
- For multi-agent systems, a de-centralized partially-observable Markov decision process (Dec-POMDP), a generalized but decentralized MDP is suggested, as that considers the uncertainty regarding the state of a Markov process in a MAS setting, allowing a state information acquisition.
- Use RL algorithm that originates from the combination of policy-based and value-based methods, this is the actor–critic approach, where the Actor is responsible for

the generating actions and the Critic assists in updating Value/Policy functions, has been well presented and captured by Canese et al. (2021).

Incidentally, the above strategy has be proposed by Mnih et al. (2016) for the A3C MARL algorithm, famously used for training Atari Games.

Multi-objective reinforcement learning (MORL) has been studied as an extension of RL in the survey paper (Felten, Talbi and Danoy, 2023), applicable in cases where policies need to be formulated in a compromised situation of conflicting objectives. These cater to complex real-life industrial decision-making, where the optimal decision is expected to be arrived at, balancing conflicting objectives. The complex real-time decision options are characterized by their corresponding set of trade-offs, including factors like time, cost, and environmental impact, choices being influenced by conflicting KPIs in cases of wireless network operations or personal values and preferences, while considering marketing choices based on social-media. Multi-objective computational methods thus are a corner-stone in enhancing next generation business and technology decisions.

The paper (Yang, Sun and Narasimhan, 2019) introduce a novel generalized version of the Bellman equation-based multi-objective reinforcement learning (MORL) algorithm with linear preferences, with the goal of enabling few-shot adaptation to new tasks, aiming to learn policies over multiple competing objectives with their relative importance (preferences) unknown to the agent. The proposed algorithm reduces dependence on scalar reward design and further learns a single parametric representation for optimal policies over the space of all possible preferences; post an initial learning phase, the agent executes the optimal policy under any given preference, or automatically infer an underlying preference with very few samples. These novel MORL designs are expected to enable automatic closed loop control of industrial systems.

This paper (Nguyen et al. 2020) introduced an updated version of MORL, a scalable multi-objective deep reinforcement learning (MODRL) framework based on deep Q-networks, aimed at developing a MODRL high-performant framework, supporting both single-policy and multi-policy strategies, as well as both linear and non-linear approaches to action selection. The proposed framework was claimed to be generic and highly modularized, allowing the integration of different deep reinforcement learning algorithms in different complex problem domains, overcoming various disadvantages involved with standard multi-objective reinforcement learning (MORL) methods. The novel framework was proposed as a testbed platform that envisioned accelerating the development of MODRL for solving increasingly complicated multi-objective problems, an important requirement to deliver efficient futuristic autonomous networks.

From the perspective of application of reinforcement learning in business, the papers (Zhu et al., 2023; Bär et al., 2019) register evidences of their usage for semi-autonomous real-time scheduling of robotic agents in factory floors targeting Smart Factories and Manufacturing systems.

MARL is also being extensively experimented for futuristic industry 4.0 application like autonomous driving and evidences of it being proposed exist (Boyali, Hashimoto and Keihanna, 2020; Zhou et al., 2020).

Hierarchical Reinforcement Learning (HRL) has been a well-researched topic (Ribas-Fernandes et al., 2011) and its various algorithmic options have been studied and compared in the blogsite (Yannis, F-B., 2019).

However lately, it has come into prominence through novel industrial application as proposed in the abstract of the paper (Luo, Zhang and Fan, 2021) as a solution approach with hierarchical multi-agent DRL-based real-time scheduling, named hierarchical multi-agent proximal policy optimization (HMAPPO), applicable to automation on Smart factory settings and ideal for industry 4.0 application. Summarily, the proposed method, containing three proximal policy optimization (PPO)-based agents, operated by objective agent, job agent, and machine agent, with the objective agent acting as a higher controller periodically determining the temporary objectives to be optimized; job and machine agents, acting as lower actuators, choose job selection rule and machine assignment rule respectively, to achieve the temporary objective at each re-scheduling point.

We can thus draw the conclusion that the innovation may be considered timely and apt for industry 4.0 applications and use-cases including smart manufacturing factories in the domains of aerospace product manufacturing and steel manufacturing, where dynamic events frequently occur, and each job may contain several operations subjected to the no-wait constraint.

Federated Learning with Distributed Multi-Agents Agents and DRL

The paper (Qi et al., 2021) clearly expresses the aims of Federated Learning (FL) as an effort aiming to "build a joint ML model without sharing local data, involving technologies from different research fields such as distributed systems, information communication, ML and cryptography". The paper captures the two prevalent architectures, as shown below:

The key characteristics of FL are explained to achieve distributed, reliable and secure communication of data, enabling peer-to-peer or central server-based machine learning model update as per demand.

Qi et al. (2021), in their paper further explore the lacuna of distributed RL and parallel RL algorithms that usually need to collect all the data, parameters, or gradients from each agent in a central server for model training, which thus have chances of agent information leakage and possibility of not protecting agent privacy during the application of RL. They further note that in an environment of dis-trust can be a major bottleneck for such RL applications. FL, as discussed, has privacy preserving secure architecture that can adapt to agents and environments. Again, RL presents a "simulation-reality gap", as the algorithms require pre-training in simulated environments as a prerequisite for application, deployment, thus are unable to accurately reflect the environments of the real world. FL, aggregating information from both environments are more likely to bridge the gap between them. Finally, while RL may use POMDP for observing partial features, it is enough to obtain sufficient information required to make decisions, while FL, makes it possible to integrate this information through aggregation.

Given the above challenges they propose the idea of federated reinforcement learning (FRL), the genesis of which they explain below:

"As FRL can be considered as an integration of FL and RL under privacy protection, several elements of RL can be presented in FL frameworks to deal with sequential decision-making tasks. For example, these three dimensions of sample, feature and label in FL can be replaced by environment, state and action respectively in FRL. Since FL can be divided into several categories according to the distribution characteristics of data, including Horizontal Federated Learning (HFL) and Vertical Federated Learning (VFL), we can similarly categorize FRL algorithms into Horizontal Federated Reinforcement Learning (HFRL) and Vertical Federated Reinforcement Learning (VFRL)."

In the presentation at CMU, McMahan (2019) presented FL in a cross-device setting, based on the client-server model, positioning it as a means for data to thrive at the edge, targeted at billions of mobiles and IoT devices that constantly generate data, federating which, should enable better products and smarter models; the cross-device setting is captured in the figure below by McMahan (2019).

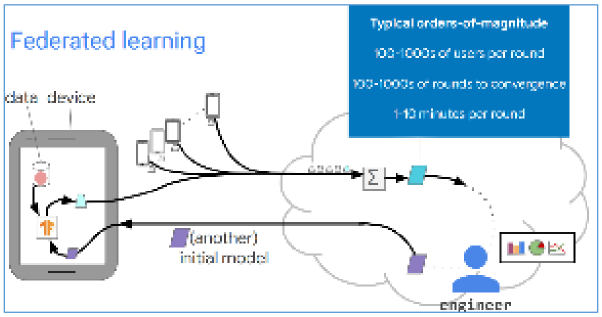

Federated learning
Typical orders-of-magnitude
100-1000s of users per round
100-1000s of rounds to convergence
1-10 minutes per round
data
device
Σ
(another)
initial model
engineer

Figure 4.9
Federated Learning in Cross-Device Setting. Courtesy: McMahan (2019)

The presentation proposes enabling on-device inference for mobile keyboards and cameras that should enable offline processing and further resulting in providing improved latency, with battery life-saving and privacy advantages.

Aligned with the cross-device setting, the paper (Bonawitz et al., 2019) reported building a scalable production system for Federated Learning in the domain of mobile devices, based on TensorFlow for large corpus of decentralized data. This paper describes the resulting high-level design, sketch some of the challenges and their solutions, and touch upon the open problems and future directions thus formalizing the FL as a scalable system.

The papers (Park, 2022; Li, He and Song, 2021) further explore FL in an alternative cross-silo setting, scenarios that are mostly presented in businesses, in medical or financial institutes. In one such scenario, multiple parties may need to collaboratively learn a model without exchanging their data. The author presented a novel one-shot federated learning (i.e., federated learning with a single communication round), contrary to the prevalent approach, applicable in cross-silo setting in practice. They further expect to alleviate the problems of existing one-shot algorithms, that use only support specific models and do not provide any privacy guarantees, which significantly limit the applications in practice, by proposing algorithm named FedKT, by utilizing the knowledge transfer technique; applied to any classification models, they can flexibly achieve differential privacy guarantees.

In the paper Yang et al. (2019) note the following in the context of privacy concerns in FL:

"One is that in most industries, data exists in the form of isolated islands. The other is the strengthening of data privacy and security."

They go on to propose a possible solution to these challenges of secure federated learning, beyond the security assumed by established organizations, by introducing "privacy-preserving decentralized collaborative machine learning techniques. The paper uses vertical federated learning operating in a cross-silo setting to achieve the purpose, and is shown in the figure below:

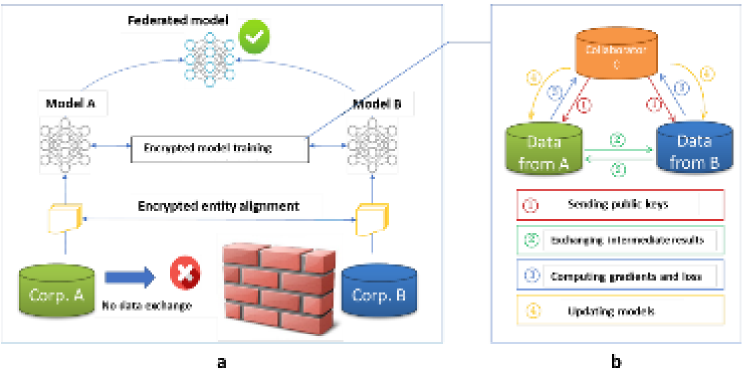


Figure 4.10
Vertical Federated Learning in Cross-silo Setting. Courtesy: Yang et al. (2019).

The architecture, as explained, is suited to a "smart retail" whose purpose to use ML techniques to provide customers with personalized product recommendation and sales services.

With FL, the competing concerns of protection, privacy and security of heterogeneous data together with that of data barriers between the three parties - banks, social networking sites, and e-shopping sites are guaranteed, as data cannot be directly aggregated to train a model.

Thus, by exploiting the characteristics of federated learning, the proposed architecture serves the co-operating purpose of three parties, without exporting the enterprise data, which not only fully protects data privacy and data security, but also provides customers with personalized and targeted services and thereby achieves mutual benefits.

The architecture also proposes to leverage transfer learning to address the data heterogeneity problem and serves as a template to build a cross-enterprise, cross-data, and cross-domain ecosphere for big data and artificial intelligence.

In the paper (Hu et al., 2021), further perspectives of classification of distributed and shared Machine Learning methods with respect to Federated Learning is shared.

The paper by Chandiramani et al. (2019) sheds light of the performance analysis of Distributed and Federated Learning Models on Private data. Both the papers are directional for Federated Machine Learning developers and testers aiming to architect and design domain-specific solutions in business and technology.

Transfer Learning

Wikipedia and Journal references (‘Transfer Learning’, (n.d.); Zhuang et al., 2020) share a common minimalistic theme of transfer learning, as a “research problem in machine learning (ML)” that, utilizing different but related stored knowledge gained from “source domains”, seek to transfer the same to a different “target domain” by solving or augmenting solution to its unique problem.

The paper by Yu, Xiu and Li (2022) points to Transfer Learning as a better alternative for solving existing lacuna in traditional ML solutions, where the training and testing data come from the same dataset and share consistent feature distributions, without guaranteed consistency marred by problems like fewer annotations in comparatively larger datasets, poor computational capability of devices, and model generalization with limited data.

The paper (Weiss, Khoshgoftaar and Wang, 2016) further offers specific transfer learning solutions, categorized, depending on similarities of the input feature space. If input feature spaces of Source and Target domains are similar, the solutions are categorized under “Homogeneous transfer learning” and if dissimilar, they are categorized under “Heterogeneous transfer learning”; and further, if input feature spaces are not well-related, solutions would be categorized under “Negative transfer Learning”.

While both the papers (Yu, Xiu and Li, 2022; Weiss, Khoshgoftaar and Wang, 2016) concur with on the categorization of TL to four types, that of instance-based, model-based, feature-based and relational-based; the paper (Yu, Xiu and Li, 2022) further comprehensively reviews the recent development of Deep Transfer Learning and are especially relevant in the current context of large-scale amalgamation of DNN and TL in the space of Generative AI techniques.

The survey paper (Zhuang et al., 2020) further experimented and captured performances of various homogeneous transfer learning models conducting experiments with three different datasets, i.e., Amazon Reviews, Reuters-21578, and Office-31. The experimental results demonstrated in their paper illustrate the importance of selecting appropriate transfer learning models for different applications in practice and served as guide to the ML practitioners and Data Scientists.

Generative AI for Business and Technology

The report by Bommasani et al. (2022) investigates the novel and emerging paradigm of building AI systems based on a general class of models, generally termed as foundation models. The report captures the basic traits of the foundation models: as those that are based on deep neural networks, and adopting self-supervised learning mechanisms at scale, are trained on broad datasets with a possibility of further "tailoring", that is to be fine-tuned to perform a wide range of domain-specific downstream tasks.

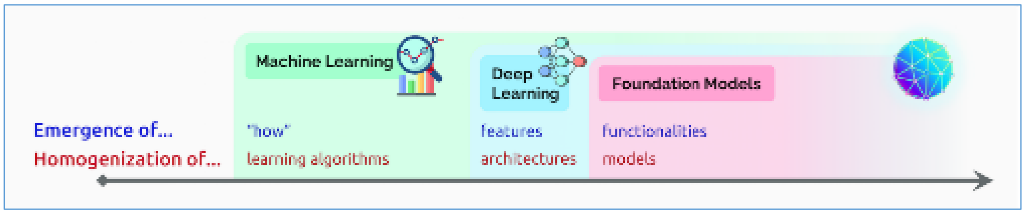


Figure 4.11
Homogenization of ML Models to Foundation Models. Courtesy: Bommasani et al. (2022)

The report further notes:

"The sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible; for example, GPT-3 has 175 billion parameters and can be adapted via natural language prompts to do a passable job on a wide range of tasks despite not being trained explicitly to do many of those tasks"

The foundation models have further been introduced by Feuerriegel et al. (2023) as "Generative AI", exhibiting "implicitly induced" emergent behavior, rather than the behavior having to be explicitly constructed, and has been defined as follows:

"The term generative AI (GenAI) refers to computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data."

From a "model-level" perspective, the author further clarifies Generative AI to be a ML architecture that uses AI algorithms to create novel data instances, drawing upon the patterns and relationships observed in the training data; a model though critically central yet requiring further fine-tuning to "domain-specific" use cases. The three-level views, that is Model, System and Application views of GenAI have been proposed in the paper.

The paper notes that GenAI is a natural consequent of evolution from statistics-based Natural Language Processing with “frequency-based embedding methods” like TF-IDF, generating text vectors by counting the frequency of frequently occurred words to NN-based language modeling with “word embeddings” in a sequence-to-sequence (seq2seq) processing of input tokens with RNN and LSTM, further to be infused with the concept of implementing the same with “Attention” mechanism.

Ultimately, parallelized processing of input sequence with the technological break-through with “Transformer with positional encoding” led to architectures only with Transformers, leading to large-scale implementation and adoption of Generative AI in various domains of business and technology.

While attention mechanism, determines the contextual word embeddings for words in a corpus in its various avatars like “self-attention” and “multi-headed attention” (Campesato, 2023), the implementation breakthrough of parallelized sequences was brough about by the proposition of Transformer, introduced by Vaswani et al. (2017; 2023).

The architecture of a Transformer is explained by Campesato (2023) in his book:

“In highly simplified terms, a transformer consists of an encoder and a decoder, each if which involves a stack of attention layers that perform attention-related operations".

Further, the concept of “positional encodings”, to feed information about the position of the token embeddings in the input sequence, has been captured in the updated paper by Vaswani et al. (2023) and also referred to in Campesato (2023).

Naveed et al. (2023) noted the recent idea of “Selective Attention” for pre-trained language models (PLMs) as below:

"While conventional language modeling (LM) trains task-specific Attention, particularly selective attention, has been widely studied under perception, psychophysics, and psychology."

While different transformer architecture models exist, namely "encoder only", "decoder only" and "encoder-decoder" based, each leading to the exploitation of the attention mechanism like "self-attention" and "cross-attention", the illustration of the "encoder-only" block architecture, exploiting the "self-attention" layer has been captured in the below figure, illustrated with an input statement (Kokab, Asghar and Naz, 2022):

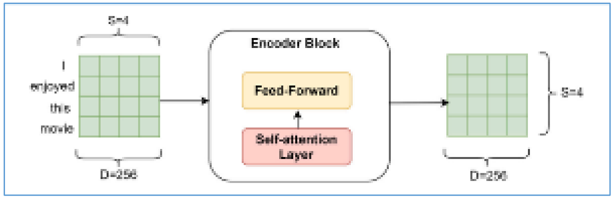


Figure 4.12
Schematic diagram of self-attention mechanism and its further exploitation for Sentiment Analysis. Courtesy: Kokab, Asghar and Naz (2022)

Transformer-based "Generative AI with varying "Large Language Models" (LLMs) is re-defining business and technology eco-systems with their ingenuine use-cases for adoption in different domains with their "ability to achieve general-purpose language understanding and generation" as per Wikipedia (Large language model, n.d).

Further, while evolving the Conversational AI tasks like "intent recognition, entity extraction, and dialogue management" by augmenting them with contextual understanding and intuitive "human-like" responses, the key concept of LLMs being "Generative" has been defined by their ability to be "pre-trained" with training data, created with their associated probability distributions, further allowing them to "generate" text and be used for other broader "generative" tasks (Campesato, 2023):

"…like story writing, code generation, poetry and creating content in specific style or mimicking certain authors, showcasing their generative capabilities."

A snap-shot view of the leading available generative AI tools, generating text, audio, images, videos, and 3D models, which can be further "taught" to perform domain-specific generation as mentioned above, pose an exciting new era of "training" of Artificial Intelligence to their domain-specific tasks, and are presented in the blog (Hiter, 2023), for "Top 9 Generative AI Applications and Tools", with information of their origin Company, summary capabilities as "Use-cases" they support and customer "on-boarding" prices, as of May, 2023.

Further, Jacob and Nair (2022) draw a compelling yet brief history of LLMs and their possible use cases in their blogpost at Exemplary and a similar yet more recent study by Toloka Team (2023) has been conducted in their blogpost on "The history, timeline and future of LLMs".

Finally, we summarily delve into the operations of the "Large-Language" Models to investigate the business relevance and impacts, basing on the work-flow as shown in the figure below (Naveed et al. 2023):

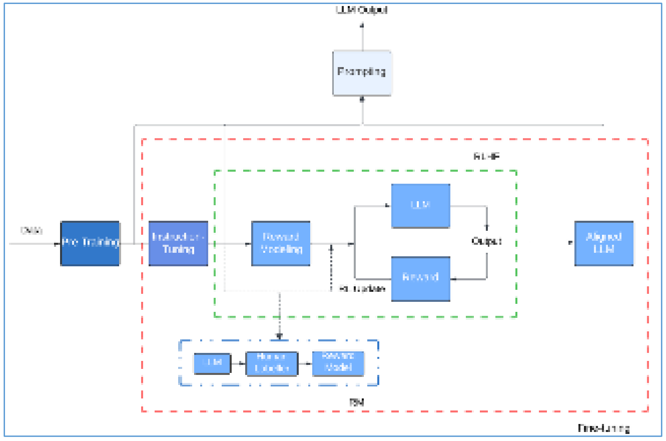


Figure 4.13
Operations of Large Language Models (LLMs). Courtesy: Naveed et al. (2023)

The above figure allows us to summarily capture the various processing modules of the LLM training work-flow.

Pre-training module, as explained in the papers (Naveed et al. 2023; Devlin et al., 2019; Humeau et al., 2020), is the initial stage where “model is trained in a self-supervised manner on a large corpus to predict the next tokens given the input”. The architectural and design choices of this module depend on the Natural Language (NL) use-case - encoder only for “NL Understanding” tasks, or decoder-only for “NL Generation” tasks; and encoder-decoder for sequence-to-sequence modeling. The paper by Chronopoulou, Baziotis, Potamianos (2019) explains the use of Transfer Learning to the Pre-train LLMS, which is briefly yet interestingly explained in the Spiceworks blogpost (Kanade, 2023).

From the enterprise business perspective, it is important to relate the domain of the business and which set of pre-trained LLM to further investigate and conduct “domain-specific training on; and incorporate in its day-to-day operations.

Further, as per the figure above (Naveed et al. 2023), further model adaptation approaches are adopted which would include Instruction-tuning and alignment-tuning.

To note, Instruction-tuning “fine-tunes” the pre-trained model in a “supervised” manner (method also known as SFT), to respond effectively to users; and is based on multi-task instruction data in plain natural language to guide the model to respond, according to the prompt and the input, improving “ask-only” or “zero-shot” generalization; Further, alignment-tuning is conducted with an aim to “Align model with human feedback”, ensuring sanitized response to the queries, fulfilling the “HHH” criteria of the model being helpful, honest, and harmless (Hao, 2023; @Masteringllm, 2023).

The paper classifies SFT mechanisms with manually created datasets as “Zero-Shot Prompting”, “Few-shot learning” and “Chain-of-Thought” (CoT) Prompting; the last case classified as a special case to enable reasoning in LLMs. The concept of CoT and its importance in “reasoning” has been further captured in the paper by Zhang et al. (2023).

Further, RLHF and RLAIF (Naveed et al. 2023**;** Hao, 2023; Lee et al., 2023) are employed for further “model alignment”, where models, already fine-tuned by human feedback, are further trained with reward model (RM), to enable classified ranking of responses that humans would prefer where the classifier is trained with humans annotating LLMs generated responses based on HHH criteria.

Ultimately, RL combines with the reward model for alignment; previously trained reward model ranking LLM-generated responses into preferred vs. unpreferred, which is used to align the model with the “iteratively repeating” proximal policy optimization (PPO) process for achieving convergence (Hua et al., 2023) and generalization and diversity with RLHF (Kirk et al., 2023).

Further, the paper by Hua et al. (2023) thoroughly investigates the RLHF framework, emphasizing on the exploration of “how the parts comprising PPO algorithms impact policy agent training” resulting in their identification of “policy constraints” as the key factor for the effective implementation of the PPO algorithm.

To note, as an efficient alternative to general fine-tuning method, “Parameter-Efficient Tuning”, has also been suggested, which “aims to achieve performance comparable to fine-tuning, using fewer trainable parameters” for LLM updates, preserving computing and memory resources with several strategies having been proposed (Naveed et al. 2023; Chen et al. 2023). Proposed techniques mentioned in the papers include “Prompt Tuning”, “Prefix Tuning” and “Adapter Tuning” that are employed to achieve specific requirements.

However, the paper by Fu et al. (2022) notes that “as the parameter number grows exponentially to billions, it becomes very inefficient to save the fully fine-tuned parameters", going on to propose a "novel Second-order Approximation Method (SAM) to approximate the NP-hard optimization target function with an analytically solvable function".

Lately, alternatives to RLHF have been researched (Dong et al., 2023) to respond to its "inherent limitations stemming from a complex training setup and its tendency to align the model with implicit values that end users cannot control at run-time". Also, the reward models in RLHF stage "commonly rely on single-dimensional feedback as opposed to explicit, multifaceted signals that indicate attributes such as helpfulness, humor, and toxicity". The paper addresses these limitations of RLHF by proposing "STEERLM", a supervised finetuning (SFT) method that empowers end-users to control and condition responses during inference, to conform "to an explicitly defined multi-dimensional set of attributes, thereby empowering a steerable AI capable of generating helpful and high-quality responses while maintaining customizability". The responses of STEERLM have been found to be "preferred by human" and "automatic evaluators" compared to "many state-of-the-art baselines trained with RLHF while being much easier to train".

Further, the paper (Yao et al., 2023) explores amalgamation of generation of reasoning traces and task-specific actions, through a mechanism appropriately named "ReAct" for interleaved operation of "reasoning" and "acting" (based on those reasoning), and note:

"…reasoning traces help the model induce, track, and update action plans as well as handle exceptions, while actions allow it to interface with and gather additional information from external sources such as knowledge bases or environments."

From the perspectives of image generation tasks, Generative Adversarial Networks (GANs) have proved themselves to be superior in most metrics, however, the aspect of capturing “diversity” is less performant, and generally are more difficult to train, making them less scalable, hindering application to new domains. Lately, Diffusion models, a class of likelihood-based models that generate samples by gradually removing noise from a signal with their training objective expressed as “a reweighted variational lower-bound”, have been shown to out-perform GANs (Dhariwal and Nichol, 2021), producing high-quality images while offering desirable properties such as distribution coverage, a stationary training objective, and easy scalability. These models are currently preferred to GAN in many image generation use-cases.

Enhancements to Diffusion Model operation have been suggested in the paper (Wallace et al., 2023), where it proposes Direct Preference Optimization (DPO), a method to align diffusion models to human preferences by directly optimizing on human comparison data, and posed as a simpler alternative to RLHF which directly optimizes a policy that best satisfies human preferences under a classification objective.

Different Multimodal LLMs (MLLMs) inspired by the success of LLMs, and offering “substantial benefits compared to standard LLMs by incorporating information from various modalities”, achieving a deeper understanding of context, leading to more intelligent responses infused with a variety of expressions have been suggested (Naveed et al. 2023; Yin et al., 2023). The importance of MLLMs as a step forward, towards Artificial General Intelligence (AGI), may be summarized as below (Yin et al., 2023):

“Importantly, MLLMs align closely with human perceptual experiences, leveraging the synergistic nature of our multisensory inputs to form a comprehensive understanding of the world. Coupled with a user-friendly interface, MLLMs can offer intuitive, flexible, and adaptable interactions, allowing users to engage with intelligent assistants through a spectrum of input methods.”

In a recent development (Wu et al., 2023) “any-to-any” Multi-modal-LLM (MM-LLM), touted as “NExT-GPT”, capable of accepting and delivering content in any modality, catering to the needs of “human-level AI” has been reported “as a breakthrough”.

The paper by Lewis et al. (2021) and the blogpost by Medium (Upadhyay, 2023) covered respectively, the theoretical and practical aspects of “Augmented LLMs” and Retrieval Augmented Generation, as an emergent capability to learn from “context augmentation” from “few-shot prompting” without needing the costly process of fine tuning. Their strength, as explained in the paper (Lewis et al., 2021) and the implementation demonstrated with Langchain and Hugging Face in the blogpost (Upadhyay, 2023), promising excellent generalization to unseen tasks with few-shot prompting, enabling LLMs to answer queries beyond the capacity acquired during training and also allowing the LLM-generated responses to avoid hallucination, inaccuracy and factual incorrectness. The paper (Lewis et al., 2021) further, in great depth, explains “Retrieval Augmented LLMs” generally known as RAGs, that allow LLMs to remain contextual and current based on updated external storage information.

Further, fine-tuning of LLMs in Amazon SageMaker JumpStart on Financial data, as a demonstration of adaptation to financial domain has been conducted showcasing the business adoption of Generative AI in the Amazon blogpost (Huang et al., 2023).

Finally, with the plethora of LLMs becoming available for use and experimentation, benchmarking their performance is currently in focus, as noted in the blogpost by Deepgram (Rowley, 2023):

“As language models have become more powerful, there's an increasing focus on benchmarks that measure both performance and ethical aspects like fairness and bias. There's also interest in explainability, or how well a model can provide understandable reasons for its outputs. The landscape of benchmarks covering these emerging areas of inquiry is rapidly evolving.”

Further the blogpost captures various popular Benchmark Tools like ‘API-Bank’, capturing the tool-usage; ‘ARC’, evaluating LLM’s reasoning abilities; and ‘HellaSwag’, an LLM benchmark for commonsense and reasoning; and ‘HumanEval’, a tool to benchmark code generation.

Finally, as noted by Naveed et al. (2023), MMLU, a benchmark measuring “knowledge acquired by models during pretraining and evaluates models in zero-shot and few-shot settings across 57 subjects, testing both world knowledge and problem-solving ability” has superseded GLUE and SuperGLUE. To note SPPERGLUE “includes a variety of language understanding tasks, such as question answering, natural language inference, and coreference”.

Edge AI Chips and Accelerators and their Impact to Technology & Business

> As per a recent market research (Fortunebusinessinsights.com, 2022), the edge AI market size has been valued at USD 11.98 billion in 2021 and was expected to reach USD 107.47 billion by 2029, exhibiting a CAGR of 31.7% during the forecast period as against global artificial intelligence market that was valued at $428.00 billion in 2022 & is projected to grow from $515.31 billion in 2023 to $2,025.12 billion by 2030.

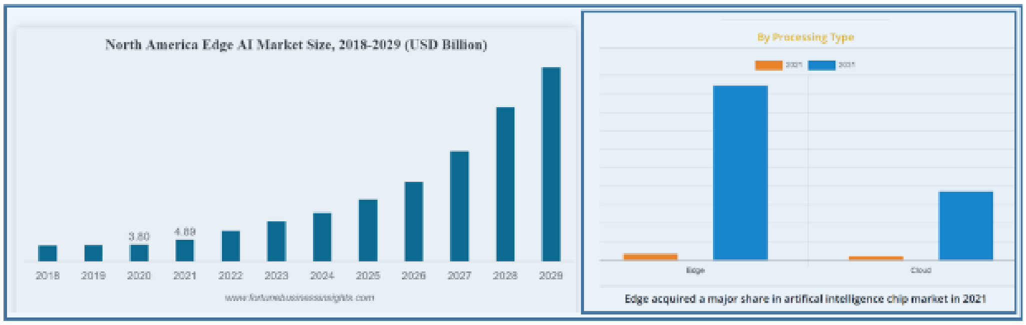
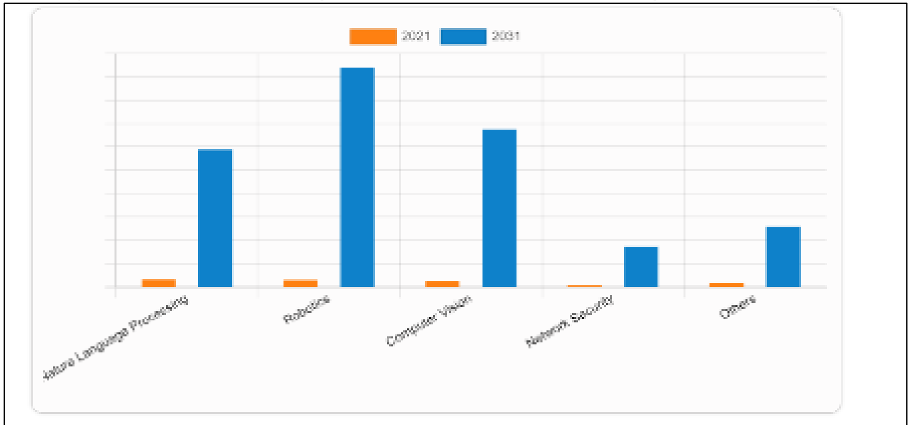


Figure 4.14

AI Chip and Market Size and Projections. Courtesy: Fortunebusinessinsights.com (2022)

Notwithstanding the difference in base years, Edge AI did acquire a major share in the AI chip market and the growth is expected to continue from a long-term perspective and it is very likely that the market pull of the edge AI chips, dictated by edge IOT, Robotics and Autonomous driving, are expected to act as a significant driver for the overall growth of the chip industry.

These edge AI chips thus can enable companies to implement next-generation IoT applications with their support as per the prediction by Deloitte (Stewart et al., 2020). Smart machines powered by AI chips are likely to assist in market creation and expansion for entrepreneurs and incumbents alike and invigorate the formation of a new market eco-system, with revenues and profits sharing re-aligned across the industries such as manufacturing, construction, logistics, agriculture, and energy The edge chips, allowing "in-place" vast amount of data collection, interpretation and the ability to immediately respond is critical for many of the data-heavy applications that the next-generation businesses and technologies will heavily rely on, like video monitoring, virtual reality, autonomous drones and vehicles, and more. With the advent of the edge AI chips, leads us to be enthusiastic as the "future seems to be now", and this is further corroborated by the market study for the Robotics growth that will drive the growth of edge AI Chips.



Robotics is expected to emerge as a significnat segment during 2022-2031

Figure 4.15
Robotics Market Projections. Courtesy: Deloitte - Stewart et al. (2020)

From an academic and technology perspective, an abridged "lecture note" on influence of Accelerators for Machine Learning with AI (Cornell University, 2019) observes:

"…an exciting new generation of computer processors is being developed to accelerate machine learning calculations. These so-called machine learning accelerators (also called AI accelerators) have the potential to greatly increase the efficiency of ML tasks (usually deep neural network tasks), for both training and inference. Beyond this, even the traditional-style CPU/GPU architectures are being modified to better support ML and AI applications. Today, we'll talk about some of these trends."

In accordance with the above, a similar context in DNN development for Accelerator has been captured in the paper (Chen et al., 2020), where they review AIML chips and accelerators as ''domain-specific computing" platforms for edge, needing specific customization for AI and also provide a vision of the roadmap of edge AI chips.

While summarizing the recent advances in accelerator designs for deep neural networks accelerators and discussing various architectures that support DNN executions in terms of computing units, dataflow optimization, targeted network topologies, architectures on emerging technologies, and accelerators for emerging applications, it also captures the journey of Google's TPU architecture towards supporting edge AI.

The paper by Google Inc. (Seshadri et al., 2022) captures the latest development and charts the design roadmap of TPU1, which, with a systolic array, focused on inference tasks and were deployed in Google's datacenters in 2015; followed by TPU2, upgraded to handle both training and inference in the datacenters. TPU2, while adopting a systolic array, introduced vector-processing units. In 2018, TPU3 was introduced with "liquid cooling" feature and further, edge TPU availability was announced targeting the inference tasks of the Internet of Things (IoT). This study thus allowed us to capture evidences for the vibrant AI Accelerator field and its influence in the Machine Learning and Inference domains.

The paper by Shlezinger et al. (2022) presents leading approaches for studying and designing “model-based” deep learning systems which comprises of methods to combine principled mathematical models with data-driven systems, benefit from the advantages of amalgamated approach allowing them to exploit both partial domain knowledge, via mathematical structures designed for specific problems, as well as learning from limited data. This can be seen as a precursor to the template-based design, Google uses to design its edge TPU, as noted in the paper:

“The Edge TPU accelerators leverage a template-based design with highly parameterizable microarchitectural components. The parameterized design of Edge TPU accelerators enables exploring various architecture configurations for different target applications.”

The above papers thus act as corroborative evidence on the exciting future trend of AI accelerator and chip design markets as the industry’s focusses on steady efforts to move to Accelerator-based specialized ML Training and Inference chips.

The latest growth of Generative AI has further fueled the growth of AI Chips aimed at training and inference of LLM applications in the data-centers as is evidenced by the Chip support for Generative AI by the big-tech market leaders (MSV, 2023):

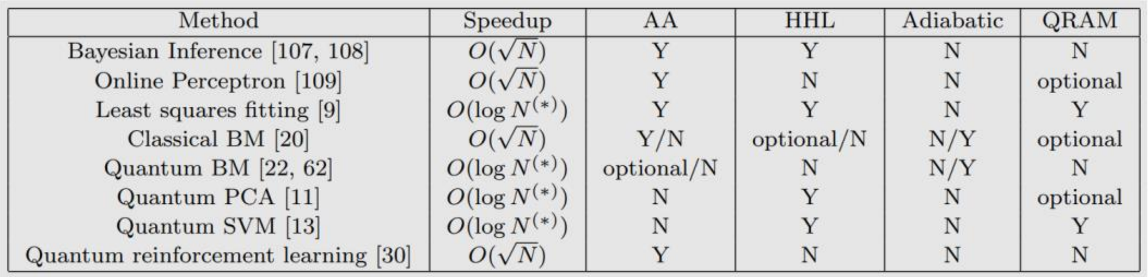
The latest announcement of Microsoft’s entry with Maia Chip (Hetzner, 2023) has further opened up competition in the niche data-center chip market, focusing on Generative AI model training.

Quantum Machine Learning

The survey paper on quantum algorithms (Dalzell et al., 2023) captures the history of QML, with Deutsch proposing his first quantum algorithm in 1985, followed through the decade, with larger efficiency compared to classical algorithms by Deutsch–Jozsa, Bernstein–Vazirani, and Simon's algorithms. They capture the break-through, heralding the advent of "first truly end-to-end quantum algorithm" Shor's algorithm (Shor, 1996), for factoring integers and computing discrete logarithms deeply affecting the future of quantum computing (QC) for real-world applications, specifically cryptography development. Since Shor's seminal discovery, various practical QC-based solutions for simulation and linear algebra have been created, optimized and generalized and the paper by Biamonte et al. (2018) captures the "Quantum speedup", achieved by the Quantum algorithms, noting:

"Quantum computers use effects such as quantum coherence and entanglement to process information in ways that classical computers cannot. The past two decades have seen steady advances in constructing more powerful quantum computers. A quantum algorithm is a step-wise procedure performed on a quantum computer to solve a problem, such as searching a database. Quantum machine learning software makes use of quantum algorithms to process information."

The paper further tabulates the speedup as shown in the figure below:



Method	Speedup	AA	HHL	Adiabatic	QRAM
Bayesian Inference [107, 108]	$O(\sqrt{N})$	Y	Y	N	N
Online Perceptron [109]	$O(\sqrt{N})$	Y	N	N	optional
Least squares fitting [9]	$O(\log N^{(*)})$	Y	Y	N	Y
Classical BM [20]	$O(\sqrt{N})$	Y/N	optional/N	N/Y	optional
Quantum BM [22, 62]	$O(\log N^{(*)})$	optional/N	N	N/Y	N
Quantum PCA [11]	$O(\log N^{(*)})$	N	Y	N	optional
Quantum SVM [13]	$O(\log N^{(*)})$	N	Y	N	Y
Quantum reinforcement learning [30]	$O(\sqrt{N})$	Y	N	N	N

Table 4.2
Quantum Speedup achieved by Quantum Algorithms. Biamonte et al. (2018)

In his paper, Osvaldo Simeone (2022) captures the nuances of quantum-gate based machine learning algorithms, a NISQ-based approach that is followed by most companies like IBM and Google to develop quantum processors. Contrarily, Adiabatic quantum computation and annealer-based processor developments have been pioneered by D-Wave (‘Adiabatic Quantum Computation’, 2019; Dargan, 2023).

However, lately Quantum computer maker D-Wave has embraced the quantum-gate approach to develop hybrid quantum hardware as reported in the CNET blogpost (Shankland, 2021) probably due to the industry’s significant weight and push behind the NISQ-based hardware development (Perdomo-Ortiz et al., 2018).

The open-access Springer e-book by Hughes et al. (2021) note:

“In 2019, Google claimed to have performed the first quantum computation that a classical computer could not do—a milestone known as “quantum supremacy”. Quantum supremacy means that a quantum computer can solve a problem that a classical computer cannot. However, the solution of the problem may not be of practical use. As such, it is important to note that Google has demonstrated quantum supremacy, not the “quantum usefulness” milestone. Google performed their task on a 53- qubit quantum computer, which took 200s.”

However, though IBM suggested an improved classical supercomputing technique could theoretically perform the same task in 2.5 days, that did not over-throw the claim of Goggle, thus heralding the era of Quantum Machine Learning as an active area of globally acknowledged practical research that could impact technology and businesses.

Glisic and Lorenzo (2022) further emphasize on the physics of quantum computing stating that “entanglement and superposition of the basic qubit states provides an edge over classical ML” and goes on to discuss in detail the various QML algorithms. The authors further note some important ML algorithms, already implemented in the Quantum realm using NISQ-based processors:

- Quantum Clustering technique, using the “quantum Lloyd’s algorithm to solve the k-means clustering problems” and speeding up “the process in comparison to the classical algorithm”;
- Quantum Decision Tree, employing “quantum states to create the classifiers in ML.”
- The quantum HHL algorithm, “for solving linear systems of equations.”
- Quantum Support Vector Machines for solving classification problems, created using qubits.

From the perspective of application of QML for implementing Neural Networks, the paper by Tacchino et al. (2018) noted:

“In practical applications, artificial neural networks are mostly run as classical algorithms on conventional computers, but considerable interest has also been devoted to physical neural networks, i.e. neural networks implemented on dedicated hardware.”

The authors further propose a model for implementing “perceptrons” on a NISQ device and note:

“…we have experimentally tested it on a 5-qubits IBM quantum computer based on superconducting technology. Our algorithm presents an exponential advantage over classical perceptron models, as we have explicitly shown by representing and classifying 4 bits strings using 2 qubits, and 16 bits strings using only 4 qubits.”

However, they note that the possibility of “exponential advantage” may not be addressed in the short-term with the NISQ-based processors:

“In principle, generic quantum states or unitary transformations require an exponentially large number of elementary gates to be implemented, and this could somehow hinder the effective advantages brought by quantum computers for machine learning applications. This currently represents a general problem of most quantum machine learning algorithms.”

Cong, Choi and Lukin (2019) introduced and analyzed a novel quantum machine learning model motivated by CNNs, using only O(log(N)) variational parameters for input sizes of N qubits, allowing for its efficient training and implementation on realistic, near-term quantum devices. The QCNN architecture combined the multi-scale entanglement renormalization ansatz and quantum error correction.

Given the direction of novel QML-based NN algorithm implementation on NISQ,

the exciting promises QML still needs implementations to fulfill the “exponential large number of elementary gates”, which is limited in NISQ processors by their limited size and “coherence time”, as has been noted in the blogsite of The Quantum Insider (Jakob, 2023),

“…Superconducting qubits – which use superconducting circuits to maintain and manipulate quantum information – are considered one of the most viable and scalable quantum computing modalities. However, short T1/relaxation and T2/ coherence, and lifetimes in superconducting qubits challenge the modality’s use in practical quantum computing applications”.

They further report an exciting breakthrough in recent material science research by Fermilab engineers of Yale Quantum Institute, as noted in their paper (Read et al., 2023), suggesting that the adoption of HEMEX sapphire as substrate instead of silicon, extends the relaxation (T1) and coherence (T2) times, will pave the path to a bright future in Quantum Computing research.

While exciting research developments to enhance the performance of the Quantum processors continue, in another front, QML is also being tested to enhance the capabilities of Generative AI (Gao, Zhang and Duan, 2018).

Finally, to note, various software frameworks currently exist for research and experimentation and the blog by Medium (Editorial@TRN, 2020) notes the top few as Cirq by Google's Quantum AI Team, Quantum Development Kit & Q# Programming Language by Microsoft and QisKit open-source quantum software development framework by IBM.

The paper (Broughton et al., 2020) specifically introduces TensorFlow Quantum (TFQ), an open-source library framework for the rapid prototyping of hybrid quantum-classical models for classical or quantum data, offering high-level abstractions for the design and training of both discriminative and generative quantum models supporting high-performance quantum circuit simulators.

To note, the paper by Kiwit et al. (2023) have also announced enhancements to **QU**antum computing **A**pplication benchma**RK** (QUARK) framework that already simplifies and standardizes benchmarking studies for quantum computing applications.

From a business application perspective, QML frameworks and novel hybrid classical–quantum neural networks to improve binary classification models for noisy datasets have been experimented on financial datasets (Schetakis et al., 2022). The metric for assessing the performance of the quantum classifiers is the generally accepted ML metric of AUC–ROC. An extensive benchmarking of the novel FULL HYBRID classifiers has shown to exhibit better learning characteristics to asymmetrical Gaussian noise in the dataset against existing quantum classifier models while and performing equally well for existing classical classifiers, with a slight improvement over classical results in the region of the high noise.

Further, the rapidly developing field of QML in business context has been emphasized by DwaveSystems (www.dwavesys.com, 2023; www.dwavesys.com, n.d.) who have recently reported:

"The Advantage™ system is the first and only quantum computer designed for business. Our 5 generation Advantage quantum computer was built from the ground up with a new processor architecture with over 5,000 qubits and 15-way qubit connectivity, empowering enterprises to solve their largest and most complex business problems."

QML thus continues its long and exciting performance enhancement journey towards it providing promised computational efficiencies, with rapid development, in the decades to come.

Cloud-Native Platforms and Frameworks for enabling Next Generation AIML

The march towards intelligent platforms is exemplified by the rapid transitioning and superseding of the early generation open European AIML Platform, represented by the "DEEP-Hybrid-Data Cloud" (DHDC) project till early 2023 (DHDC, 2023) with AI4EOSC platform and the AI4OS stack. The infrastructure layer, as shown in the below figure is expected to be primarily supported by the Intelligent Kubernetes Orchestrator engine, while support for other big-data orchestration platforms, like Apache Mesos also exist.

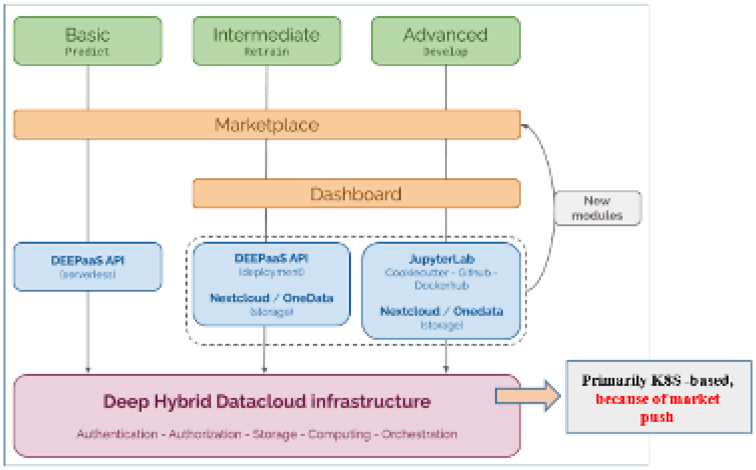


Figure 4.16
Architecture overview — DEEP-Hybrid-Data Cloud. Courtesy: DEEP-2 documentation. Deep Hybrid Data Cloud [online] Available at: https://docs.deep-hybrid-datacloud.eu/en/latest/user/overview/architecture.html (Accessed 5 Jan. 2024).

As documented in the paper, the supported platforms range from user workstations, computing servers all the way to HPC systems, and Container Orchestration Engines (such as Kubernetes or Mesos) or serverless frameworks such as OpenWhisk.

In Private data centers and private cloud settings enabling "cloud-native microservices-based" business and AIML applications mainly rely on CNCF-promoted (as noted in Wikipedia ('CNCF', 2023) open-source "container orchestration platform" Kubernetes, as evidenced by the noted technology blogger Jankiraman (MSV, 2023) in his blogpost:

"Red Hat, VMware, Canonical, Mirantis, Rancher and other vendors offer Kubernetes-based platforms that can run in both enterprise data centers and the public cloud. The rise of Kubernetes forced hyperscale cloud vendors such as Alibaba, AWS, IBM, Google, Huawei, Microsoft and Oracle to offer managed Kubernetes services."

The existence of Kubernetes-based open-source platforms allows exploitation of the same by Entrepreneurial Organizations to setup AIML-based services based on open-source software.

Kubernetes and CNCF thus offer a level playing field for entrepreneurs, intrapreneurs and cloud companies to develop value-based service models and coopete based on their offerings.

CNCF, as noted in Wikipedia ('CNCF', 2020), currently includes a plethora of working Projects, Kubernetes ('Kubernetes', 2020), being the most well-known, promoting other key components of the cloud native ecosystem: Prometheus, Envoy, Helm, Fluentd, gRPC, and many more.

Kubernetes orchestrates the containerized applications, running in pods, takes advantage of the distributed and decentralized nature of the cloud (Poulton and Joglekar, 2020), thus allowing the cloud native applications (latest generation cloud software development model) to be composed of multiple, cooperating, distributed microservices.

The domination of Kubernetes as a platform for enabling AIML services, as CNCF provides the evidences of contribution to K8S development by the Cloud/AIML Technology leaders, clearly exemplifying its adoption, as presented in "Kubernetes Project Journey Report*"* by cncf.io (2023).

Given the above evidences of industry adoption, we are led to believe that Kubernetes-based platform eco-system promises to dominate the "cloud-native" based AIML/SaaS application software development, as a de-facto standard, thus providing the economy of scale to the industries at large.

The book by Poulton and Joglekar (2020) and the Journal paper (Ogbuachi et al., 2020) further explore and capture the key architectural and functional attributes of Kubernetes Platform reflecting the intelligent orchestration, scheduling and management of the pods (the smallest deployable units in the nodes) by the Controller Manager, Scheduler and the database in Kubernetes Master node and its application in edge application in the context of 5G.

In the Public cloud setting, Kubernetes services are provided as EKS, AKS and GCE for AWS, Azure and GCP services respectively (offered by Amazon, Microsoft and Google). The focused comparisons on the services offered are presented in a reputed technology blog (Sam-Solution, 2022).

To note, very recently K8sGPT (Jones, A., (202x), an automated analysis tool based of Generative AI has been inducted in CNCF, with codified SRE experience, to serve for Kubernetes workload health analysis aiding fast automated scanning of Kubernetes clusters, diagnosing and sorting and allocating issues. These automations serve further to establish Kubernetes as a “go to” infrastructure platform to host for modern container-based applications for businesses.

The Kubernetes platform is also a boon to any entrepreneurial endeavor for AIML development and deployment in a startup organization, the intermediation of various Service provides to deploy K8S services do exist, due its complexity of continuous maintenance. The various blogposts (Sherrer, 2021; Modi, 2023) capture provide various alternative routes to provide similar service, albeit at an addition cost.

To further emphasize, given the wide-spread adoption of Kubernetes in platforming AIML container-based services, we further explore Kubeflow. As explained in a blog by CNCF (Kubeflow, 2023) aims to provide a Kubernetes-native MLOps project for deploying and managing a Machine Learning (ML) stack on Kubernetes, as noted below:

“The Kubeflow community actively develops and supports Kubernetes-native MLOps for its users who develop and deploy distributed machine learning in popular frameworks, including TensorFlow, PyTorch, XGBoost, Apache MXNet, and more.”

Automated model deployment in production environment

Creating multi-tenant environment for effective sharing of ML environments and resources

Allow data scientists run Jupyter Notebooks in combination with python code in docker containers providing secure workflow environments on GPUs, that provide powerful computational power to execute complex ML algorithms.

Further, the blog (run.ai, 20xx) compactly captures the main architectural components and elements of Kubeflow, emphasizing on the Intelligent Platform service-aspects, with the Kubernetes resources and controllers, intelligently and autonomously orchestrating the pipeline services, aided by intelligent persistence agent and ML metadata.

Finally, supporting evidence from “The AI Index Report” from Stanford University (HAI, Pg. 195, 2023) corroborates our emphasis and analysis of K8S/Kubeflow as an enabler for transforming the cloud industry’s data management and processing with the following table capturing the fact that maximum private investment in focus area in 2021/2022 were led by Data Management and processing in the Cloud, enabled by “de-facto” k8S based-AIML platforms.

Given the wide-spread adoption of Kubernetes/Kubelow for AIML platform and MLOps, as already explored, we may be confident that they will remain dominant technologies enabling large-scale industry adoption of AIML in the years to come.

It is thus imperative from the above analysis of market direction and industry adoption that AIML forms the secret sauce Digital Transformation with Distributed Cloud.

Deep Learning Tools and Frameworks

As explained in the online Deep Learning eBook (Zhang et al., 2022), the first generation of open-source frameworks for neural network modeling consisted of ‘Caffe’, ‘Torch’, and ‘Theano’; and ultimately, they were superseded by Tools like ‘TensorFlow’, ‘PyTorch’ and ‘TensorRT’ (Hadjer, 2020).

Further the paper (Pouyanfar, Sadiq and Yan, 2018) studies the available traditional and the latest frameworks, comparing various support, notably for that of Deep Neural Networks.

Hadjer (2020) further presents a view of TensorFlow as an open-source neural network modeling tool for creating ML applications, consisting of several “Python and C APIs that have been expanded since its release in 2015”, by Google Brain Team. Aligned with the reputation of being “one of the most popular frameworks”, TensorFlow justifies that by having functionality allowing the user end-to-end control of the entire machine learning chain, from defining a network to running inference. An updated version released in 2019, TensorFlow 2.0, using Keras APIs, are currently in use as noted by Google Machine Learning Team (Abadi et al., 2016).

Another notable tool presented is ‘TensorRT’, an SDK build by NVIDIA on “CUDA” parallel programming model, including an inference optimizer. Notably, trained network from a supported platform, for example TensorFlow, can be run by the TensorRT optimizer to increase performance.

In the technology blogpost of Medium, Galarnyk (2021) presents ‘PyTorch’, a popular for constructing DNNs because of its ability to optimize mathematical expressions for computations on multidimensional arrays utilizing graphics processing units (GPUs); and the latest applications of DNNs are largely applied in the context of distributed Computing for all facets including training networks, tuning hyperparameters, serving models and processing data.

Galarnyk also points to the popular and open-source Framework Ray, with source library support for parallel and distributed Python, allowing to rapidly ML applications when paired with PyTorch. Notably, Ray eco-system consisting of three parts: the core Ray system, scalable libraries for machine learning (both native and third party), and tools for launching clusters on any cluster or cloud provider.

Entrepreneurial businesses, adopting Deep Neural Networks to develop their distributed computing-based AIML applications, can adopt any of the above framework and tools.

Analytics Platforms for Batch and Stream Processing

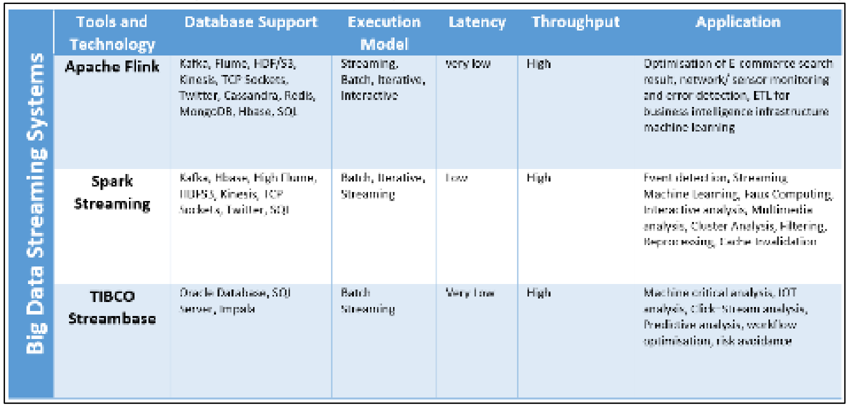
Stream processing architectures have been increasingly adopted by the industry due to deficiencies in traditional batch processing, as noted in the survey paper (Kolajo, Daramola and Adebiyi, 2019):

"Big data batch processing is not sufficient when it comes to analyzing real-time application scenarios. Most of the data generated in a real-time data stream need real-time data analysis. In addition, the output must be generated with low-latency and any incoming data must be reflected in the newly generated output within seconds."

Further the paper notes:

"The demand for stream processing is increasing. The reason being not only that huge volume of data need to be processed but that data must be speedily processed so that organizations or businesses can react to changing conditions in real-time."

The survey paper further goes on to compare a few of the Streaming Platforms are shown below, based on large-scale industry adoption and first-class streaming support.



	Tools and Technology	Database Support	Execution Model	Latency	Throughput	Application
Big Data Streaming Systems	Apache Flink	Kafka, Flume, HDF/S3, Kinesis, TCP Sockets, Twitter, Cassandra, Redis, MongoDB, Hbase, SQL	Streaming, Batch, Iterative, Interactive	very low	High	Optimisation of E commerce search result, network/ sensor monitoring and error detection, ETL for business intelligence infrastructure machine learning
	Spark Streaming	Kafka, Hbase, High Flume, HDFS3, Kinesis, TCP Sockets, Twitter, SQL	Batch, Iterative, Streaming	Low	High	Event detection, Streaming Machine Learning, Faux Computing, Interactive analysis, Multimedia analysis, Cluster Analysis, Filtering, Reprocessing, Cache Invalidation
	TIBCO Streambase	Oracle Database, SQL Server, Impala	Batch Streaming	Very Low	High	Machine critical analysis, IOT analysis, Click-Stream analysis, Predictive analysis, workflow optimisation, risk avoidance

Table 4.3
Comparison of Apache Beam, Flink and Spark. Courtesy: Kolajo, Daramola and Adebiyi (2019)

The journey of adoption of real-time processing in platform solutions was kick-started by the enterprise architects with Lambda "hybrid" architecture, having separate batch and real-time layers. Lambda architecture has been superseded recently by the

Kappa architecture, having a single real-time pipeline (Waehner, 2021). The architectural comparison, has been captured and presented in the blog site by (Owczarek, 2023).

Novel ML Operation and automation Frameworks

The general connection between the manual ML pipeline and the vision for an expected automated pipe-line process between MLOps and AutoML (Automated Machine Learning) has been identified and ways of combining them has been laid out in the paper by Symeonidis et al. (2022), and the following figure lays out the difference.

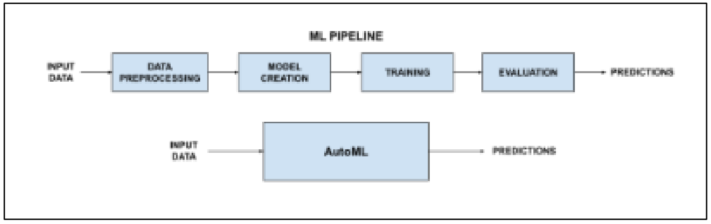


Figure 4.17
Manual ML pipeline vs AutoML. Courtesy: Symeonidis et al. (2022)

The general history of evolution of AutoML solutions in the past decade, is noted in the blogpost by Pascual in Medium.com (Pascual, 2021) illustrating open-source ‘AutoML’ solutions like ‘AutoWeka’ and ‘Auto-sklearn’ pioneered the space closely followed by TPOT, triggering a new wave of AutoML solutions including ‘Auto-ml’, and ‘Auto-Keras’. Parallelly, startups like H2O.ai and DataRobot released their versions of automated solutions augmented by solutions from big-tech companies who joined the solution marketplace with their offerings like “Amazon Sagemaker”, “Google AutoML, and “DriverlessAI”.

The following figure (Truong et al., 2019), high-light the ML activities the various stages of the AutoML process generally targets, through its “process pipeline”.

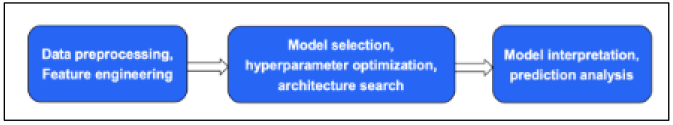


Figure 4.18
General performance measures of machine learning algorithms. Courtesy: Troung et al. (2019)

As noted in the Springer Open-Access text (Hutter, Kotthoff and Vanschoren, 2019), AutoML has allowed access to automated, hence efficient and fast, ML approaches to the researchers and business-domain engineers to compare pros and cons of the solution space, for finding the best-fit approach based on comparative statistical results, allowing easy accessibility, thus paving the path to democratization of machine learning.

The MLOps Survey paper (Zhengxin et al., 2023) attempts to reconcile both "existing literature and industrial best practices by suggesting an eight-step process, each consisting of a set of activities/tasks.

A perspective on AutoML taxonomy is captured in an orthogonal research and study requirements (Chen, Song and Hu, 2019), as shown in the figure below.

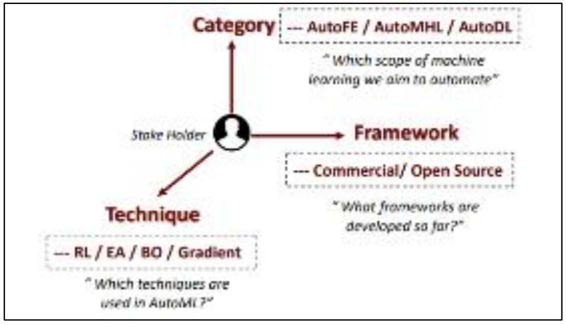

Category
--- AutoFE / AutoMHL / AutoDL
" Which scope of machine learning we aim to automate"
Stake Holder
Framework
--- Commercial/ Open Source
" What frameworks are developed so far?"
Technique
--- RL / EA / BO / Gradient
" Which techniques are used in AutoML?"

Figure 4.19
AutoML Taxonomy: Courtesy: Chen, Song and Hu (2019)

In the first axis of the above figure, AutoML is categorized under automated feature engineering (AutoFE), automated model and hyperparameter tuning (AutoMHT), and automated deep learning (AutoDL); the second axis, reviews AutoML techniques, studied under Bayesian optimization (BO), reinforcement learning (RL), evolutionary algorithm (EA), and gradient approaches (Gradient); in the third axis, AutoML frameworks and their provisioners in commercial services and open-source communities are reviewed. The recommended review method thus resents a unique way to review and research the AutoML categories, techniques and frameworks.

Further Novel Machine Learning Operation Frameworks

From a technology selection perspective, the comparison of various MLOps Platforms, made available by ThoughtWorks Inc. (MLOps Platforms, 202x) is relevant and with their key attributes captured, can aid technology and business leader to make an informed decision on the choice of Platforms in a planning stage of business transformation based in operational synergies of Intelligent Platform and AIML deployment requirements. They present their finding (available through Github), comparing a few well-known MLOPs Platforms from the perspective of business-related choices and analysis; which expectedly would lead to further elaborate engineering and operational analysis based on ML Operations.

In the context of the latest explosion of Generative AI technologies, the operational aspects of large-language models (LLMs), termed as “LLMOps” has become relevant, as noted in the article blogpost by Wandb.ai (202x):

“The steps involved in LLMOps are in some ways similar to MLOps. However, the steps of building an LLM-powered application differ due to the emergence of foundation models. Instead of training LLMs from scratch, the focus lies on adapting pre-trained LLMs to downstream tasks.”

The blog notes that the process of LLMOps differs from MLOps, focusing on adapting pre-trained models to downstream tasks, their evaluation, deployment and monitoring.

The blogpost by Caylent (Arabi, 2023) captures a reference architecture for operations on large-language models on AWS, providing a reliable snap-shot view in the journey through the fast-moving LLMOps landscape.

Decentralized Data Storage Technologies for Generative AI

For data storage in analytics and AIML applications purposes, in the age of Generative AI, a recent introduction in the context of storage. are Vector databases (Hinkle, 2023; Ghosh, 2023). The blogs note Vector databases as a key enabler for generative AI which are optimized for storing and retrieving embeddings, the vector representations of data. These databases form the basis of ultra-fast storing and retrieving data for generative AI models, excelling at fast nearest neighbor search across billions of embeddings for image search, recommendation and anomaly detection.

Novel Visual analytics Tools

While traditional leading commercial tools are presented in the blog by SelectHub (Adair, 2024) clearly showing the impact of “big data”, concentrating on the novelty perspectives, we draw attention to the Generative AI Visualization Tool RATH (kanaries.net., 2023). The tool is advertised to automate exploratory data analysis workflow with an “augmented generative AI-laced analytic engine”, assisting in efficient and fast discovery of patterns, insights and causals and presenting those captured insights with auto-generated multi-dimensional data. It boasts of various data exploratory automated tools for data wrangling, painter, charter and causal analysis and urges users to use RATH as “AI copilot” in data analysis.

Various alternatives of RATH have been made available (Sourceforge, 2024; g2.com, 2024), and they may serve as a “go to” source for decision making on choices of the best-fit visualization tool targeting the specific business requirement.

4.2 Meta-analysis of Strategic Business Frameworks for AIML

From our previous explorations and analysis of complexity of AIML technology, it can somewhat be ascertained that the strength of their offerings in business and technology domains can be viewed through the lens of 1. applying appropriate algorithmic techniques, 2. Choosing right deployment platforms and 3. Creating incubation environment to rapidly develop algorithm to beat the competitive market.

Consequently incubation, development and deployment of such complex AIML technologies (and successfully taking them to market), while needing strategic frameworks to plan competitive go-to-market plans for the enterprise businesses and with the core platform and algorithm suppliers may not be expected to be vertically integrated, eco-system integrating diverse Independent Software Vendors (ISVs), each providing, with its individual product and services for integration support to the complex engineered multi-part solution, needing to interact cohesively to succeed. This leads us to explore the various combinations of strategic best-fit options to tackle the challenge.

In this section we review the adoption of novel AIML algorithm and platform through case studies and further illustrate the application of strategic and tactical frameworks researched below, and apply them through case studies in specific technology and business management scenarios where AIML infusion is imperative.

Competition, Coopetition and Open-source in Business and Technology.

For competitive analysis in any market, the importance of SWOT analysis cannot be understated, as noted by Benzaghta et al. (2021):

“In a competitive environment, enterprises need to take advantage of any opportunity to optimize their business developments. A SWOT analysis is used more frequently than any other management technique.”

The strategic importance of the application of SWOT for analysis of competition and strategies based on internal competencies have been noted in (GÜREL and TAT, 2017; Taherdoost and Madanchian, 2021).

Further to analyze competitors in the marketplace, Meyer and Volberda (1997) in their paper argue that as per Porter, strategizing for corporates mainly involve finding answers to how they will be "composed" based on synergies and how their array of businesses be "controlled" to effectively manage those anticipated synergies. They further high-light that the emergence of BCG Matrix and other portfolio grid techniques, developed based on Porter's theory, to commonly address evaluation metrics of the corporation's businesses based on their strength and attractiveness in their respective industries.

The authors explain Porter's theory on Corporate Strategy of "Disaggregation" to manage complexity of corporate diversity emerges as a theme for "control", while opting for "Composition", to select only "those businesses that have good skill transfer and activity sharing potential" further to clarify that each move to compose or diversify should pass an array of tests to create share-holder's value based on "attractiveness" test, "cost of entry" test and "better-off" tests.

However, as suggested lately by Bruijl (2018), while being hugely successful in shaping corporate strategies, there are inadequacies in Porter's theory of Five Forces for analyzing competition in the present innovative and rapidly changing business environment. Bruijl refers to "Blue Ocean" strategy creators (Kim & Mauborgne, 2015) to point out that for value innovation, Blue Ocean Strategy (Burke et al., 2016; Wang, Lin and Chu, 2011) may be considered a more focused strategy to follow in the competitive scenarios of current business settings. The author further clarifies that the Blue Ocean Strategy generally applies when "demand is created in an environment where the rules of the game are not set" and "by not combatting one's main competitors."

However, in a complex and wider context, where the industry's solution innovation demand innovators across industry to participate to create value, coopetition strategy seems to provide a strategic forward-looking solution (Dagnino and Padula, 2009), pointing out the relevance of coopetition strategy in the form of "interfirm dynamics" to elucidate the strategic interplay among "coopetitors", further referring to complex dyadic coopetition scenarios in the automobile industry, which, resonates with the present-day dynamics and co-opetition rules (Brandenburger and Nalebuff, 2021) for dyadic or even multi-party eco-system formation between the AIML Software Development companies, Cloud Platforms Companies and Technology and Business Service Providers.

In a separate paper, Rusko (2012) finds that coopetition itself has been narrowly and implicitly defined as a "dyadic relationship" emphasizing tension in such cases, coopetition, that is cooperation between two (or more) competing firms:

"Even though dyadic coopetition might create value for the two competing firms, it will not necessarily create value for other stake-holders such as consumers or the public sector (society)."

The papers (Rusko, 2012; DiVito & Sharma, 2019) further delve into multi-party or "multilateral" coopetition, capturing its very nature as:

"a contextual coopetition network comprising of two (or more) coopetitive firms in which also at least one or more actor, such as own or foreign government, customers or other stakeholders of the firms are involved in."

The paper by Roth et al. (2019) and the above discussions strongly point to the relevance of eco-system partnership formation, even including rivals.

Divito & Sharma (2019) interestingly note:

"…multilateral coopetition is relevant for radical innovation, and dyadic coopetition is more suitable for incremental innovation, which has implications for value captured by all parties. Second, implicit in the coopetition literature is the idea that coordinating activities between competitors is important for the coopetition's success, and hence such coopetition introduces another actor to play a coordinating role."

Thus, coopetition implementation choices may be classified as either focused two-party technology creation initiative or in a complex multiparty-based service development eco-system, emphasizing the relative moderation of social capital achieved by the two eco-system partnership strategies.

The paper by Aarikka-Stenroos and Jaakkola (2011) brings in an important perspective on dyadic "co-creation" framework, in where value "co-creation" in knowledge intensive business services is undertaken jointly by the Supplier and the Consumer, and presents it as a "joint problem-solving process", as exemplified in the figure below:

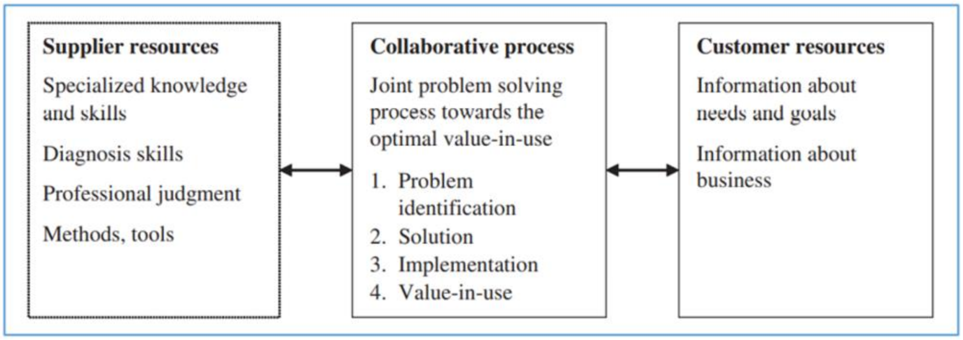

Supplier resources
Specialized knowledge and skills
Diagnosis skills
Professional judgment
Methods, tools
Collaborative process
Joint problem solving process towards the optimal value-in-use
1. Problem identification
2. Solution
3. Implementation
4. Value-in-use
Customer resources
Information about needs and goals
Information about business

Figure 4.20
Dyadic co-creation Framework. Courtesy: Aarikka-Stenroos and Jaakkola (2011)

Further the "value capture" by the firm and the customer respectively have been emphasized in the paper by Storbacka and Nenonen (2009), further streamlining and segregating the dyadic value co-creation components into "resources" and capabilities", as captured in the figure below:

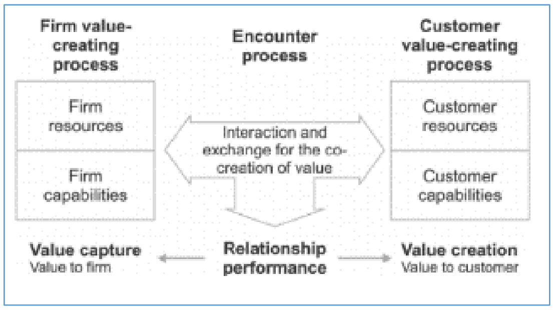


Figure 4.21
Dyadic Value co-creation components. Courtesy: Storbacka and Nenonen (2009)

Interestingly, the evidence of success of the above co-creation frameworks has been displayed by the sincere joint service creation initiatives undertaken by Samsung (Prime 4G Network Gear Supplier Firm) and Jio, the 4G Service Operator (Customer) to create and commercialize the very successful AIML-enabled 4G Wireless Network in India (RCR Wireless News, 2017).

Duc et al. (2017) study the software firms which strongly adopt the innovation strategy to extend value creation beyond the firm's boundary. They note:

"The participation in an open and independent environment also implies the competition among firms with similar business models and targeted markets. Hence, firms need to consider potential opportunities and challenges upfront."

On this novel software partnership eco-system, their below observation is significant:

"Increasingly, software products are no longer developed solely in-house, but in a software ecosystem (SECO), where developers collaborate with "distributed collaborators" beyond their firm boundary. This differs from traditional outsourcing techniques in that the initiating actor does not necessarily own the software produced by contributing actors and does not hire the contributing actors. All actors, however, coexist in an interdependent way."

This study explores how software firms interact with others in OSS ecosystems from a coopetition perspective.

The case for partnership in opensource based software systems, is further strongly presented in the article by Ruffatti (2009). Here he presents the following:

- historical evidence of a large Information Technologies (IT) firm in developing and managing free open-source software (OSS) projects.
- examines reasons, strategy, relations with the communities and results, and lastly
- provide lessons learned as a base for further developments and initiatives.

Illustrating, how OSS allows a steady movement from collaboration to coopetition, the author finally states that simultaneous cooperation (in non-monetary issues) and competition (in the same market) enable the complex relationships fosters OSS the ecosystem. This project thus presents us with an-eye-opening illustration.

Having investigated the competitive strategy landscape with a focus on innovation and co-opetition in the OSS world, we finally develop an amalgamed view, as a combined strategy to tackle novel AIML-laced, innovation-based business and technology transformation view with open-source software, based on Porter 5-forces competitive theory, Blue Ocean strategy and co-opetition strategy, as shown in the figure below.

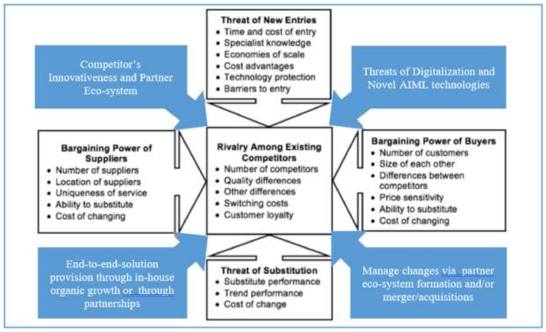


Figure 4.22
Five Forces, amalgamated with Blue Ocean & Coopetition Strategies. Picture adapted from Bruijl (2018).

From the perspective of application of Value Generation in enterprises by AIML technology, the Enterprise AI canvas based on the famous BMC model has been proposed in the paper by Kerzel (2021) and has been “designed to bring Data Scientist and business expert together to discuss and define all relevant aspects which need to be clarified in order to integrate AI based systems into a digital enterprise.”

He further explains the Canvas as consisting of two parts where “part one focuses on the business view and organizational aspects, whereas part two focuses on the underlying machine learning model and the data it uses”.

From the perspective of Return on Investment (ROI) of AIML-laced businesses, however it seems that the previous industry valuation methods as was suggested by (Visconti, 2019) are possibly not enough to tackle and take to market these new technologies and thus may not be sufficient to measure future value innovation success. Rapid innovation in an accelerator environment with demonstrated incremental value in a phased-manner for staged VC funding or corporate funding, aiming to develop product or service with a focus on creating new market adoption opportunities by the industry participation in a proof-of-concept with co-creation and testing projects, thus seem to be the order of the day.

We thus envision corporates to adopt tactical innovation gameplan based on open-source, and an “amalgamated” competition and coopetition based on accelerated innovation business strategy to succeed in the AIML-based digital transformation of their businesses.

The above view is further corroborated by the break-neck speed of innovation that Generative AI has created and thus making it apt to review where they stem from. Thus, the adoption of “Corporate Accelerator” and “Private Accelerator” environments for fast innovation is reviewed and further, a general tendency of vertical integration in the big-tech to integrate the innovations to win the race of “General AI” supremacy has been noted in our case study.

The paper in the Journal of Business Models, the paper by Bagnoli et al. (2020**)** describes the Accelerators as “…an increasingly diverse set of programs and organizations and, often, the lines that distinguish accelerators from similar institutions, like incubators and early-stage funds, become blurred…” and those that “bear some similarities to incubators and angel investors” and “all help and fund nascent ventures, offering educational components”.

They further differentiate the Accelerators as “fixed length of the program, its intensity, the provision of benefits and services, and the cohort-based nature” from the incubators, which “lack a fixed term and do not typically provide equity investment in return for cash” but draw similarity with the “business angel investors”, those that include “wealthy individuals who invest their own money into early-stage start-ups, usually having previous experience in seed investing or who might have started a few businesses on their own before” and both showing the possibilities to “improve the survival rate of the start-ups.”

The paper further studies the nine building blocks of the Accelerator’s Business Model Framework proposed by Biloslavo, Bagnoli and Edgar et al. (2018), differentiating the model from the Business Canvas Model proposed by Osterwalder and Pigneur (2010) and the Lean BMC (Maurya, A. (2012). The reviewed Accelerator framework is shown in the figure below.

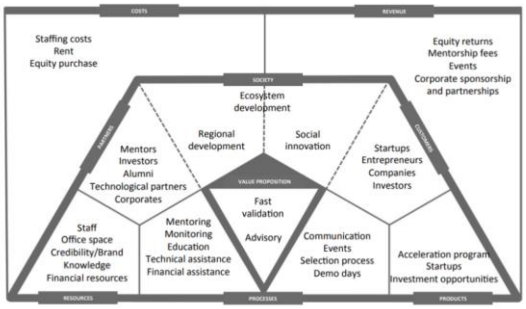


Figure 4.23
Accelerator's Business Model Framework. Courtesy: Bagnoli et al. (2020); Biloslavo, Bagnoli and Edgar et al. (2018).

As shown in the figure above, compared to the BMC of Osterwalder and Pigneur (2010), the accelerator's value creation view is broadened and includes "benefit" component, along with the "cost"; and also considers all value generating resources.

Further, the Accelerator and Incubator ecosystem models catering to value creation from a "Country-specific Strategic Innovation competence development" and the corresponding viewpoints and roles in Europe, Australia and India have been studied and captured as well (Santiso, 2013; Bliemel et al., 2016; Bhagavatula et al., 2017).

The specific requirement to develop expertise in Business Accelerator design to harness innovation and value capture as an intermediary in an entrepreneurial ecosystem has been studied by Bhagavatula et al. (2017)**.**

Finally, focusing on the Accelerator for new start-ups in San Francisco, US, where the innovation landscape has been greatly influenced by Y-combinator **(**Y-Combinator, 2023; Levy, 2021)**;** and the start-ups from the stable have created global impact in SaaS and cloud-based businesses, harnessing latest AIML technologies, harnessing "Y-Combinator's Fresh Approach to Innovation", as noted in Harvard Business Review article (Anthony, 2009).

Lately, novel innovative business models have been created; enabling formation of complex partnerships, and presented in the figure below (OpenAI, 2023).

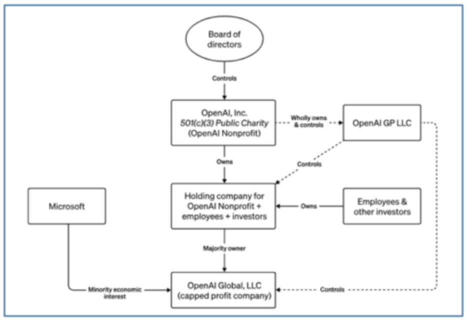

Board of directors
Controls
OpenAI, Inc.
501(c)(3) Public Charity
(OpenAI Nonprofit)
Wholly owns & controls
OpenAI GP LLC
Owns
Controls
Microsoft
Holding company for OpenAI Nonprofit + employees + investors
Owns
Employees & other investors
Majority owner
Minority economic interest
OpenAI Global, LLC (capped profit company)
Controls

Figure 4.24
Futuristic Board structure of OpenAI. Courtesy: OpenAI, 2023, available at [https://openai.com/our-structureOur structure](https://openai.com/our-structure)

The complex futuristic board structure created, allows OpenAI Inc. to accelerate Generative AI innovation and testing using Microsoft cloud infrastructure; as testing of LLM models need significant computation power, as was reported in late 2020s, as "OpenAI found that that the amount of computational power used to train the largest AI models doubled every 3.4 months since 2012" (Hao, 2019). However, lately, Piper (2024), reported that "OpenAI is attempting to transition to a more conventional organization structure", that of a for-profit public benefit corporation.

Further, using appropriate rules of engagement, offered above, allow the entrepreneurial companies to gauge their progress and based on the emergent ever-changing business environment, decide on either "pause" or opt for vertical integration or even change the nature of co-opetition strategy altogether, from a dyadic to multi-party ecosystem; but normally, the expectation for an entrepreneurial company creating the innovation, would be to graduate to the "next stage VC funding" discussions for taking the innovation to the market.

CHAPTER V:

RESULTS

5.1 Industry Trends and Impacts of Novel AIML Technologies in Business and Technology

In order to assess the impact of AIML in Business and Technology, we utilize latest research data on AI adoption in industries and undertake an analysis of early trends to support immediate impacts based on short-term data and assess possible continuing impacts in the future.

Support for AI adoption in Industry - Themes for AI Mentions in Fortune 500 Earnings Calls

From the perspective of data, specific chart of the Stanford University Report (HAI, 2023) presented in the AI impact market survey, showed increasing trend of “mentions” of the novel technology adoption in the earning calls in the Fortune 500 Companies, between years 2018 and 2022, as shown in the figure below.

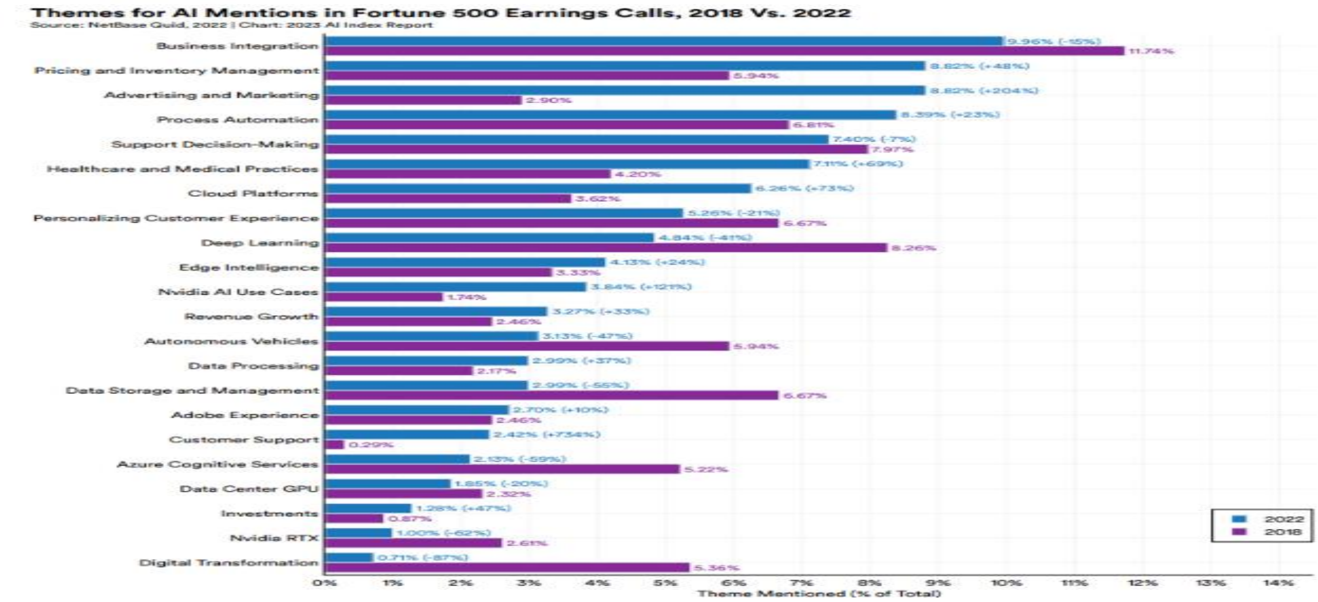


Figure 5.1
AI Earning Call and mentions of AIML-related technologies: Courtesy: HAI, 2023

This, thus signals high-levels of commitments in the industries, leading to allocation of both human (AI engineers) and financial resources to generate early leads in AIML technology innovation and their operationalization that corroborates the positive growth trend of adoption of novel AIML Technologies, as illustrated in the further trend analysis, discussed below.

Creation of "Open Access" Foundation Models – Possible Democratization of Novel AIML technologies via viral adoption of GenAI in Industries

The numbers of Open Access Generative Models created, as shown in green bars, and highlighted by the over-lapped triangle in the below figure, shows a high linear growth trend, between years 2021 and 2023.

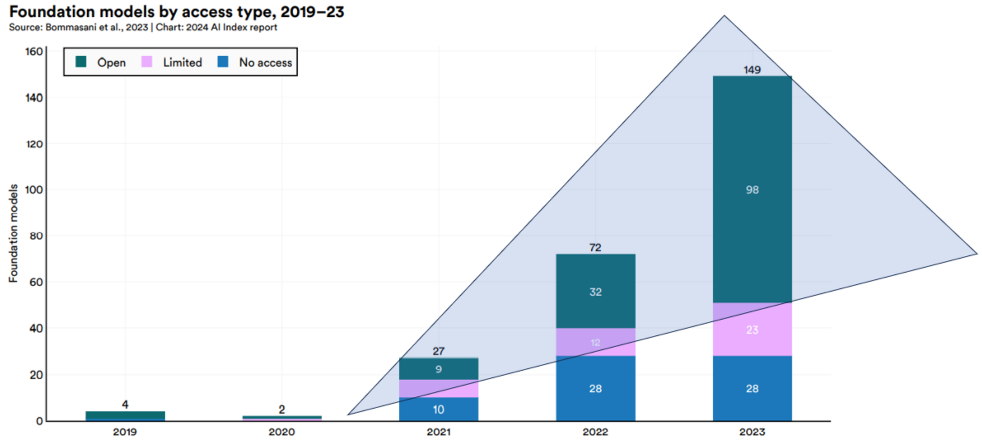
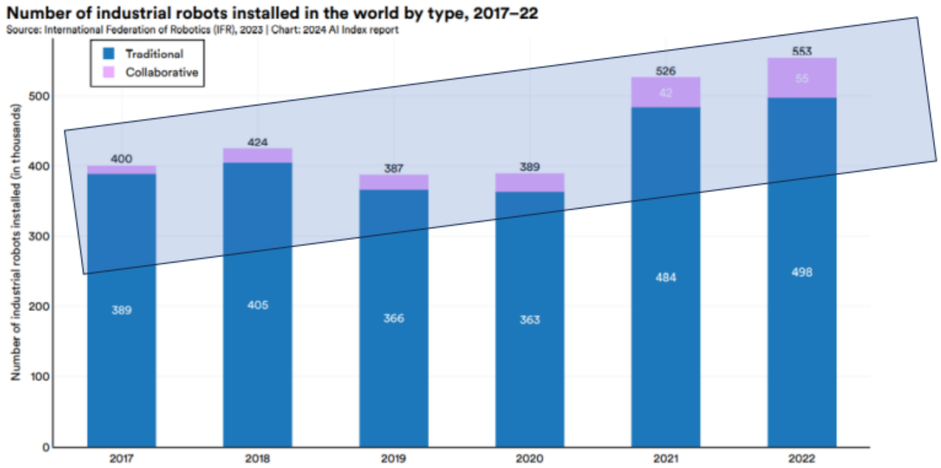


Figure 5.2
Trends in Creation of Foundation Models for Generative AI: Courtesy: HAI, 2024

While the general trend of the created AI Foundation models, is an increasing one (Figure Appendix A.1), possibly not expected to grow as much in the next few years, they still show promising signs of viral adoption as they are low-cost and open-source based, democratizing the usage of Generative AI models for experimentation and non-critical business use as well. Also, the increasing trends in creation of "limited" and "closed" source Foundation models indicate market traction for need and adoption in Industries that rely on critical infrastructures, needing high quality models (hence, costly and closed-source). This allows the company's goals of GenAI deployments to achieve immediate cost reduction and future business growth.

Global Installation of Collaborative, AI-enabled next-gen robots - establishing rapid adoption of AI in robotics

Number of global installations of "Collaborative AI-based Robots" (based on Reinforcement Learning laced AI models) in industries, shown with pink bars, and highlighted by the over-lapped rectangle in the below figure, indicates a near-exponential level in the growth trend between years 2017 and 2022.


Number of industrial robots installed in the world by type, 2017–22

Source: International Federation of Robotics (IFR), 2023 | Chart: 2024 AI Index report

Figure 5.3
Trend in Implementation of Collaborative AI Robots: Courtesy: HAI, 2024

This, though shows an early trend (Figure Appendix A.2), however, if it lives to the promises, is most likely to herald Industry 5.0 revolution, with rapid introduction of Intelligent robots in critical industry functions, reducing human and associated resource costs while hugely increasing efficiency and productivity.

However, to be cautious, it is important to note that "rapid creation and introduction" and "high-quality implementation and operationalization" are competing cost versus value imperatives; as is the rapid revenue and profit generation and cost reduction objectives of Industries versus the societal safety and overall good.

5.2 Meta-Analysis of AIML impacts through Case Studies

Noting that for innovative Platform business based on AIML, gauging their level of competency to succeed are essential, we undertake a thorough analysis of their management with case studies employing industry-leading strategic frameworks. The tools and techniques, used to explore the simulated case studies for studying the strategic and tactical and rules of engagement, are as follows:

a. Dyadic and Multiparty Coopetition Strategies are illustrated with Case Studies 1 & 2, using NWDAF, MEC and UAVs/Drones as AI-impacted industries.

b. Platform Business Model Map for eco-system analysis is illustrated in Case Study 3, utilizing MEC based on UAV/Drone technologies in a Cloud-based AIML Platform.

c. Case study 4 illustrates Roger's "Value-Train" analysis for tactical value generation/transfer in dyadic and multiparty coopetition settings respectively in Generative AI Industry, for Case Study 4;

Through the above lens, we delve into eco-system partnership, build vs. buy strategies and their adoption to navigate through this decision-making process of developing and adopting AIML in business and technology.

- **Case Study 1 – Intelligent Agent-based Framework in MEC and UAVs guided by Multiparty co-opetition Strategy**

The use case presents the application of novel AIML technologies in edge computing and analytics applications, focusing on wireless industry, where AI/ML frameworks for edge networking promises flexibility, scalability, software/hardware reuse and automated system design has been thoroughly researched in the white paper on Edge Networks by 5G Americas (Edge WP, 2019). The companies considered for the case study are Microsoft and various supporting companies for MEC Platform, as indicated in Table 5.1; as well as aerial Drone and UAV vendors.

Further from Mobile Network domain perspective, we note that Mobile Edge Computing (MEC) is an accepted 3G PP Framework and standard for global deployment and the source of the Reference Architecture for global deployments are presented in the ETSI Paper on Mobile Edge Computing (ETSI GS MEC 003, 2016).

The exploitation of AIML in a distributed multi-agent system setting for MEC has been accurately captured in the paper by Leppänen (2019), where he focuses on the strategic utilization of software agent technologies in MEC.

The paper captures the essence of complex integration and management of distributed services on heterogeneous system components with agents, and specifically for MEC, integrates the ideas and proposals of using interface agents to integrate disparate 3rd party cloud computing platforms into the system and to manage interactions between the platforms. The distributed AIML agent-based architecture of MEC has been presented in his paper, capturing the essence of agent-based coordination of the 3rd-party service delivery to IoT systems, control agents are utilized as a MAS to create adaptive and decentralize services for Fog computing, and its computing system orchestration with performance monitoring agents. The ideas of optimization of task assignment, with negotiating agents, between cloud and edge platforms has been well integrated in the paper, as is the idea of collaborative microservices being are modeled as a MAS for IoT.

Complex Multiparty coopetition strategy for MEC Platform

As already postulated by Cohen of Economic Strategy Institute (2020), 5G Private Networks with MEC will allow firms to extend their value chains beyond traditional internal or manufacturing value chains, providing greater workflows across firms and more profitability from expanded value chains.

This includes extending the firm’s value chain by providing data-based insights to its customers:

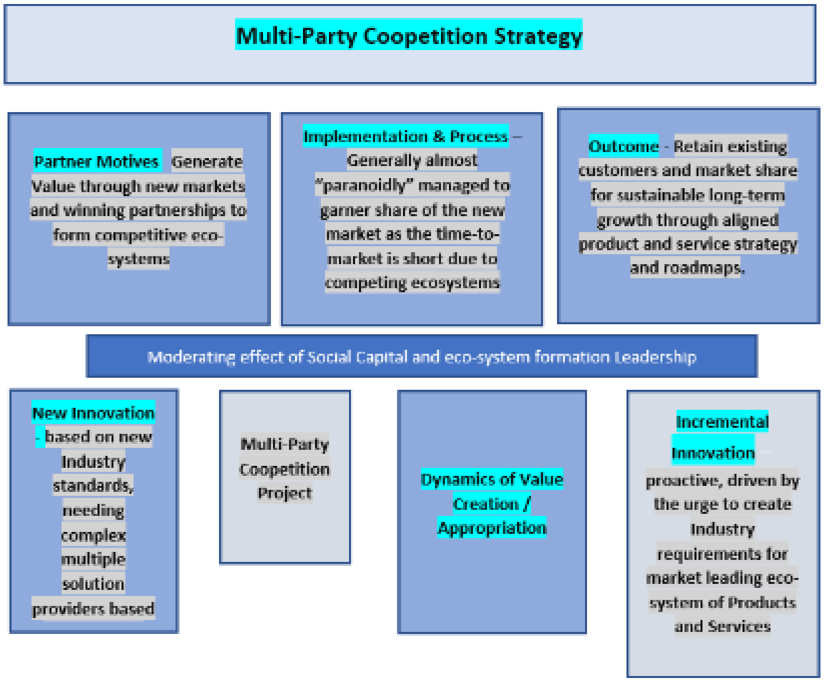
• With APIs and Edge Computing, firms can take their value chains far beyond connected manufacturing and predictive maintenance.

• They can provide new services to customers by analyzing data they access at the customer’s sites with IoT for Smart Factory and Home automation Management.

• Firms provide suppliers and customers with the ability to access data on the machines or products they sell. This supports finer control of services.

However, the services need tight partnership between the OEM (MEC service provider), Hyper-scaler (AWS, Microsoft, Google), Operator and Operator Partners which can only be controlled and exploited from the view-point of value co-creation for complex multi-party projects (Dagnino and Padula, 2009). For the MEC product and Services perspective, “Microsoft as a Prime” presents a classic example of a complex “multi-party” or “multilateral” (DiVito and Sharma, 2019) eco-system and the complexity is further noted:

The complex collaborative multi-party solution development relationship for MEC, fostering co-creation and governance for an industry leading solution can be demonstrated and illustrated by the diagram below:



Multi-Party Coopetition Strategy

Partner Motives Generate Value through new markets and winning partnerships to form competitive eco-systems

Implementation & Process – Generally almost "paranoidly" managed to garner share of the new market as the time-to-market is short due to competing ecosystems

Outcome - Retain existing customers and market share for sustainable long-term growth through aligned product and service strategy and roadmaps.

Moderating effect of Social Capital and eco-system formation Leadership

New Innovation - based on new Industry standards, needing complex multiple solution providers based

Multi-Party Coopetition Project

Dynamics of Value Creation / Appropriation

Incremental Innovation – proactive, driven by the urge to create Industry requirements for market leading eco-system of Products and Services

Figure 5.4

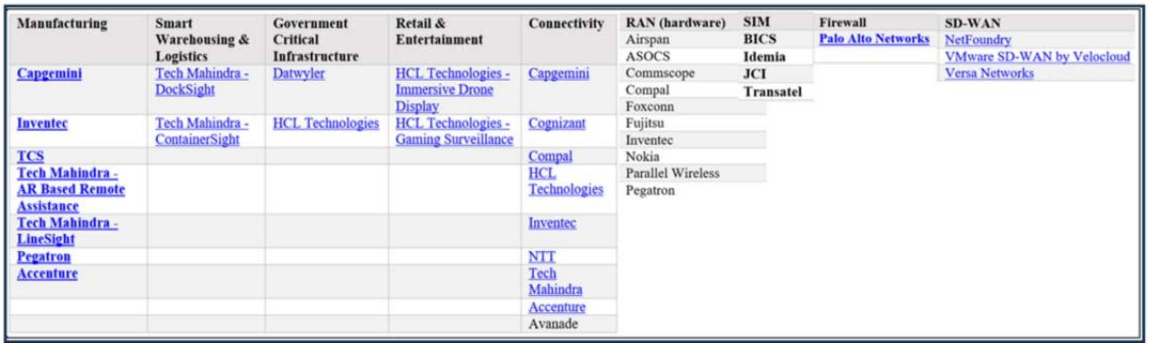
Illustration of Multi-Party Coopetition Strategy Implementation. Adapted, Courtesy: https://www.slideteam.net/coopetition-strategy-showing-implementation-outcomes-and-process.html

The high-lights are that incremental value realization is appropriated for complex new innovation projects for services or products involving multiple parties, where the over-arching motives of all partners are to generate value for the new market and rapid business generation based on evolving industry standards.

The resulting outcome for all partners is of an energizing and invigorating environment for value creation resulting in the formation of winning eco-system for business and technology partnership. Normally, the leader of the Group governs and drives the eco-system, supported by partners and operate as a Multi-sided Platform (MSP).

Further, we explore with an example of the Complex MEC ecosystem, governed and driven by Microsoft as a leader or prime in the solution co-creation project, managing the complex eco-system of Customers, Partners and their Services in a decentralized yet tightly governed manner. Here multiple parties compete in the service arena, yet cooperate for the specific client solution, as governed by the Customer and Azure and through diverse marketing channels of each partner, who "bring in the business" and increases the network effect.

Thus, Azure private multi-access edge compute partner solutions demonstrate a rich ecosystem management with multiple Partners involved in co-creating the MEC solution, which the Customer, together with Azure can partner with, forming a winning ecosystem as noted by Liu, Ovhal and Dwivedi, K. (2023); the evidence of the availability of the rich ecosystem is thus well advertised in their website and is further illustrated in the below table.



Manufacturing	Smart Warehousing & Logistics	Government Critical Infrastructure	Retail & Entertainment	Connectivity
Capgemini	Tech Mahindra - DockSight	Datwyler	HCL Technologies - Immersive Drone Display	Capgemini
Inventec	Tech Mahindra - ContainerSight	HCL Technologies	HCL Technologies - Gaming Surveillance	Cognizant
TCS				Compal
Tech Mahindra - AR Based Remote Assistance				HCL Technologies
Tech Mahindra - LineSight				Inventec
Pegatron				NTT
Accenture				Tech Mahindra
				Accenture
				Avanade

RAN (hardware)	SIM	Firewall	SD-WAN
Airspan	BICS	Palo Alto Networks	NetFoundry
ASOCS	Idemia		VMware SD-WAN by Velocloud
Commscope	JCI		Versa Networks
Compal	Transatel		
Foxconn			
Fujitsu			
Inventec			
Nokia			
Parallel Wireless			
Pegatron			

Table 5.1
Azure MEC Ecosystem. Courtesy: Microsoft, available at https://learn/microsoft.com/en-us/azure/private-multi-access-edge-compute-mec/partner-programs

Complex multi-party Co-opetition strategy also comes into play in the aerial MEC using UAVs/Drones (Yazid et al., 2021; McEnroe, Wang and Liyanage, 2022) that are expected to invigorate the adoption of 6G Wireless Networks, especially for Digital-Twin-based aerial edge computing systems for Next Generation Wireless Networks (Abir and Chowdhury, 2023).

It is to be noted that UAV/Drones use various AI Technologies like Machine Learning, Deep Learning, Computer Vision (CV), Reinforcement Learning, Natural

Language Processing, Swarm Intelligence, Genetic Algorithms, Sensor Fusion, Edge AI and Anomaly Detection to provide disparate services like Precision Agriculture, Mapping and 3D Modeling, Erosion and Coastal Monitoring to name a few, has been well explained in the book by Shinde, More and Chaudhari (2025).

Needless to say, that each Service will need multiple OEM vendors to provide the niche technology and services and integration by a Prime Vendor and the value generation success will depend on rich ecosystem and partnership formation to provide cost-effective services.

For MEC Services, the UAVs extend the service range. A view of multi-UAV assisted MEC system is shown in the following figure (ref. Service architecture using Drones is shown in the below figure (Chen et al., 2023).

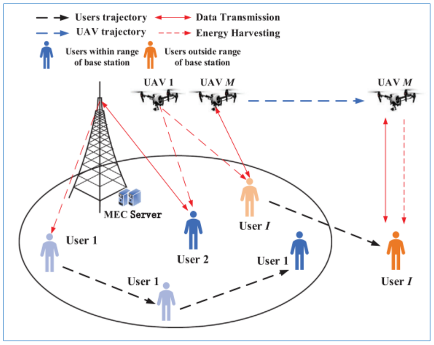


Figure 5.5
Illustration of Multi-UAV assisted MEC system Courtesy: Chen et al., 2023, DOI: 10.1049/cmu2.12596

The architecture aids to offload and allocate computation workload to the Base Station and the UAV, optimizing Energy efficiency by freeing the MEC Server and also aiding to extend the MEC services to the users who are outside the range of the Base Stations.

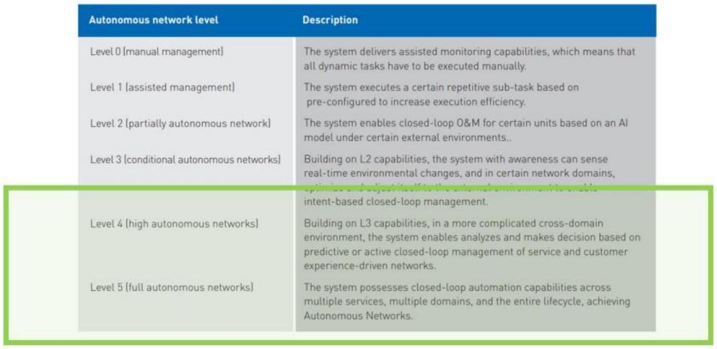
Thus, the multi-party Coopetition Strategy as analyzed in this Case Study, will be essential for a “tear-down analysis” of the complex matrix of Technologies and Services for UAV-based systems for strategic management of future generation technologies.

- **Case Study 2 – Intelligent Autonomous Networks for NG Mobile Network Data Analytics governed by dyadic co-opetition strategy**

This case study considers two leading companies, co-developing Network Data Analytics Platform solutions - SaS and DigitalRoute (SAS NWDAF, 2022; DigitalRoute, 2022).

For a background on associated technologies of MEC Platform, we note that the ETSI White Paper on GANA (Meriem et al., 2016) originally proposed as "Autonomic network engineering for the self-managing Future Internet (AFI)", specifying the instantiation and implementation of the GANA Model onto Heterogeneous Wireless Access Technologies using Cognitive Algorithms have evolved in step with the IETF standards and further updated and evolved to Standards paper (ETSI GANA, 2020).

As per the over-arching vision in the GANA white paper, it is thus expected that with the "automation oversight", serving to provide to the underlying network which needs to function autonomously, employing the AIML methods, possibly like those of MORL and HRL, to "close the machine learning, analysis and eventual deployment and feedback loop"; this would thus achieve autonomous network behaviors.



Autonomous network level	Description
Level 0 (manual management)	The system delivers assisted monitoring capabilities, which means that all dynamic tasks have to be executed manually.
Level 1 (assisted management)	The system executes a certain repetitive sub-task based on pre-configured to increase execution efficiency.
Level 2 (partially autonomous network)	The system enables closed-loop O&M for certain units based on an AI model under certain external environments..
Level 3 (conditional autonomous networks)	Building on L2 capabilities, the system with awareness can sense real-time environmental changes, and in certain network domains, optimize and adjust itself to the external environment to enable intent-based closed-loop management.
Level 4 (high autonomous networks)	Building on L3 capabilities, in a more complicated cross-domain environment, the system enables analyzes and makes decision based on predictive or active closed-loop management of service and customer experience-driven networks.
Level 5 (full autonomous networks)	The system possesses closed-loop automation capabilities across multiple services, multiple domains, and the entire lifecycle, achieving Autonomous Networks.

Figure 5.6
Maturity Levels of autonomous networks run from level 0 to level 5, Courtesy Analysis Mason: Source: TM Forum (2020), Autonomous Networks: Empowering Digital Transformation for The Telecoms Industry. Available at: https://www.tmforum.org/resources/whitepapers/autonomous-networks-empowering-digital-transformation-for-smart-societies-and-industries/

The idea of enabling the CSPs to move up in the level of network autonomy maturity level, was further defined by the TMForum (2020), and re-defined by Analysis Mason (Rao et al., 2020), as captured in the figure above.

In accordance with the above figure, the paper by Analysys Mason (Rao et al., 2020) further discussed the stepwise journey of the Communication Service Provides from automatic towards autonomous networks, as discussed with the figure dictated by the proposal by TM Forum, defining the maturity levels of autonomous networks, thus providing an industry standard for evaluating the maturity of autonomous networks.

According to interviews conducted by Analysis Mason:

"…most CSPs are at level 2 in TM Forum's classification system, with more advanced CSPs operating some domains at level 3. Most CSPs have started to lay the foundations to take their whole network to level 3 and will continue to expand autonomous functionality through their networks incrementally to progress to the higher levels.

The journey to the fully autonomous network is expected to take at least a decade. New technologies and services will be added to the network over the years to come, which will create additional requirements for what constitutes a fully autonomous network."

Echoing the above Industry sentiment, the paper by Arzo et al., (2022) very recently defined a novel Agent-Based Intelligent Network Architecture, integrating the ideas of the ETSI-GANA model to propose a 6G autonomous network, based on agent framework and captures essence of the futuristic journey of network autonomy.

Application of Federated Learning in NWDAF for Autonomous 5G Networks

In the Core of the 5G Network the architecture of Network Data Analytics Function (NWDAF) has been described in ETSI technical specification (3G PP TS

23.501, 2018) as one that “represents operator managed network analytics logical function. NWDAF provides slice specific network data analytics to a Network Function (NF).”

NWDAF primarily collects data from 5G Core Network Functions and the operations support system (OSS); hosting AIML algorithms and deploying their services in order to translate them to network operations related insights. NWDAF insights are mainly applied to 5G core networks to enhance their functionality. Optimized data collection and storage has been specified in the Standard, together with training and ML model retrieval.

NWDAF responds to the urgent need to reduce operational costs while supporting the rapid introduction of new services and products and identifying and leveraging monetization opportunities. Centrally or distributed intelligent standard specified platform architecture and strategy thus form the key to unlock the opportunities for the BSS/OSS and Telco vendors.

The journal paper by Sevgican et al., (2020) elaborates NWDAF as a “newly defined data analytics function in 5G cellular networks that provides network analysis upon request from other Network Functions, using any other NF as a data source”. This “two-way” relation between NWDAF and NFs as depicted in the below figure.

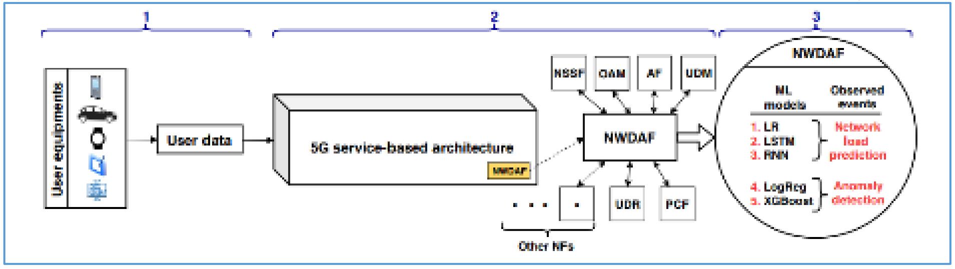


Figure 5.7
5G Network Components and NWDAF. Courtesy: Sevgican et al., 2020

The paper further explains the two available services, namely, events subscription (i.e., analytics subscription), and analytics information retrieval service that requires the application of AIML techniques. To note, the paper further captures the important event that can be observed, which could be useful to develop the possible AIML algorithms and applications that may need to be deployed to trigger the AIML data collection and analysis by the AIML algorithm and retrieve the results related to a specific event.

The paper by Pateromichelakis et al. (2019), referring to 3GPP discussions about "How SA5/ SA2 / RAN3 Could Work Together to Guarantee Network Slice SLA", provides an overview of how analytics can be used across CN, RAN, and OAM to enable network automation, specifically pointing out the areas of centralized analytics and distributed analytics, as illustrated in the below figure.

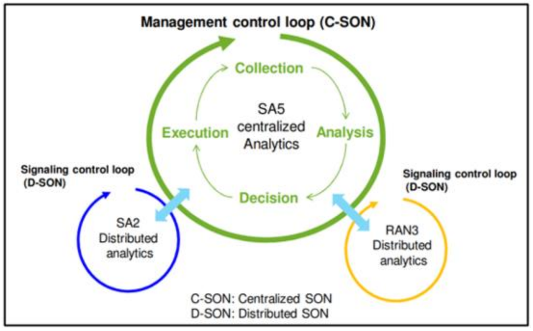


Figure 5.8
Cooperation of Network Components to guarantee Network Slice SLA. Courtesy: Pateromichelakis et al. (2019)

The above discussion paper thus conclusively sets the industry-direction for current and future evolution of design and development based on new requirements for C-SON and hence helping to set NWDAF's architectural roadmap for various OEM vendors.

We specifically point to an industry architecture (TM Forum Insight, 2020) in the figure below that expectedly aligns with the recommendation of a distributed NWDAF architecture for 5G Networks, with a Central and multiple Edge NWDAFs

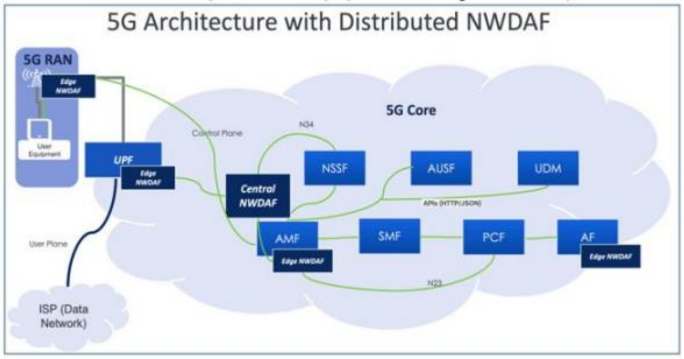


Figure 5.9
5G Architecture with Distributed NWDAF. Source TMForum Insight, 2020: *https://inform.tmforum.org/features-and-opinion/nwdaf-automating-the-5g-network-with-machine-learning-and-data-analytics*

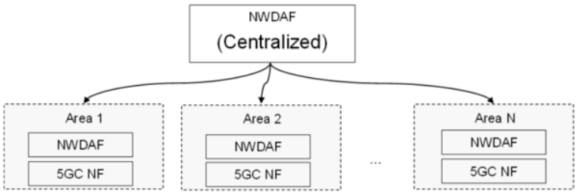


Figure 5.10
Hierarchical NWDAF Deployment with Federated Learning in a PLMN. Courtesy: 3GPP Technical Report 23.700-91., 2020 (Page 133)

With the above figure and example, we conclude our technical research on the case study, providing evidence of application of state-of-the-art machine learning methods to implement autonomous next generation wireless networks.

Dyadic Partnership Eco-System Formation for NWDAF

It is evident that CSPs are futureproof their 5G analytics solution strategies by taking a platform-based approach to NWDAF architectural implementation.

The relevance of a well-architected NWDAF based on an intelligent Platform, leading to the formation of a winning dyadic eco-system is best demonstrated by taking a deep-dive at the industry-leading NWDAF solutions and specifically identifying SaS and DigitalRoute (SAS NWDAF, 2022; DigitalRoute, 2022) and demonstrate their collaborative effort to foster co-creation partnership to co-create an industry leading NWDAF solution based on AIML-based services.

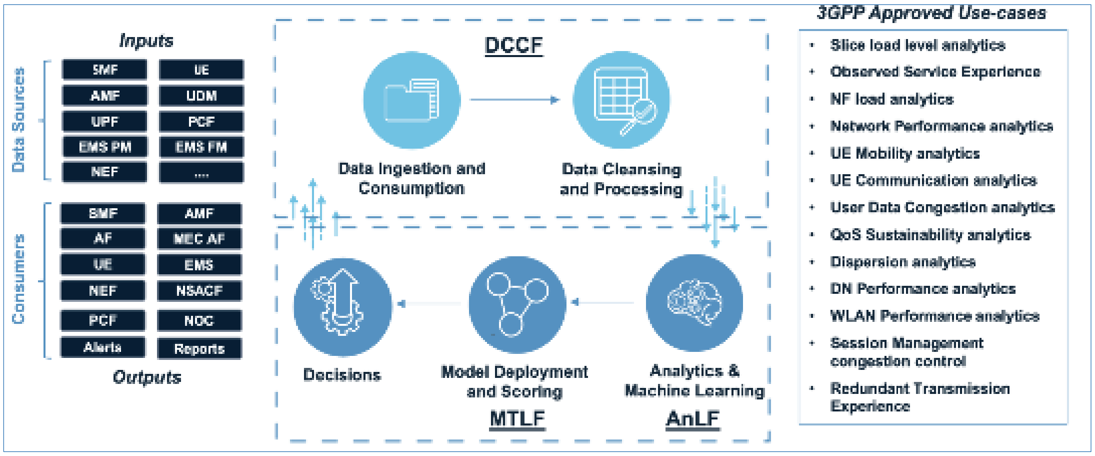

Inputs
Data Sources
SMF
UE
AMF
UDM
UPF
PCF
EMS PM
EMS FM
NEF
....
Consumers
SMF
AMF
AF
MEC AF
UE
EMS
NEF
NSACF
PCF
NOC
Alerts
Reports
Outputs
DCCF
Data Ingestion and Consumption
Data Cleansing and Processing
Decisions
Model Deployment and Scoring
Analytics & Machine Learning
MTLF
AnLF
3GPP Approved Use-cases
Slice load level analytics
Observed Service Experience
NF load analytics
Network Performance analytics
UE Mobility analytics
UE Communication analytics
User Data Congestion analytics
QoS Sustainability analytics
Dispersion analytics
DN Performance analytics
WLAN Performance analytics
Session Management congestion control
Redundant Transmission Experience

Figure 5.11

Disaggregated NWDAF solution from SAS and DigitalRoute – demonstrating dyadic coopetition. Courtesy SAS: https://www.sas.com/offices/pdf/mx/20221103-nwdaf.pdf

SaS declares the collaborative effort as follows:

“DigitalRoute collects and processes any type of data from any system in real-time. DigitalRoute provides multiple data management steps to turn raw data into useful information. Data is normalized and transformed into a consistent and usable format. Data quality steps clean and correct data. Data is combined, aggregated, evaluated and enriched with other data sources.

SAS then provides the advanced analytics, enabling algorithmic selection, model training, model monitoring, model deployment, and intelligent decisioning. Automation in model deployment is enabled with auto-tuning and hyperparameter selection. SAS can update models seamlessly as the data changes in real-time at the edge.”

From a dyadic coopetition perspective, the above business relationship between the two firms stands as evidence for a classic application of the framework, as illustrated below.

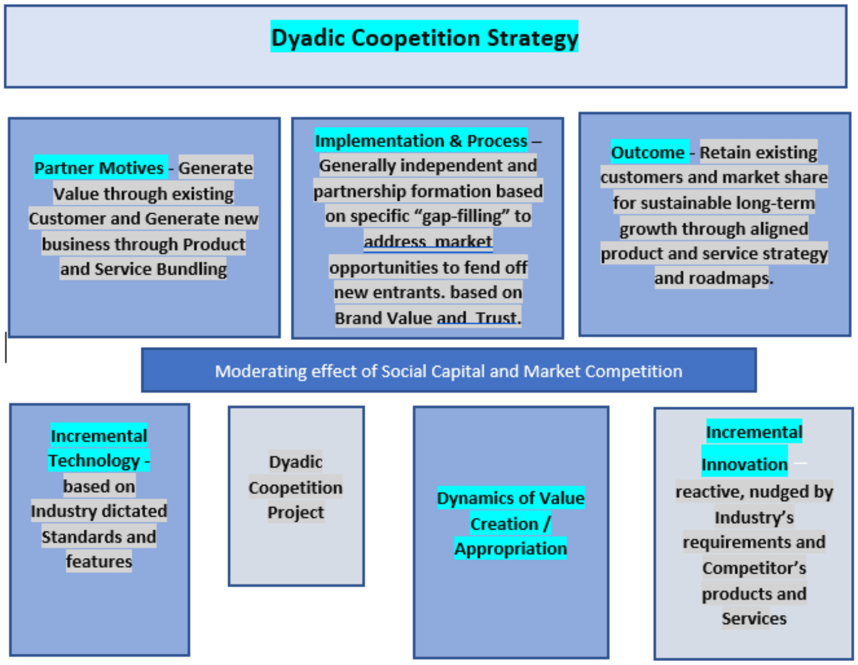

Dyadic Coopetition Strategy
Partner Motives - Generate Value through existing Customer and Generate new business through Product and Service Bundling
Implementation & Process – Generally independent and partnership formation based on specific "gap-filling" to address market opportunities to fend off new entrants. based on Brand Value and Trust.
Outcome - Retain existing customers and market share for sustainable long-term growth through aligned product and service strategy and roadmaps.
Moderating effect of Social Capital and Market Competition
Incremental Technology - based on Industry dictated Standards and features
Dyadic Coopetition Project
Dynamics of Value Creation / Appropriation
Incremental Innovation – reactive, nudged by Industry's requirements and Competitor's products and Services

Figure 5.12
Illustration of Dyadic Coopetition Strategy Implementation. Adapted, Courtesy: *https://www.slideteam.net/coopetition-strategy-showing-implementation-outcomes-and-process.html*

In this tightly coupled dyadic coopetition relationship (Rusko, 2012), each partner draws on its market reach and already generated value for existing customers to further its growth strategy in an existing industry where incremental Standards and features are dictated by Industry Standards. The incremental innovation is nudged by the intent to stay ahead of the competition in a competitive marketplace, each utilizing the value of its partner to co-create solution for the clients as illustrated in the figure above.

Further, we may note that while the industry leading NWDAF solution development companies, namely like Guavus (2022) and Radcom (2020) are competing by exposing standardized APIs to comply with the diverse 5G eco-system, the dyadic ecosystem partnership between SAS and DigitalRoute is indeed a unique solution provisioning methodology in the competitive NWDAF marketplace.

- **Case Study 3 – Multi-sided Platform (MSP) Analysis for Drone-as-a-Service Provisioning ecosystem through Platform Business Model Map**

The lead companies considered for this case study are industry leading AIML Platform vendor Amazon, through their offering of SageMaker Platform and Drone/UAV supplier companies like Aerodyne and others. Since the study is about multi-sided platform there are various other vendors, partners and users in various government and private agencies who deploy aerial Drones/UAVs for their services.

The initial trend of adopting Digital Business Platforms in businesses progressed as a necessary requirement for digital transformation through automation, based on AIML-laced Intelligent Platforms follow similar traits for both ICT and Retail industries, has been introduced as a general framework for 5G by PwC (2020) and is further corroborated by the explicit support in the text on Retail Management by Berman et al. (2018, p 416):

“Retailers envision using the software algorithms that define artificial intelligence (AI)—including pattern recognition, deep-learning neural networks, and computer vision—to help generate the major decisions that help shape consumers’ user experience. It can be used for product recommendations, dynamic pricing, and promotions, and it gets smarter over time, learning from the data. While AI will become more rooted in the retail industry, virtual reality (VR) is also knocking on the door. The goal is to use VR to create more immersive, contextual experiences, customized to highlight every brand’s asset and strengthen their customer relationships.”

The authors go on to predict that the new generation, who are more exposed to video-gaming, are more likely to embrace the new shopping experience based on VR, which is being dubbed as V-commerce.

Thus, the “Intelligent” Digital Platforms, with a singular agenda to enhance

Customer's digital experience and synergize the multi-sided platform (MSP) partner eco-system for both Telcos and Enterprises through a unified Digital Experience Platform (DXP) with an eye to monetize the opportunities, necessitates a migration from product to a platform mindset, leading to an inevitable outcome for the next generation business ecosystem formation to evolve and thrive successfully.

Choudary (2021, p 27) in his latest book, "Platform Scale", emphasize the importance of AIML as one of the key drivers to realize a winning platform ecosystem in a "post-pandemic world", as noted below:

"Improvements in artificial intelligence and machine learning during the 2010s have important implications for the future of platform businesses. Machine Learning creates value through learning effects. The platform gathers data from the activity in its ecosystem and learns more from this data, thereby improving its ability to facilitate future interactions. As the platform captures more usage data, it is able to train and improve its learning models, which in turn allow it to provide more value back to the ecosystem. This leads to more platform usage, as users see more value in engaging with the platform, which leads to the capture of even more data."

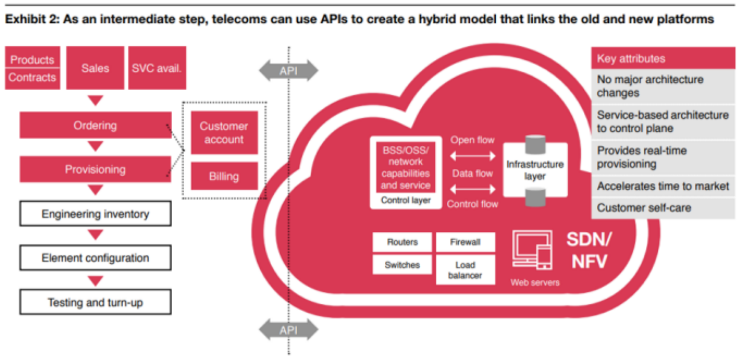


Figure 5.13
Platform automation and staged Transformation of telecoms' internal ecosystems. Courtesy: PwC (2020)

TMForum community, however, proposes a hybrid approach to the evolving platform strategy for the BSS/OSS 5G eco-systems as they need to evolve in stages, without service disruption, from traditional to Service-based Platforms, as shown above. They propose the co-existence and gradual staged evolution be executed through an evolving set of APIs, as illustrated.

Monetization through various marketing channels using DXPs are the essential goal of the Enterprise and Telco Intelligent Platform businesses.

From AIML perspective, in all these platform endeavors, the recurrent theme that emerges is the possibilities to automate the marketing and their operations activities using AIML algorithms in possibly cloud-native environments with programmatic services (Katsov, 2018). The text provides contextual algorithmic frameworks for converged Technology and Business Platforms, including those of product recommendations, dynamic pricing, recommendations and promotions; of which we delve into the recommendation aspect, which is a key pillar of monetization in telecom and retail industries, as its influence in contextual sale and RoI is undeniable as noted by Hinz and Eckert (2010) in their research paper:

“Huge assortments, however, are only beneficial for consumers if their search for appropriate products is supported by tools which help them to identify products that fit to their preferences. Therefore, search and recommendation tools play a crucial role in e-commerce.”

Further, they explain the consequences of search and recommendation systems on Sales and hence RoI impacts on businesses:

“First, decreasing search costs can lead to higher sales based on additional consumption; second, there can also be a shift in demand from blockbusters to niche products and vice versa, so that substitution effects can be observed. These two different consequences (additional consumption and substitution) are of high importance for online retailers: While additional consumption always leads to higher sales and potentially to higher profits, substitution is only advantageous if a low-margin product is substituted by a product with a higher profit margin. However, if providers know about margin differences between products, sales can systematically be shifted to more profitable products by appropriate search and recommendation tools. The basis for such a sales shift, however, is precise knowledge about how various search and recommendation tools affect sales."

ML Platform Ecosystem Analysis through Business Model Map

Amazon Sagemaker is a well-established platform for ML development and engineering. Given the latest updates with Sagemaker Jump-start model (Huang, Lokanatha, Karp and Das, 2023), it has also established itself as Novel platform for experimentation with Generative AI. Especially, development of LLM and RAGs are well documented in the Caylent blogpost (Arabi, 2023), where the aspects of platform enablement of "distributed training" for generative AI has been captured.

As shown in the figure above, even for training of recommendation systems based on Generative AI and ReAct-like systems, that synergize reasoning and acting in Language Models (Yao et al., 2023), training using Sagemaker can ease the process due to the distributed nature of the platform (distributing the model or data) and providing high processing power that training Generative AI needs, with either GPUs or Amazon's customized processors.

Focusing on the Processing Power, the blogpost by Caylent notes that:

"AWS has additional options beyond GPUs like the AWS Trainium instance which is a machine learning (ML) accelerator purposely built for deep learning training of 100B+ parameter models. You can further leverage AWS's custom silicon for inference with AWS Infrentia2, which provides high performance at the lowest cost for deep learning inference."

To note, the Platform supports other frameworks and packages such as PyTorch Distributed Data Parallel (DDP).

Sagemaker from Amazon is a Platform for Machine learning with a great market traction and attraction from the ecosystem (Domo.com, 2024). Recently build support for Generative AI on Sagemaker has been announced through an AWS Cloud Development Kit (AWS CDK) as “asynchronous SageMaker JumpStart foundation model” which furthers the possibility of the platform eco-system growth (Huang, Lokanatha, Karp and Das, 2023). At the same time, the SageMaker Platform is expected to be exploited in the “Drone-as -a Service” systems, with exciting use-cases to fuel the 6th Generation of Mobile Technology Business (Besada et al., 2019).

To note, the economic potential for multi-sided Platform (MSP) for remote operation of Drones and UAVs (drone-sharing platform) over the internet have been demonstrated for providing technology solutions for remote internet-based operation and advanced autonomy. The main roles of the stakeholders in the service field have been articulated in the paper by Besada et al. (2019), but from the context of AIML, we find the articulation of the Swoop Arrow Platform (Swoop Arrow, 2023), a global Drone Deliver Service Provider as relevant – the platform uses AWS IoT Core to connect UAVs as IoT devices, further using the socket capability in Amazon API Gateway to provide two-way communication between UAVs and remote pilots. Swoop Arrow leverages computer vision technology at the edge for drones to ascertain the potential to land safely in auto-pilot mode. The technology leverages custom ML models, trained on Amazon Sagemaker and SageMaker Ground Truth data labelling services; further the company relies on Amazon SageMaker Neo for model optimization to run onboard the UAVs in a poor communication environment, using Federated Learning.

From a platform strategy management view, using Amazon Sagemaker,

SageMaker Ground Truth and Neo works cooperatively to fulfill advanced Drone-as-a-Service Operation.

The combined platform can be used as a reference, and further using Roger's Platform business model map, can illustrate, articulate and conduct value analysis of the multi-sided Platform (MSP) eco-system; and by visualizing the engagement scenarios of the eco-system partners, diverse businesses may employ novel ways to derive the full benefits using the "value-based rules of engagement".

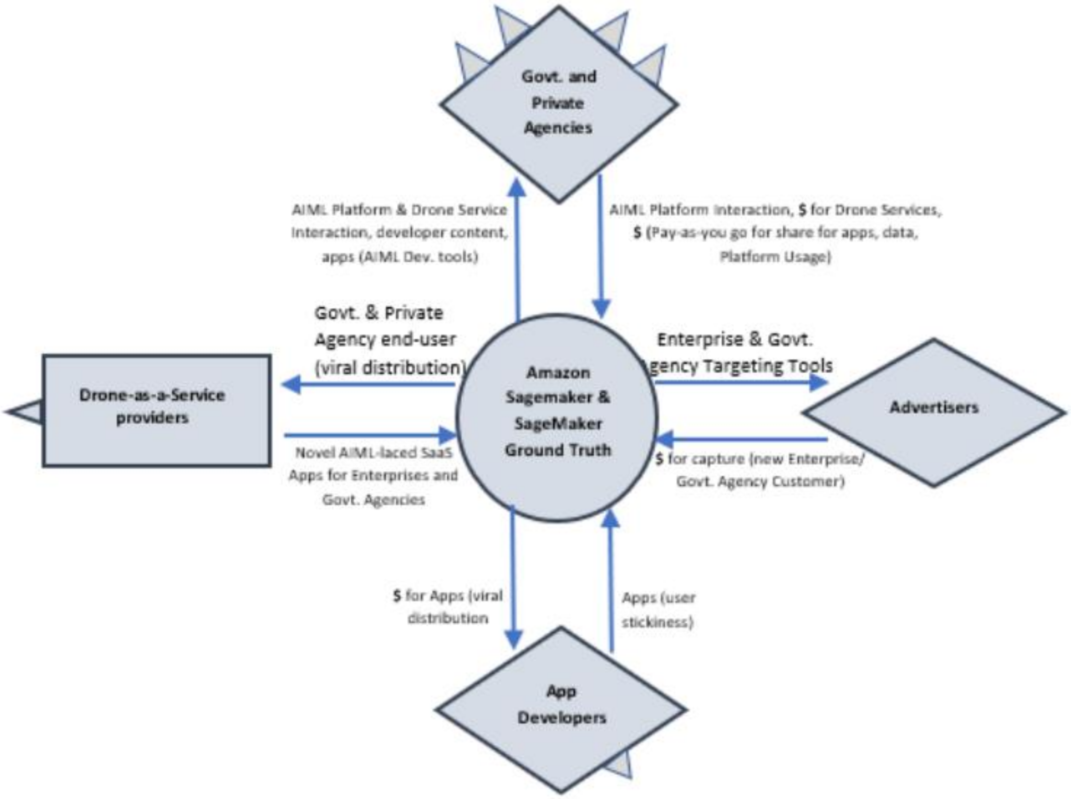

Govt. and
Private
Agencies
AIML Platform & Drone Service Interaction, developer content, apps (AIML Dev. tools)
AIML Platform Interaction, $ for Drone Services, $ (Pay-as-you go for share for apps, data, Platform Usage)
Govt. & Private Agency end-user (viral distribution)
Enterprise & Govt. gency Targeting Tools
Drone-as-a-Service providers
Amazon Sagemaker & SageMaker Ground Truth
Advertisers
Novel AIML-laced SaaS Apps for Enterprises and Govt. Agencies
$ for capture (new Enterprise/ Govt. Agency Customer)
$ for Apps (viral distribution
Apps (user stickiness)
App
Developers

Figure 5.14
AWS SageMaker, Sagemaker Ground Truth and "Drone-as-a Service"-based Platform eco-system analysis from the perspective of Platform Business Model Map. Courtesy: The Digital Transformation Playbook, David Rogers

In the above figure, illustrating the Platform Business Model Map, the Publishers, can be considered as various Drone-as-a-service Independent Software Vendor (ISV) Companies (Swoop Arrow, 2023, Aerodyne, 2023) using the Platform to deliver or

“Publish” SaaS-based Drone Services, generating value which is then consumed by the Government and Private Agencies, as “users”. Amazon regularly addresses the connection between the Publishers with their platform innovations as exemplified in the blogpost (Bakkaloglu, Kadiyala, 2023).

The “Platform users” community in the above figure, may include users like Analysts of the Government and Private Agencies utilizing the Drone-based services like Precision Agriculture, Homeland Security or Asset monitoring Services directly using the Platform on a “pay-as-you-go” model for entrepreneurial AIML-based service provisioning to other businesses; or the Machine Learning Scientists or engineers using Sagemaker for domain-specific product/service development for technology or enterprise businesses with the possibility of attractive and value-added rewards from Amazon and possibly even receive attractive cost-efficient bundled services, based on a deep B2B relationship (Amazon Sagemaker Canvas, 2024)

Amazon Sagemaker may be considered as a “self-advertising” and “self-servicing” Product amongst the AIML community in the enterprises using Amazon Marketplace to derive value. Their guidance in evident from their blogpost (Amazon Sagemaker, 2024; ‘Apps Run the World’, 2024), addressing the OEM and ISV and Customer community.

The “Advertisers” in the proposed model above, would, majorly include AWS Partner Network (APN) (Sagemaker Partners, 2024; AWS Partner Network, 2024). who would market and deliver AIML products and services in a stand-alone mode for AIML developer or user community in the SMB segment or to complement cloud-based services in big enterprises, depending on the end-customer or Enterprise’s choice. As a network effect perspective of multi-sided Platforms, the Publishers also assist in developing the eco-system by “Advertising” the usage of SageMaker platform to the users and App developers alike.

The “App developer” community, as in the figure, would exploit the stickiness of the eco-system and would further develop apps using Amazon Sagemaker tools, specific to the combined AIML-based Drone-as-a-Service Platform, generating value for themselves and the Combined Platform as a whole. Amazon would publish the latest developer tools and enhancements for the Developer Community (by their blogpost advertising the “Developer Guide” (Sagemaker Developer Guide, 2024; AWS Marketplace, 2024) to allow developers easy access to Sagemaker framework and supporting tools to develop and “*Sell algorithms and packages in the AWS Marketplace”.* as per the guidance of the documented development guidelines from Amazon Sagemaker.

The value provided and derived by each eco-system partner is noted in the figure above and thus is generic enough to serve as a win-win rule of engagement for each party engaged in the platform ecosystem partnership. This case study thus, using the framework of “Platform Business Model Map” is exemplary for any domain-specific eco-system formation analysis for any AIML multi-sided Platform engagement scenario.

- **Case Study 4 – Dynamic Strategic and Tactical partnership analysis of Generative AI Model-based businesses using Value Train Analysis Framework**

This case study considers unique analysis of business methodology via study of the Generative AI offerings of Microsoft and OpenAI. These companies are currently accepted as the industry leaders in providing the Generative AI solutions, like "ChatGPT" (OpenAI) and "AutoPilot" (Microsoft).

Reports from McKinsey's (2023) and VentureBeat (Franzen, 2023) find generative AI could add up to $4.4 trillion a year to the global economy latest research estimates that generative AI has the potential to add the equivalent of $2.6 trillion to $4.4 trillion annually across the 63 use cases they analyzed, increasing the impact of all artificial intelligence by 15 to 40 percent. This estimate would roughly double if the impact of embedding generative AI into software that is currently used for other tasks beyond those use cases are included.

McKinsey concluded that about 75 percent of the value that generative AI use cases could deliver falls across four areas: Customer operations, marketing and sales, software engineering, and Product R&D.

McKinsey aimed to understand the technology's potential to deliver value to the economy and society (value potential by function) at large will help shape critical decisions using two complementary lenses to determine where generative AI with its current capabilities could deliver the biggest value and how big that value could be, as shown in the below table.

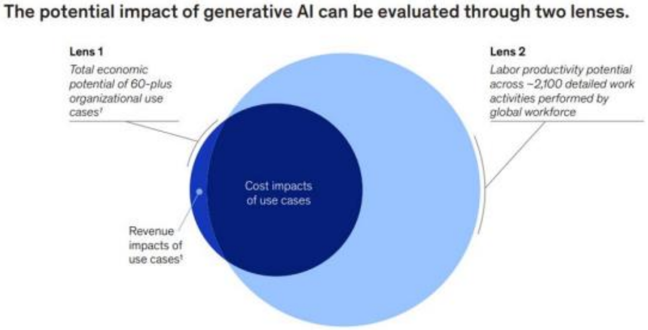

The potential impact of generative AI can be evaluated through two lenses.
Lens 1
Total economic potential of 60-plus organizational use cases[1]
Lens 2
Labor productivity potential across ~2,100 detailed work activities performed by global workforce
Cost impacts of use cases
Revenue impacts of use cases[1]

Figure 5.15
Evaluating potential impact of GenAI with 2 specific lenses. Courtesy: McKinsey (2023)

The first lens, as McKinsey explains, “scans use cases for generative AI that organizations could adopt. McKinsey defined a “use case” as a targeted application of generative AI (GenAI) to a specific business challenge, resulting in one or more measurable outcomes. Our second lens complements the first by analyzing generative AI’s potential impact on the work activities required in some 850 occupations, enabling McKinsey to estimate how the current capabilities of GenAI could affect labour productivity across all work currently done by the global workforce.

The summary to the main themes of their findings on the prospects and impacts of GenAI may be drawn as follows:

- Generative AI’s natural language capabilities increase the automation potential previous generation management automation tasks (related to collecting and processing data these types of activities). As GenAI is fundamentally engineered to do cognitive tasks, hence efficiency of automation activities and RPA activities will be boosted and impacted, resulting in biggest impact on knowledge work, particularly activities involving decision making and collaboration, which hitherto had the lowest potential for automation.
- McKinsey’s estimate of the technical potential to automate the application of expertise jumped 34 percentage points, while the potential to automate management and develop talent increased from 16 percent in 2017 to 49 percent in 2023. This is

explained by GenAI’s much superior ability to understand and use natural language for a variety of activities and tasks explaining the steep rise of automation potential. They corroborate with their research finding that around 40 percent of the activities that workers perform in the economy require at least a median level of human understanding of natural language. The eventuality that potentially many of the work activities involving communication, supervision, documentation, and interaction with people, be automated by GenAI, accelerating and hastening the transformation of work in occupations such as education and technology are already showing tell-tale signs of coming to reality.

▪ McKinsey’s findings, contrary to the labor economist of earlier generation, convey that GenAI is likely to have the most incremental impact through automating some of the activities of more-educated workers. McKinsey describes Generative AI as a “skill-biased technological change, but with a different, perhaps more granular, description of skills that are more likely to be replaced than complemented by the activities that machines can do.” GenAI’s thus will impact higher-wage knowledge workers significantly because of advances in the technical automation potential of their activities that were previously considered relatively immune from automation.

Deloitte (Deloitte AI Institute, 2023) further confronts us with the “key questions” for various personas in the respective domains led by the identified Stakeholders, on how to incorporate the latest GenAI technologies.

- **SWOT Analysis for analyzing competition landscape**

The importance of SWOT analysis has already been emphasized; however, based on the specific requirements of a AIML companies, we may sketch and draw up a general SWOT analysis template, applicable to Companies operating in Novel AIML product and service development domains using standard templates (Forbes SWOT Analysis

Template, n.d.)

- **Partner Analysis for assessing tactical dis-intermediation using Value Train**

Computational cost of testing Generative AI is very high (Hao, 2019).

As already noted, testing Generative AI LLMs need high computational power, leading to a rapid cost escalation on progressive releases of new versions, supporting more Model parameters.

For an innovative LLM developer, a win-win situation may be created by a futuristic dyadic coopetition structure (Rusko, 2012)**,** allowing rapid testing of innovation by OpenAI; and in the process, offering the test infrastructure provisioner the opportunity to harness the innovation and incorporate the same in its own products and thus create and capture early market share in the new technology space (Warren, 2023).

The above situation may lead to a coopetitive dyadic business environment and guide to create rules of engagements enabling to measure incremental successes.

So, while the “big-tech” tactical strategy could be to upgrade and vertically integrate the Value Producer’s deliverable in its own development initiative, similar to that of “CoPilot” and also train its Chip development initiative, that is the Computation Resource (AI Chip) and also, as all other Big-Tech Companies, developing the “Competitive” aspect of the relationship, a dyadic “cooperative” partnership with the Value Creator, may result in an initial stable delivery of value to its own end-user customer base. Also, this tactical innovation and intermediation by utilizing its new chip the Parent Company may “vertically integrate” the “innovation” into its offering.

Similarly, the innovative Generative AI LLM development Company, will likely demonstrate its “competing” urge to chart its own course by forging new alliance with a possible emergent AI Chip Vendor aiming to rival against the data-center chip provided Nvidia as reported by Nolan (2021), to provide “lower-cost competition to Nvidia and reduce the cost for running services like ChatGPT".

Further, future ROI of Prime Service and Infrastructure Provider is likely to get disrupted if the "value train" is perceived to be "risky" in terms of value delivery as the "Value Producer" may wish to chart its own course with a new AI Chip Vendor. Similarly, the risk of "uncertainty" of the produced "end-value" may force end-users to leave the platform.

Following an analogous situation of the latest announcement of Microsoft to incorporate the support for multiple LLMs other than OpenAI's ChatGPT allows the Platform Innovator to move away from the dependence of OpenAI and thus the risk, forming a multiparty eco-system for LLM-based value generation, as reported in Business Insider (Nolan, 2021). Also, with an upgraded powerful AI Chip catering to aid the testing of newer more powerful Generative AI models, the Parent Company may successfully "lock-in" the value Generator and thwart the possibility of eco-system disruption, nullifying the possibility of value erosion for the entire eco-system. However, in Parallel, OpenAI continues to nurture its ambition to have its own AI chip/accelerator as revealed in the blogpost by The Register (Mann, 2024).

To illustrate the above tactical "snap-shot" views of rapidly developing coopetitive business environment, any possible emergent situation may be captured by the Value-Train (VT) analysis tool to exemplify a "rapidly developing intermittent" coopetitive scenario, the like of which can be conducted in any complex and dynamic business eco-system, with various dynamic equilibrium structures at various stages of coopetitive eco-system partnership, similar to a possible rapidly emergent business situation, as depicted below, noting that the "innovative GenAI Company" may want to forge new partnerships with established mobile vendors to expand its market-share (Espósito, 2024; Axon, 2024).

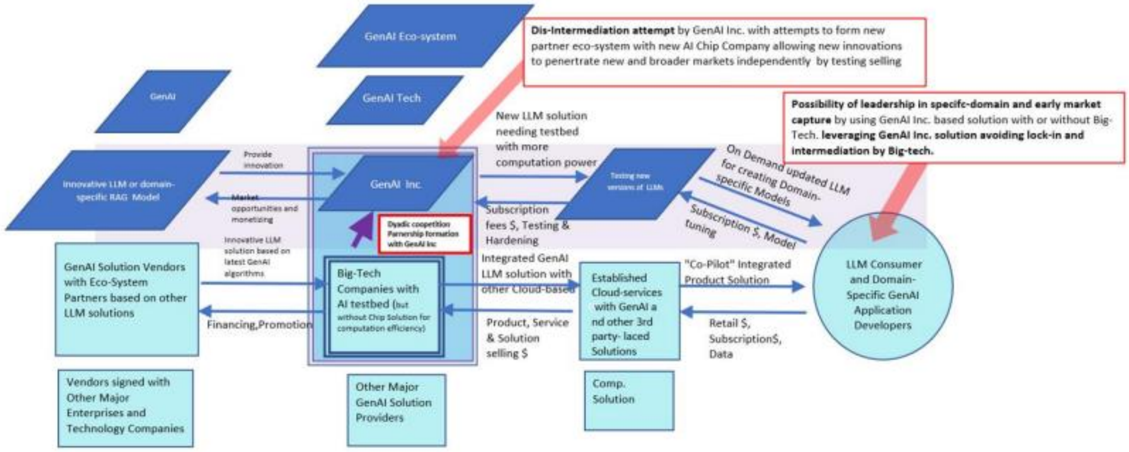

GenAI Eco-system
Dis-Intermediation attempt by GenAI Inc. with attempts to form new partner eco-system with new AI Chip Company allowing new innovations to penertrate new and broader markets independently by testing selling
GenAI
GenAI Tech
New LLM solution needing testbed with more computation power
Possibility of leadership in specifc-domain and early market capture by using GenAI Inc. based solution with or without Big-Tech. leveraging GenAI Inc. solution avoiding lock-in and intermediation by Big-tech.
Innovative LLM or domain-specific RAG Model
Provide innovation
Market opportunities and monetizing
GenAI Inc.
Testing new versions of LLMs
On Demand updated LLM for creating Domain-specific Models
Subscription $, Model tuning
Subscription fees $, Testing & Hardening
Dyadic coopetition Parnership formation with GenAI Inc
Innovative LLM solution based on latest GenAI algorithms
GenAI Solution Vendors with Eco-System Partners based on other LLM solutions
Financing,Promotion
Big-Tech Companies with AI testbed (but without Chip Solution for computation efficiency)
Integrated GenAI LLM solution with other Cloud-based
Product, Service & Solution selling $
Established Cloud-services with GenAI a nd other 3rd party- laced Solutions
"Co-Pilot" Integrated Product Solution
Retail $, Subscription$, Data
LLM Consumer and Domain-Specific GenAI Application Developers
Vendors signed with Other Major Enterprises and Technology Companies
Other Major GenAI Solution Providers
Comp. Solution

Figure 5.16
Value Train Analysis. Courtesy: The Digital Transformation Playbook, David Rogers

The Big-Tech companies with cloud-based AI test-beds have an upper-hand in negotiation with the new GenAI Inc.-like companies that are innovating LLM models and solutions. Due to the huge computational requirements to test the LLM solutions, necessitates the formation of tight dyadic coopetition model-based ecosystem.

However, that may not hinder the innovative GenAI companies to forge new Chip-partnerships (Shilov, 2024) and develop their own testbeds to march towards leadership roles in search of huge profits. Further, the GenAI Entrepreneur Company may even mull changing from non-profit to for-profit Organization Structure to garner more economic and financial benefits. This complex tension-filled rapid innovation playground is well tactically analyzed by the above VT Model, where intermediation/dis-intermediation attempts by the coopetitors to generate value and deliver them to the "last-mile" clients progresses, altering their organic/inorganic development strategies based on their negotiations with competing business houses in adjacent domains.

5.3 Summary of Findings

From the deep research of the industry trends and technological disruptions that are shaping the Transformation of both Enterprise and Technology companies, the findings from the perspective of management of Novel AIML Technologies (NAIMLTs) and their impacts can be summarized as below:

1. Novel AIML Technologies have broadly found their usage in the niche distributed cloud operating models, adopting distributed Hybrid cloud and cloud-native modes of operations, collaborated with role of opens-source SAAS applications; further Coopetition and Co-creation models form significant tools for strategic and tactical analysis of these novel AIL tools and techniques that are being adopted in the industries.

2. Frameworks of BDI and DAI frameworks form the two bedrocks based on which all next generation Novel AIML Algorithms, Models and operation are being innovated, adopted and deployed.

3. Deep Learning Frameworks used for Reinforcement and Transfer Learning, Large Language Models used in ChatGPT, BARD etc. are expected to revolutionize the general mass-adoption of AIML in the digitally-operating industries.

4. There are enough software frameworks and tools available for innovators to succeed in creating innovative solutions for the various vertical domains based on the novel AIML Techniques; and unique business models can be strategized based on SWOT, BMC, Platform Model Map and Value Train analysis etc. to strategize and tactically analyze the next steps of innovation, with nimble and agile product road-maps.

5. The possibility of value erosion for the innovative small companies, whose business models and data can be compromised based on General Prompt Engineering tools used by companies deploying LLMs which voiced slightly differently by Meta’s Chief Scientist, as penned by Forbes as “Generative AI Sucks” and calling for

a more “Objective-Driven AI” (Marr, 2024).

6. SWOT Analysis, BMC and VT frameworks can be used successfully to assess new product launch success that heavily integrate AIML tools, creating and business model that can generate profits and creating a successful value delivery model for the end-user and Partner Vendors respectively. Generative AI has shown extraordinary promise in the last few years, in the areas of Marketing Chatbots and Research and Product Ideation fields and other creative and artistic domains. They are actively under PoC stages in Enterprises, aiming to enhance productivity with LLMs, and lately with SLMs and RAG-based Generative AI applications (Superannotate.com, 2024).

7. For any new AIML-related innovations with Generative AI, for initial planning, the adoption of “Accelerator Business Model” Framework (Bagnoli et al. (2020); Biloslavo, Bagnoli and Edgar et al. (2018) is recommended for its rigor, allowing detailed analysis of impacting parameters, thus establishing a well-planned approach for fast innovations and more importantly, prompting an equivalent consideration of socio-economic impacts of any new technology for which societal and economic risks have already been highlighted.

8. With respect to the latest hype on adoption of prompt engineering for “Teaching and Training” the LLMs, focusing on educational research arena, Dr. An (pp 81, 2023) however advises to exercise caution and not blindly trust the results from “prompting”, for undertaking Research-related work at the current stage of Gen AI development. He finds that due to its probabilistic nature of decision making that the technology adopts, “for the same title and abstract, ChatGPT’s response may not be same if asked twice”. Also, very recently Microsoft has updated its service agreement with a view to warn the everyone not to take its current state of Generative AI services seriously

(Claburn, 2014).

9. Further, with respect to the ethical processes pertaining Academics and research, Eke (2023) has expressed concerns on Academic Integrity due to Generative AI, stating:

"The general fear is that students as well as researchers can start outsourcing their writing to ChatGPT."

Eke also shares the concerns of the general administration, that "a harmonized and responsible way of acknowledging the use of ChatGPT is yet to be established" and some confirmed views of non-acceptability of ChatGPT-aided research by some reputed journals and publishers:

"However, both Nature (Nature, 2023) and Science (Thorp, 2023) journals have made their stance clear that no LLM can be accepted as a credited author in their journals. The current lack of guidance for users on how to acknowledge the use of ChatGPT raises a lot of concerns."

Further, Capilano University (2024), acknowledging the concerns raised by Eke (2023) and others, stated:

"The widespread application of Generative AI in academic contexts has caused waves of concern about its potential for misuse and the serious challenge it poses to academic integrity", further stating:

"In response, educational institutions are making efforts to uphold academic integrity while promoting safe and responsible use of Generative AI".

The Institution also went on to release a string of recommendations and guidelines for the instructors, to assuage the threats.

10. In business settings, there is a frenzy in the Generative AI eco-system formation as demonstrated by Google and Nvidia new announcement to expand

partnership (Buchanan, 2024) and also by Intel’s organic move to create its latest offering of AI accelerator, Gaudi 3 (Alcorn, 2024) and AI Chip (Leswing, 2024). All these frenzies show that market is yet to settle down to market-winning competing eco-systems, probably constrained by the “train-ability” of the LLMs in the near future and this tempered view has been vented by O'Brien (2024) in Washington Times where he warns that “AI ‘gold rush’ for chatbot training data could run out of human written text”. However, adoption of clean Nuclear Power facilities by the Hyper-scalers (Amazon, Google and Microsoft) to counter the energy costs of Data Centers needed for astronomical computation power needed for Generative AI (Bathgate, 2024), shows their commitment to take Generative AI to take to its next level, conclusively.

11. From regulatory framework development perspective, Volpicelli (2023) noted that “ChatGPT broke the EU plan to regulate AI”; emphasizing the challenges that the Generative AI content brings in for the existing and developing regulatory frameworks like GDPR and XAI respectively; the nuances of the European Union deal was further critically analyzed by Clifford (2023).

12. Research of novel and automated AIML algorithms, online/incremental and federated learning with streaming data for real-time platforms & processes common to key areas of next generation technology, like 6G wireless and Drone Technology; and allied businesses and assist in synergizing their common value propositions.

13. Frameworks for winning ecosystem partnership formation, for businesses adopting AIML as the value for transformation using Value Train Analysis and Lean Business Models & Platforms have been introduced based on latest AI technologies.

14. Strategies for “build or buy” decisions for adoption AIML solutions should be weighed against the option of eco-system partnership formation and may be adopted, mixed and matched and tailored based on specific business scenarios.

15. Recently reasoning-based LLM has made its mark with STaR (Zelikman, 2022) from Google Research Team, followed by variant from OpenAI, with its release of Latest LLM Multi-Modal Model, GPT-4o, using “test-time compute” technique for abstract reasoning (OpenAI, 2024). The technique of the "Self-Taught Reasoner" (STaR), working in a loop to generate rationales to answer many questions, prompted with a few rationale examples, paves the way to providing reasoning-based AI robots working as cooperative agents. This step is a mile-stone towards shaping up businesses with “Agentic LLM-based” smart domain-specific robots, trained in a responsible manner with XAI for cooperative operations. Thus, this promises to herald, with “self-driving AI”, an era of implementation of serious Artificial General Intelligence. In the words of AI Thought Leader Aschenbrenner (2024):

“The AGI race has begun. We are building machines that can think and reason. By 2025/26, these machines will outpace college graduates. By the end of the decade, they will be smarter than you or I; we will have superintelligence, in the true sense of the word.”

5.4 Conclusion

It is imperative to state that AI and ML has influenced the Cloud-based transformation of modern B2B and B2C businesses and technologies alike. The all-encompassing nature of AIML adoption for autonomic functioning can thus be considered to be the “promised key” for a successful march towards next-gen mobile technologies and herald the much-awaited next-gen industrial revolution, “Industry 5.0”. However, as noted in our Research, socio-economic benefits have to be analyzed by Global Statutory Bodies to guide and steer the adoption of these novel and exciting technologies in business and technology with proper risk analysis and installing adequate “guard-rails”, to ensure the “general well-being” of the society at large.

CHAPTER VI:

DISCUSSION

6.1 Discussion of Results

The research has thus been conclusively conducted with systematic meta-analyses through technology deep-dives, followed by trend analysis of current uptake of these new AIML techniques and further, through impact analysis via business case studies. Adoption trends of novel AIML technologies revealed that it is highly likely that the novel technologies, will continue to define the state of the distributed cloud-based technologies and enterprise businesses.

6.2 Discussion of Research Question One

Various novel AIML technologies and business methodologies have been researched and analyzed based on credible industry reports on current usage trends of novel AIML techniques, allowing to align the business requirements of achieving faster innovations, heightened productivity and cost-effectiveness in both current and futuristic next-gen business settings.

The novel AI Techniques have been studied, collaboratively driving and leading to Generative AI boom, creating new job roles and displacing older ones. The latest trend marches towards embracing “Agentic AI”, made viral through the open-source tool, “AgentPy” (Foramitti, 2021), with its future promises in aiding enterprises to build sophisticated “Domain-specific” AI Agents to solve industry-specific business problems in RAG-based private data centers, ensuring data security and privacy. AI Agent workflows also show promises for integration with the novel LLM models to bring in additional benefits of generalization in agentic communications, as described in the new “Autogen” tool (Gannon, 2024).

The next generation technologies like NWDAF and Drones, studied in this thesis, are revolutionizing the established industries with their adoption in 5G/6G technologies and enterprise/Government businesses respectively – all powered by the novel AI techniques, discussed as a response to the above research question.

Finally, in the "Results" section, we assessed the trends and impacts of Novel AIML techniques utilizing Industry-supported AI report (HAI, 2023; HAI, 2024). We conducted analysis of initial trends of adoption of these technologies in industries and social communities, utilizing data tables for 1. "Themes for AI Mentions in Fortune 500 Earnings Calls", 2. industry trend of adoption of "Open Access Foundation AI models" and 3. Trend analysis of rapid implementation of AI in robotics using "Collaborative next-gen robots".

These studies, thus allowed us to analyze early, yet noticeable impacts on AIML adoption between the years 2018 and 2023, raising our expectation that they will continue to re-define the industries in the future, provided the innovations adhere to the guidelines prescribed by the regulatory bodies.

To note, post-2022, research and experimentation of adoption of Generative AI in enterprise and technology businesses have seen a meteoric rise in PoC and experimentation; and Clark, in his blogpost in The Register (2024) notes that Gartner expects full-fledged adoption of Generative AI in office environments around 2026, furthering the impact of AIML adoption in Technology and Business.

6.2 Discussion of Research Question Two

To corroborate the findings of the AI research for business and technology impacts, specific focused Industries have also been researched via specific case studies that appear in the "Results" section, show tremendous potential for future application for analysis while navigating fast-changing AIML impacts in business and technology.

In “Case Study 1”, multi-party coopetition business aspects, prevalent in the highly complex business solution delivery scenarios, is aptly demonstrated in innovative Multi-access Edge Communication business for 5G/6G, which need multiple niche technology solutions to be delivered by diverse eco-system vendors. The co-opetition stems from the fact that each party of the eco-system is free to be co-operative for some client projects while also competing with them, for forging alliances with external clients, thus widening the ambit of their business opportunity funnel.

This strategy however, needs high level of governance by the lead-partner for the eco-system to bring timely business value to external Clients. “Case study One” aptly analyses and illustrates, using visual tool, industry collaboration efforts that are currently under way, for Multi-Access Edge Computing using 5G Edge Routers and aerial UAVs/Drones.

In “Case Study 2”, dyadic (or two-party) coopetition in businesses is explored, typically prevalent between established vendors with high brand values, who complement each other’s offerings, to provide Industry leading service offerings and solutions. The technologies in these co-opetative areas are normally highly regulated by Industry standards and each of the vendors are expected to be lead vendors in their own solution space; however, they only jointly bring in the value to a very fast-developing technology area, where their joint solutions and road-maps can out-smart entrepreneurial ventures. Further, the dyadic coopetitive vendors assist each other to develop competitive market expansion opportunities, integrating their individual solutions. The dyadic co-opetition case has been studied in the business case, using the NWDAF solution requirements, where established vendors coopete to establish the new 5G-based Platform for Network Data Analytics.

In “Case Study 3”, we explored an innovative technique to visually analyze Platform-based businesses with “Platform Model Map”, to establish the network effects of Multi-Sided Platforms, governing the various players in the Platform business eco-system. We showed, with the use of AIML Platform and the “Drone as a Service business” as a producer, how a thriving business may be conceived, only if the values, generated and consumed are opportunistically balanced, each partner in the eco-system bringing in the value of network effects, to generate business growth and increasing value proposition from the marketplace. This points to the fact that a high level of Platform governance is needed to ensure that values thus created, are rightly disseminated to all eco-system partners for viral adoption and growth of the Platform business.

“Case Study 4” show-cased the utilization of novel business techniques, called “Value-Train Analysis” to aid dynamic strategic and tactical plays encountered in rapidly changing entrepreneurial business environments. While the scope of long-term value and profit generation recur as a theme of Alliance and Partnership formation and management (Lewin,1990), the dynamic of nature of current generation AI-laced businesses need rapid variation to the modalities of engagement in managing already established partnerships and alliances for undertaking rapid decisions on future scopes of value creation, management and even exit. Accordingly, a Value-Train Analysis provides a visualization analysis technique to assess business tensions in a dyadic coopetitive business environment, where an established partner forges alliance with an entrepreneurial new entrant to offer enhanced business value. Ideally, the technique visualizes the scenarios periodically, to assess the dynamic business engagement tactically, in an already competitive market environment. The assessment conducted periodically with the Value-Train (VT) Analysis tool, allows the established incumbents to re-strategize and re-formulate its own alliance and in-house development plans based on the counter-alliance and partnership formation of the entrepreneurial company.

The dyadic partnership tension and possible avenues to get relief from those, are thus made evident through the application of VT analysis in the “Case Study 4”, where the Prime Generative AI Test Infrastructure Provider is forced to explore its own inhouse development strategy and strengthen its own offering “Copilot” technology to the end-users, allowing it to be less reliant on the Novel LLM vendor’s “GPT Technology” by “developing a cheaper, less powerful AI model” as evidenced by the report in the renowned blogposts (Stradling, 2023; Nolan, 2023).

The above case studies thus allowed to explore novel business techniques to harness the powers of fast-developing AI-laced businesses to avoid risks, reduce cost of adoption by end users by forging winning partnerships and alliances and most importantly governing them ethically to generate sustainable profits.

The thesis thus establishes novel techniques to handle novel AIML technologies, application of which may ignite specific domain analysis and research in both academia and industry alike, and provide an early and timely analysis of the opportunity and threat landscapes and undertake appropriate business decisions in order to etch out winning and sustainable eco-systems for Platforms or Products. Importantly, a futuristic view of the journey of AI to AGI has been studied, aligning to which, businesses and technologies will be able to derive and generate economic value and reduce economic costs through the proper and timely use of business tools explored and illustrated in this Thesis.

CHAPTER VII:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

7.1 Summary

Novel AIML technologies have established themselves as a harbinger of value generation for next generation businesses and technologies. Henceforth, no new technology or business can be conceived without the careful adoption of AIML technologies, aiding them to be innovative, efficient and cost-effective.

The possible negative impacts or “collateral damages” on society, however can be likened to the recurring themes of “creative destruction” (Pfarrer and Smith, 2015), that accompanies each new generational shift in industrialization; whenever innovative tools, techniques and processes are adopted for enhancing business efficiency.

7.2 Implications

The implications of the impact of the novel AIML techniques are expected to be far-reaching and probably heralds the beginning of the “conscious machine” age where Man and Machine will coopete in ways hitherto unknown.

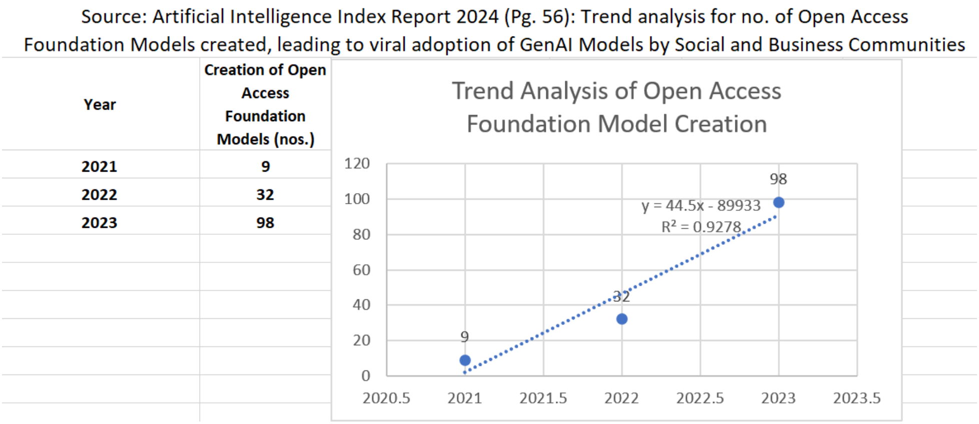
7.3 Recommendations for Future Research

With the concept of lifelong-learning, as noted by Gartner (Wiles, 2023) in its futuristic post on what lies beyond ChatGPT and supported by the concept of adaptive ML (Takyar, 2023), we have the promises of an exciting future ahead for AIML-based research for future adoption in industries and hence is recommended as a direction for future research.

7.4 Conclusion

The tempered and rounded view of the research undertaken suggests that while where the future is predictably bright for AIML-laced businesses where profit-generation and operating cost will find a positive balance while at the same time, dwindling human value and “not so positive” work-force impacts could cause an upheaval in the society if these technologies are not well regulated for the greater economic good.

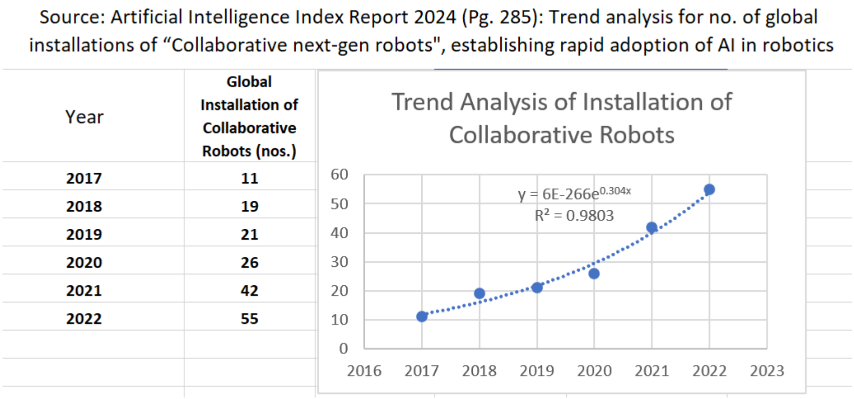
APPENDIX A – TREND ANALYSIS CHARTS



Source: Artificial Intelligence Index Report 2024 (Pg. 56): Trend analysis for no. of Open Access Foundation Models created, leading to viral adoption of GenAI Models by Social and Business Communities

Year	Creation of Open Access Foundation Models (nos.)
2021	9
2022	32
2023	98

Figure A.1
Trend Analysis for creation of Open Access Foundation Models



Source: Artificial Intelligence Index Report 2024 (Pg. 285): Trend analysis for no. of global installations of “Collaborative next-gen robots", establishing rapid adoption of AI in robotics

Year	Global Installation of Collaborative Robots (nos.)
2017	11
2018	19
2019	21
2020	26
2021	42
2022	55

Trend Analysis of Installation of Collaborative Robots
60
50
40
30
20
10
0
2016
2017
2018
2019
2020
2021
2022
2023
$y = 6E\text{-}266e^{0.304x}$
$R^2 = 0.9803$

Figure A.2
Trend Analysis for number of Collaborative Robots installed Globally

REFERENCES

Aarikka-Stenroos, L., Jaakkola, E. (2012) 'Value co-creation in knowledge intensive business services: A dyadic perspective on the joint problem solving process', [online]. Available at: https://www.researchgate.net/publication/220043055

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C. (2016) 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems', *(Preliminary White Paper, November 9, 2015) Google Research Team* [online]. Available at: https://arxiv.org/abs/1603.04467

Abir, A.B.S, Chowdhury, M. Z (2023) Digital Twin-based Aerial Mobile Edge Computing System for Next Generation 6G Networks Conference Paper. DOI: 10.1109/EICT61409.2023.10427658
[online]. Available at: https://www.researchgate.net/publication/378194844

Adair, B. (2024) 'Best Visual Analytics Tools 2024', *SelectHub* [online]. Available at: https://www.selecthub.com/business-intelligene/visual-analytics-tools/ (Accessed: 5 January 2024).

Adamopoulou E, Moussiades L. (2020) 'Chatbots: History, technology, and applications' [online]. Available at: https://doi.org/10.1016/j.mlwa.2020.100006

'Adiabatic Quantum Computation' (2019) Wikipedia Contributors [online]. Available at: https://en.wikipedia.org/wiki/Adiabatic_quantum_computation (Accessed 5 January 2024).

Aerodyne (2023) 'Aerodyne and Amazon collaborate to build drone data solutions', AWS:ReInvent [online]. Available at: https://www.digitalnewsasia.com/business/aerodyne-teams-aws-solve-complex-industrial-issues-drone-data

Agarwal, Anurag & Jaiswal, Deepak. (2012) 'When Machine Learning meets AI and Game Theory' *Stanford University Project* [online]. Available at: https://cs229.stanford.edu/proj2012/AgrawalJaiswal-WhenMachineLearningMeetsAIandGameTheory.pdf

Aggarwal, C.C. (2018) *Neural Networks and Deep Learning – A Textbook:* Springer International Publishing AG.

Akraino (2020) ‘Cloud Interfacing at the Telco 5G Edge’ [online]. Available at: *https://www.lfedge.org/wp-content/uploads/2020/09/Akraino_Whitepaper2.pdf*

Akraino (2023) LF Akraino [online]. Available at: *https://www.lfedge.org/projects/akraino/*

Alcorn, P. (2024). ‘Intel details Gaudi 3 at Vision 2024 — new AI accelerator sampling to partners now, volume production in Q3’, Tom’s Hardware [online]. Available at: https://www.tomshardware.com/pc-components/cpus/intel-details-guadi-3-at-vision-2024-new-ai-accelerator-sampling-to-partners-now-volume-production-in-q3 [Accessed 13 Apr. 2024].

Alpaydin, E. (2014). *Introduction to Machine Learning:*3rd edn. Prentice Hall India.

Amazon Sagemaker Canvas (2024) ‘Build highly accurate ML Models using a virtual interface, no code required’, *Amazon Web Services* [online]. Available at: https://aws.amazon.com/sagemaker/canvas/ (Accessed 5 Jan. 2024).

AWS Marketplace (2024) ‘Sell algorithms and packages in the AWS Marketplace: Developer Guide’, *Amazon Sagemaker* [online]. Available at: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-marketplace.html (Accessed 5 Jan. 2024).

AWS Partner Network (2024) ‘AWS Partner Network (APN)’, *Amazon Sagemaker* [online]. Available at: https://aws.amazon.com/blogs/apn/ (Accessed 5 Jan. 2024).

An, R. (2023) *Supercharge your Research Productivity with ChatGPT – A Practical Guide*: Amazon.com (Independently Published and distributed)

Analysys Mason (2021) ‘Near-real-time RIC: enabling AI/ML-driven extreme automation and granular control of Open RAN’ [online]. Available at: https://cdn.brandfolder.io/D8DI15S7/at/968x3jjhmp48q7wwkrbfq5qz/Analysys-Mason_RIC-perspective_final_20210601.pdf

Anthony, S.D. (2009) ‘Four Lessons from Y-Combinator’s Fresh Approach to Innovation*’*, *Harvard Business Review* [online]. Available at: https://hbr.org/2009/06/four-lessons-from-ycombinators (Accessed 5 Jan. 2024)

Arabi, A. (2023) ‘LLMOps Reference Architecture - MLOps for Large Language Models on AWS’, *Caylent*. October 12, 2023 [online] Available at: https://caylent.com/blog/ml-ops-for-large-language-models (Accessed: 5 January 2024).

Arzo, S.T., Scotece, D., Bassoli, R., Granelli, F., Foschini F., and Fitzek, F.H.P. (2022) ‘A New Agent-Based Intelligent Network Architecture’, Available at: DOI: arXiv:2211.01924v1 [cs.NI]

Aschenbrenner, L. ‘SITUATIONAL AWARENESS – The Decade Ahead’ [online]. Available at: https:situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

AutoML (2018) ‘Automated Machine Learning’ [online]. Available at: https://www.fast.ai/2018/07/23/auto-ml-3/

Axon, S. (2024) ‘Apple and OpenAI have signed a deal to partner on AI’, *Report - Ars Technica* [online]. Available at: https://arstechnica.com/gadgets/2024/05/report-apple-and-openai-have-signed-a-deal-to-partner-on-ai/
[Accessed 27 Jun. 2024]

Bagnoli, C., Massaro, M., Ruzza, D. and Toniolo, K. (2020) ‘Business Models for Accelerators: A Structured Literature Review. Journal of Business Models’, Vol. 8, No. 2, pp. 1-21 [online]. Available at: https://core.ac.uk/download/pdf/327193447.pdf

Bakkaloglu, M., Kadiyala, R. (2023) ‘Integrate SAAS platforms with Amazon Sagemaker to enable ML- powered applications’, *AWS Machine Learning Blog* [online]. Available at: https://aws.amazon.com/blogs/machine-learning/integrate-saas-platforms-with-amazon-sagemaker-to-enable--ml-powered-applications/ (Accessed 5 Jan. 2024).

Bär, S., Bakakeu, J., Meyes, R., Meisen, T. (2019) ‘Multi-Agent Reinforcement Learning for Job Shop Scheduling in Flexible Manufacturing Systems’ [online]. Available at: https://www.researchgate.net/publication/336253552

Bathgate, R. (2024) ‘Hyperscalers go nuclear’, itpro.com [online]. Available at: https://www.itpro.com/infrastructure/data-centres/hyperscalers-go-nuclear

Benzaghta, M. A., Elwalda, A., Mousa, M. M., Erkan, I., & Rahman, M. (2021) ‘SWOT analysis applications: An integrative literature review’, *Journal of Global Business Insights*, 6(1), 55-73. https://www.doi.org/10.5038/2640-6489.6.1.1148

Berman, B., Evans, J. R., Chatterjee, P., Srivastava, R. (2018) *RETAIL MANAGEMENT – A STRATEGIC APROACH*: 13th edition, Pearson India.

Besada, J.A., Bernardos, A.M., Bergesio, L., Vaquero, D., Campaña, I., Casar. J.R (2019) ‘Drones-as-a-service: A management architecture to provide mission planning, resource brokerage and operation support for fleets of drones’ [online]. Available at: http://sig-iss.work/percomworkshops2019/papers/p931-besada.pdf

Bhagavatula, S. et al. (2017) ‘Accelerator Expertise: Understanding the Intermediary Role of Accelerators in the Development of the Bangalore Entrepreneurial Ecosystem’, *Article in Strategic Entrepreneurship Journal* [online]. Available at: https://www.researchgate.net/publication/320917394

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. and Lloyd, S. (2018) ‘Quantum Machine Learning’, arXiv:1611.09347v2

Bigelow, Stephen J. (Sep, 2021) ‘Multi-cloud vs. hybrid cloud similarities and differences’, *Techtarget.com* [online]. Available at: https://www.techtarget.com/searchcloudcomputing/feature/Multi-cloud-vs-hybrid-cloud-and-how-to-know-the-difference downloaded on Jan 2022

Bishai, A. (2018) ‘How to Optimize cost savings in AWS Marketplace’, *Pay-as-you-go, AWS* [online]. Available at: https://aws.amazon.com/blogs/awsmarketplace/tag/pay-as-you-go/

Bliemel, M.J., Flores, R., Klerk, S.D., Miles, M.P. (2016) ‘The role and performance of accelerators in the Australian startup ecosystem Technical’, *Report* · February, 2016 [online]. Available at: https://www.researchgate.net/publication/305307540

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S. et al. (2022) ‘On the Opportunities and Risks of Foundation Models’, *Center for Research on Foundation Models (CRFM) Stanford Institute for Human-Centered Artificial Intelligence (HAI) Stanford University* [online]. Available at: https://arxiv.org/abs/2108.07258

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V. et al. (2019) ‘*TOWARDS FEDERATED LEARNING AT SCALE: SYSTEM DESIGN’,* arXiv:1902.01046v2

Botchkarev, Alexei. (2018) ‘Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio’, *Article in SSRN Electronic Journal,* DOI: 10.2139/ssrn.3177507

Boyali, A., Hashimoto, N. and Keihanna. (2019) ‘Multi-Agent Reinforcement Learning for Autonomous on Demand Vehicles’, DOI: 10.1109/IVS.2019.8813876 [online]. Available at: https://www.researchgate.net/publication/332446506

Borg (2015) ‘The predecessor to Kubernetes’, *Kubernetes.io* [online]. Available at: https://kubernetes.io/blog/2015/04/borg-predecessor-to-kubernetes/

Brandenburger, A. and Nalebuff, B. (2021) ‘The Rules of Co-opetition’, *Harvard Business Review* [online]. Available at: https://hbr.org/2021/01/the-rules-of-co-opetition

(Accessed 10 Jan. 2024).

Broughton, M., Verdon, G., McCourt, T., Martinez, A.J., Yoo, J.H., Isakov, S.V. et al. (2020) ‘TensorFlow Quantum: A Software Framework for Quantum Machine Learning’ [online]. Available at: https://arxiv.org/abs/2003.02989

Bruijl, G. (2018) ‘The Relevance of Porter's Five Forces in Today's Innovative and Changing Business Environment’, *Article in SSRN Electronic Journal* · January 2018 DOI: 10.2139/ssrn.3192207 [online]. Available at: https://www.researchgate.net/publication/326026986

Buchanan, N. (2024) ‘Nvidia, Google Expand Partnership with Nvidia Blackwell Coming to Google Cloud in 2025’, *Investopedia* [online]. Available at: https://www.investopedia.com/nvidia-google-expand-partnership-with-nvidia-blackwell-coming-to-google-cloud-in-2025-8628792 [Accessed 13 Apr. 2024].

Burke, Stel, A-V. and Thurik, R. (2016) ‘Testing the Validity of Blue Ocean Strategy versus Competitive Strategy: An Analysis of the Retail Industry’, *International Review of Entrepreneurship, Article* #1529, 14(2): pp. 123-146. *Senate Hall Academic Publishing* [online]. Available at: https://www.researchgate.net/publication/254803904

Campesato, Oswald. 2023. *Transformer, BERT, and GPT: Including ChatGPT and Prompt Engineering*. MERCURY LEARNING AND INFORMATION. Amazon Kindle.

Canese, L., Cardarill, G. C., Nunzio, L. D., Fazzolari, R., Giardino, D., Re, M., Spanò, S. (2021) ‘Multi-Agent Reinforcement Learning: A Review of Challenges and Applications*’, ppl. Sci*. 2021, 11, 4948 [online]. Available at: https://dx.doi.org/10.3390/app11114948

Capilano University. (2024) “Generative AI and Academic Integrity” [online]. Available at: https://cte.capilanou.ca/wp-content/uploads/sites/19/2024/09/Generative-AI-and-Academic-Integrity.pdf (Accessed 23 Sep. 2024)

Casey, M. (2023) ‘GenAI most impactful tech of the decade’, *Gartner AI Hype Cycle* [online]. Available at: https://snorkel.ai/genai-most-impactful-tech-of-the-decade-gartner-ai-hype-cycle-2023/

Chakraborty, S., Sayan, D., Syamal, P. (2017) ‘A Brief Study to Cloud*’* [online]. Available at: http://oru.diva-portal.org/smash/get/diva2:1413267/FULLTEXT01.pdf

Choudary, S. P. (2021) *PLATFORM SCALE, FOR A POST-PANDEMIC WORLD*: Penguin Random House India Pvt. Ltd.

Chen, Y., Zhao, Y., Hi, X., Xu, Z. (2023) Resource allocation method for Mobility-Aware and Multi-UAV-Assisted mobile edge computing systems with energy harvesting. IET Communication, WILEY.
DOI: 10.1049/cmu2.12596

Clifford, C. (2023) ‘THE EU’S ARTIFICIAL INTELLIGENCE ACT: WHAT DO WE KNOW ABOUT THE CRITICAL POLITICAL DEAL’, *Briefing* [online]. Available at: https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2023/12/the-eus-ai-act-what-do-we-know-about-the-critical-political-deal.pdf

Chandiramani, K., Garg, D., Maheswari N. (2019) ‘*Performance Analysis of Distributed and Federated Learning Models on Private Data’, Published by Elsevier B.V.*, DOI 10.1016/j.procs.2020.01.039 [online]. Available at:
https://www.sciencedirect.com/science/article/pii/S1877050920300478?via%3Dihub
https://doi.org/10.1016/j.procs.2020.01.039

Chen, J., Zhang, A., Shi, X., Li, M., Smola, A., Yang, D (2023) ‘PARAMETER-EFFICIENT FINE-TUNING DESIGN SPACES’, [online]. Available at: https://arxiv.org/abs/2301.01821

Chen, Y., Xie, Y., Song, S., Chen, F., Tang, T. (2020) ‘A Survey of Accelerator Architectures for Deep Neural Networks’, *published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company* [online]. Available at: https://doi.org/10.1016/j.eng.2020.01.007

Chen, Y-W., Song, Q., Hu, X. (2019) ‘Techniques for Automated Machine Learning] [online]. Available at:
https://kdd.org/exploration_files/7._CR._27._Techniques_for_Automated_Machine_Learning-2.pdf

Cheung, Bryan. (201x) ‘What is Digital Experience Platform’, Liferay Inc. [online]. Available at: https://www.liferay.com/resources/l/digital-experience-platform

Choi, RY, Coyner, AS, Kalpathy-Cramer J, Chiang, MF., Campbell JP. (2020) ‘Introduction to machine learning, neural networks, and deep learning’, *Trans Vis Sci Tech.* 2020;9(2):14, https://doi.org/10.1167/tvst.9.2.14

Chronopoulou, A., Baziotis, C., Potamianos, A. (2019) ‘An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models’, *Association for Computational Linguistics. Proceedings of NAACL-HLT* 2019, pages 2089–2095 [online]. Available at: https://arxiv.org/abs/1902.10547

Clark, L. (2024) ‘Gartner mages: Payback from office AI is expected in around two years’, *The Register* [online]. Available at: https://www.theregister.com/2024/08/15/gartner_see_payback_from_office/?td=keepredading

Claburn, T. (2014) ‘Microsoft tweaks fine print to warn everyone not to take its AI seriously’, *The Register.com* [online]. Available at: https://www.therergister.com/2024/08/14/microsoft_services_update_warns

‘CNCF’ (2023). Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/Cloud_Native_Computing_Foundation (Accessed 5 Jan. 2024)

cncf.io. (2023) ‘Kubernetes Project Journey Report’, *Cloud Native Computing Foundation* [online]. Available at: https://www.cncf.io/reports/kubernetes-project-journey-report/ (Accessed: 5 January 2024)

Cohen, R. B. (2020) ‘5G Private Networks and Changes in Value Chains through Accelerating Data Analysis and Providing an Opportunity to Create New Services’, *Preprint. Economic Strategy Institute*. DOI: 10.13140/RG.2.2.15313.76640
https://doi.org/10.13140/RG.2.2.15313.76640
https://www.researchgate.net/publication/343481396

Collier, R. W., O’Neill, E., Lillis, D., O’Hare. G. M. P. (2019) ‘MAMS: Multi-Agent MicroServices’, *ACM ISBN* 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316509

Cong, I., Choi, S. and Lukin, M.D. (2019) ‘Quantum Convolutional Neural Networks’, arXiv:1810.03787v2

Cornell University. (2019) ‘Machine Learning Accelerators’ [online]. Available at: http://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture25.pdf

Cotton, R. (Apr 2022) ‘Machine Learning Cheat Sheet’ [online]. Available at: https://s3.amazonaws.com/assets.datacamp.com/email/other/ML+Cheat+Sheet_2.pdf

Cui, Z., Ke, R., Pu, Z., Wang, Y. (2019) ‘Deep Bidirectional and Unidirectional Recurrent Neural Network for Network-wide Traffic Speed Prediction’, [online]. Available at: https://arxiv.org/abs/1801.02143

Dagnino, G. B., Padula, G. (2009) ‘COOPETITION STRATEGY A NEW KIND OF INTERFIRM DYNAMICS FOR VALUE CREATION’, [online]. Available at: https://www.researchgate.net/publication/228605296

Dalzell, A. M., McArdle, S., Berta, M., Bienias, P., Chen, C-F, Gily´, A. et al. (2023) ‘Quantum algorithms: A survey of applications and end-to-end complexities’, *[quant-ph]* arXiv:2310.03011v1

Datarevenue, (20xx) ‘Task orchestration tools and workflows’, *datarevenue.com* [online]. Available at: https://www.datarevenue.com/en-blog/airflow-vs-luigi-vs-argo-vs-mlflow-vs-kubeflow

Dargan, J. (2023) ‘D-Wave Quantum Annealer Practical Usage in 2023’, *Quantum Insider* [online]. Available at: https://thequantuminsider.com/2023/05/05/d-wave-quantum-annealer-practical-usage-in-2023/ (Accessed: 5 January 2024)

Daugherty, P., and Wilson, J. (2018) *Human+Machine: Reimagining Work in the Age of AI*: Harvard Business Review Press, Boston.

Deng, Y. (2019) ‘Deep Learning on Mobile Devices – A Review’ [online]. Available at: https://www.researchgate.net/publication/331533064

Deloitte AI Institute. (2023) ‘Generative AI is all the rage’, *Deloitte Development LLC*. [online]. Available at: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-gen-ai-for-enterprises.pdf (Accessed 20 Jan. 2024).

Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *Google AI Language,* arXiv:1810.04805v2

Dhankhar, P. (2018) ‘RNN and LSTM based Chatbot using NLP’, *International Journal of Innovations in Engineering and Technology (IJIET),* Volume 10 Issue 2 [online]. Available at: http://dx.doi.org/10.21172/ijiet.102.32

Dhariwal, P., Nichol, A. (2021) ‘Diffusion Models Beat GANs on Image Synthesis’ [online]. Available at: https://arxiv.org/abs/2105.05233

DHDC. (2023) ‘Our different user roles - DEEP-Hybrid-DataCloud DEEP-2 documentation’, *Deep Hybrid Data Cloud, EU* [online] Available at: https://docs.deep-hybrid-datacloud.eu/en/latest/user/overview/user-roles.html (Accessed 5 Jan. 2024).

DigitalRoute (2022). ‘Telecom solutions for BSS, policy control and OSS’*, DigitalRoute* [online] Available at: https://www.digitalroute.com/solutions/telecom/ (Accessed 5 Jan. 2024)

Digout Jacques, Senechal Sylvain, Salloum Charbel (2019). *Methods and Tools for Completing Doctor of Business Administration (DBA) Theses*: Cambridge Scholars Publishing.

DiVito, L., Sharma, G. (2019) ‘Strategies of Multilateral Coopetition: Experienced Tensions and Coopetition Capabilities’, *Article in Academy of Management Proceedings*. DOI: 10.5465/AMBPP.2019.15049abstract [online]. Available at: https://www.researchgate.net/publication/334851615

Domo.com (2024) ‘Amazon Sagemaker Adoption’, *Domopalooza* [online]. Available at: https://www.domo.com/domopalooza/resources/amazon-sagemaker-adoption (Accessed 5 Jan. 2024)

Dong, Y., Wang, Z., Sreedhar, M. N., Wu, X., Kuchaiev, O. (2023) ‘SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF’, *NVIDIA*. [online]. Available at: https://arxiv.org/abs/2310.05344

Duc, A. N., Cruzes, D. S., Hanssen, G. K., Snarby, T and Abrahamsson, P. (2017) ‘Coopetition of software firms in Open-source software ecosystems’, *Lecture Notes in Business Information Processing, vol 304. Springer, Cham*, [Online]. Available at: https://doi.org/10.1007/978-3-319-69191-6_10

@Masteringllm. (2023) ‘LLM Training: A Simple 3-Step Guide You Won’t Find Anywhere Else!*’*, *Medium*. Oct 1, 2023 [online]. Available at: https://medium.com/@masteringllm/llm-training-a-simple-3-step-guide-you-wont-find-anywhere-else-98ee218809e5

Editorial@TRN. (2020) ‘Top Frameworks to Explore Quantum Computing’, *Medium.com* Jul 4, 2020 [online]. Available at: https://medium.com/the-research-nest/top-frameworks-to-explore-quantum-computing-2485c678a15a
(Accessed: 5 January 2024)

Edge WP. (2019) ‘5G AT THE EDGE’, *Whitepaper 5G Americas* [online]. Available at: https://www.5gamericas.org/wp-content/uploads/2019/10/5G-Americas-EDGE-White-Paper-FINAL.pdf

Espósito, F. (2024) ‘OpenAI is helping Apple fix Siri, and that has Microsoft worried. 9to5Mac’, [online]. Available at: https://9to5mac.com/2024/05/29/openai-apple-fix-siri-microsoft-worried/ (Accessed 27 Jun. 2024)

ETSI GANA. (2020) ‘Autonomic network engineering for the self-managing Future Internet (AFI); An Instantiation and Implementation of the Generic Autonomic Network Architecture (GANA) Model onto Heterogeneous Wireless Access Technologies using

Cognitive Algorithms'. *ETSI TR* 103 626 V1.1.1 (2020-02). Reference: DTR/INT-001-AFI-0027 [online]. Available at: http://www.etsi.org/standards-search

ETSI MEC, (202x) 'Multi-access Edge Computing (MEC)', *ETSI* [online]. Available at: https://www.etsi.org/technologies/multi-access-edge-computing

ETSI GS MEC 003. (2016) 'Mobile Edge Computing (MEC); Framework and Reference Architecture', *ETSI GS MEC* 003 V1.1.1 (2016-03). [Online] Available at: http://www.etsi.org/standards-search

Fautrero, Valerie & Gueguen, Gael. (2013) 'The dual dominance of the Android business ecosystem', *Toulouse Business School* [online] Available at: https://www.researchgate.net/publication/260105011_The_dual_dominance_of_the_Android_business_ecosystem

Felten, F., Talbi, E-G. and Danoy, G. (2023) 'Multi-Objective Reinforcement Learning based on Decomposition: A taxonomy and framework', [online]. Available at: https://arxiv.org/abs/2311.12495

Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P. (2023) 'Generative AI', *Springer Publishing* [online]. Available at: https://doi.org/10.1007/s12599-023-00834-7

Flach P. (2012) *Machine Learning- The Art and Science of Algorithms that Make Sense of Data*: Cambridge University Press.

Foramitti, J., (2021) 'AgentPy: A package for agent-based modeling in Python', Journal of Open-Source Software, 6(62), 3065 [online]. Available at: https: //doi.org/10.21105/joss.03065

Forbes SWOT Analysis Template. (n.d.) Forbes [online]. Available at: https://www.forbes.com/advisor/wp-content/uploads/2022/01/SWOT_Analysis_Template.pdf

Franzen, C. (2023) 'McKinsey report finds generative AI could add up to $4.4 trillion a year to the global economy', *VentureBeat* [online]. Available at: https://venturebeat.com/ai/mckinsey-report-finds-generative-ai-could-add-up-to-4-4-trillion-a-year-to-the-global-economy/ (Accessed 5 Jan. 2024).

Fu, Z., Yang, H., So, A. M-C., Lam, W., Bing, L. and Collier, N. (2022). 'On the Effectiveness of Parameter-Efficient Fine-Tuning', [online]. Available at: https://doi.org/10.48550/arXiv.2211.15583 g2.com. (2024) ‘Top 10 RATH Alternatives & Competitors’, G2.com [online]. Available at: https://www.g2.com/products/rath/competitors/alternatives (Accessed: 5 January 2024).

Galarnyk, M. (2021) ‘Machine Learning with PyTorch and Ray’, *Medium.com* [online]. Available at: https://medium.com/distributed-computing-with-ray/getting-started-with-distributed-machine-learning-with-pytorch-and-ray-27175a1b4f25

Gannon, D. (2024) ‘A Brief Look at Autogen: a Multiagent System to Build Applications Based on Large Language Models’, Technical Report [online]. Available at: https://www.researchgate.net/publication/377411718

Gao, X., Zhang, Z.-Y. and Duan, L.-M. (2018) ‘A quantum Machine Learning algorithm based on generative models’, *Science Advances*, Volume 4, Issue 12. [Online]. Available at: https://www.science.org/doi/pdf/10.1126/sciadv.aat9004

Gaur, Nitin. (2021) ‘Why Kubernetes is de-facto choice for every company going cloud native*’,* [online]. Available at: https://www.linkedin.com/pulse/why-kubernetes-de-facto-choice-every-company-going-cloud-nitin-gaur

Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M. (1998) ‘The Belief-Desire-Intention Model of Agency’, [online]. Available at: https://www.cs.ox.ac.uk/people/michael.wooldridge/pubs/atal98b.pdf

Georgeff, M., Wooldridge, M., Tambe, M (1970) ‘The Belief-Desire-Intention Model of Agency’, [online]. Available at: https://www.researchgate.net/publication/2596320

Ghosh, Ashish & Chakraborty, Debasrita & Law, Anwesha. (2018) ‘Artificial Intelligence in Internet of Things*’, IET Research Journals.*

Ghosh, B. (2023) ‘Power of Vector Databases for Gen AI Applications’, *Medium.com* [online]. Available at: https://medium.com/@bijit211987/power-of-vector-databases-for-gen-ai-applications-a63d4cf7e352 (Accessed: 5 January 2024).

Girardi, R. and Leite, A. (2013) ‘A Survey on Software Agent Architectures’, *IEEE Intelligent Informatics Bulletin*, Vol.14 No.1 [online]. Available at: https://www.comp.hkbu.edu.hk/~iib/2013/Dec/article2/iib_vol14no1_article2.pdf

Glisic S., Lorenzo, B. (2022) *Artificial Intelligence and Quantum Computing for Advanced Wireless Networks*: WILEY.

Gronauer, S., Diepold, K. (2021) 'Multi-agent deep reinforcement learning: a survey', *Artificial Intelligence Review* 55:895–943 [online]. Available at: https://doi.org/10.1007/s10462-021-09996-w

Guavus (2022) 'NWDAF Autonomous 5G Networks', *Guavus* [online]. Available at: https://www.guavus.com/wp-content/uploads/2022/11/Guavus_NWDAF_Autonomous_5G_Networks_eBook_2021.pdf

Guerra-Hernández, A., Fallah, A. E., Sorbonne, S., Soldano, H. (2004) 'Learning in BDI Multi-agent Systems', DOI: 10.1007/978-3-540-30200-1 [online]. Available at: https://www.researchgate.net/publication/226976978

GÜREL, E., TAT. M. (2017) 'SWOT ANALYSIS: A THEORETICAL REVIEW' [online]. Available at:
Doi Number: http://dx.doi.org/10.17719/jisr.2017.1832

Hadjer, B. (2020) 'Comparison of Deep Learning Frameworks and Compilers', *Master Thesis*, DOI: 10.13140/RG.2.2.15094.22085 [online]. Available at: https://www.researchgate.net/publication/343320122

HAI. (2023) 'The AI Index Report – Artificial Intelligence Index', *Stanford University* [online]. Available at: https://aiindex.stanford.edu/report/
& at: https://aiindex.stanford.edu/report/HAI_AI_Index_Report_2023.pdf
(Accessed: 5 January 2024)

HAI. (2024) 'The AI Index Report – Artificial Intelligence Index', *Stanford University* [online]. Available at: https://arxiv.org/abs/2405.19522
(Accessed: 22 November 2024)

Hammad, I., El-Sankary, K., Gu, J. (2019) 'A Comparative Study on Machine Learning Algorithms for the Control of a Wall Following Robot', *IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE*, 2019. (Pages: 2995 – 3000) [online]. Available at: https://arxiv.org/pdf/1912.11856.pdf

Han, J. and Kamber, M. (2006) *Data Mining, Concepts and Technologies:* 2nd ed. Amsterdam: Elsevier.

Hao, K. (2019) 'The computing power needed to train AI is now rising seven times faster than ever before', *MIT Technology Review. November* 11, 2019 [online]. Available at: https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/ (Accessed: 5 January 2024).

Hao, S. (2023) ‘Reinforcement Learning in the Era of LLMs: What is Essential? What is needed? An RL Perspective on RLHF, Prompting, and Beyond’, *[Preliminary Work]* [online]. Available at: https://arxiv.org/abs/2310.06147

Hastie, T., Tibshirani, R., Friedman, J. (2017) *Elements of Statistical Learning*: 2nd Edition, Springer.

He, X., Zhao, K. & Chu, X. (2021) ‘AutoML: A Survey of the State-of-the-Art’, arXiv:1908.00709v6 [cs.LG]

Higashino, M., Kawato, T., Kawamura, T. (2018)
‘A Design for Application of Mobile Agent Technology to MicroService Architecture’, *International Science Index, Computer and Information Engineering*, Vol:12, No:2 [online]. Available at: https://publications.waset.org/Publication/10008517

Hinkle, M. (2023) ‘Vector Databases: Long-Term Memory for Artificial Intelligence’, *Thenewstack* [online]. Available at: https://thenewstack.io/vector-databases-long-term-memory-for-artificial-intelligence/ (Accessed: 5 January 2024).

Hinz, O and Eckert, J. (2010) ‘The Impact of Search and Recommendation Systems on Sales in Electronic Commerce’, *Research Paper, Business & Information Systems Engineering*. DOI 10.1007/s12599-010-0092-x [online]. Available at: https://link.springer.com/article/10.1007/s12599-010-0092-x

Hiter, S. (2023*)* ‘Top 9 Generative AI Applications and Tools’. *eweek.com* [online]. Available at: https://www.eweek.com/artificial-intelligence/generative-ai-apps-tools/ (Accessed: 5 January 2024)

Ho, H-N, Lee, E. (2015) ‘Model-based reinforcement learning approach for planning in self-adaptive software system’, *Proceedings of the 9th international conference on ubiquitous information management and communication*; 2015. p. 1–8. DOI: 10.1145/2701126.2701191

Hu, K., Li, Y., Xia, M., Wu, J., Lu, M., Zhang, S., Weng, L. (2021) ‘Federated Learning: A Distributed Shared Machine Learning Method’, *Research Article from Hindawi Complexity* Volume 2021, Article ID 8261663 [online]. Available at: https://doi.org/10.1155/2021/8261663

Hua, Y., Shen, Wei., Wang, B., Liu, Y., Jin, S., Liu, Q. et al. (2023) ‘Secrets of RLHF in Large Language Models Part I: PPO’ [online]. Available at: https://arxiv.org/abs/2307.04964

Huang, X., Lokanatha, A.K., Karp, M., Das, S. (2023) 'Domain-adaptation Fine-tuning of Foundation Models in Amazon SageMaker JumpStart on Financial data', *Amazon SageMaker* [online]. Available at: https://aws.amazon.com/blogs/machine-learning/domain-adaptation-fine-tuning-of-foundation-models-in-amazon-sagemaker-jumpstart-on-financial-data/ (Accessed: 5 January 2024)

Hetzner, C. (2023) 'Satya Nadella instructed Microsoft to design its own silicon chip—and it could end Nvidia's stranglehold over the sector', *fortune.com* [online]. Available at:
https://fortune.com/2023/11/16/microsoft-satya-nadella-semiconductor-chips-nvidia-chatgpt-openai/?utm_source=search&utm_medium=advanced_search&utm_campaign=search_link_clicks
(Accessed: 5 January 2024)

Hughes, C., Isaacson, J., Perry, A., Sun, R. F., Turner, J. (2021) *Quantum Computing for the Quantum Curious. Springer:* (Open-Access eBook) [online]. Available at: https://doi.org/10.1007/978-3-030-61601-4

Humeau, S., Shuster, K., Lachaux, M-A., Weston, J. (2020) 'Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring'. *Conference paper at ICLR 2020*. arXiv:1905.01969v4

Hutter, F., Kotthoff, L., Vanschoren, J. (2019) *Automated Machine Learning Methods, Systems, Challenges*: Springer.

3GPP Technical Report 23.700-91. (2020) 'Study on enablers for network automation for the 5G System (5GS) Phase 2 (Release 17)', *3GPP TR 23.700-91-h00* [online]. Available at: https://www.3gpp.org/ftp/Specs/archive/23_series/23.700-91

IMF. (2023) 'ARTIFICIAL INTELLIGENCE What AI means for economics. Finance & Development', *A Quarterly Publication of the International Monetary Fund* December 2023 | Volume 60 | Number 4 [online]. Available at: https://www.imf.org/en/Publications/fandd/issues/2023/12/Macroeconomics-of-artificial-intelligence-Brynjolfsson-Unger

Intel and Nokia Siemens Networks, (2013) 'Increasing mobile operators' value proposition with edge computing', *Technical brief* [online]. Available at: https://www.intel.co.id/content/dam/www/public/us/en/documents/technology-briefs/edge-computing-tech-brief.pdf

Jacob, J. and Nair, S. (2022) ‘LLMs, a brief history and their use cases’, *exemplary.ai* [online]. Available at: https://exemplary.ai/blog/llm-history-usecases

Jakob, P. (2023) ‘Improved Performance of Superconducting Qubits Makes Investigation of Sapphire Substrates Compelling as An Alternative to Silicon’, *The Quantum Insider* [online]. Available at: https://thequantuminsider.com/2023/12/14/improved-performance-of-superconducting-qubits-makes-investigation-of-sapphire-substrates-compelling-as-an-alternative-to-silicon/

Jones, A. (202x) ‘K8sGPT - Cloud Native Sandbox’, *k8sgpt.ai* [online]. Available at: https://k8sgpt.ai/ (Accessed: 5 January 2024)

Kafka. (201x) ‘Apache Kafka’ [online]. Available at: https://kafka.apache.org/

Kanade, V. (2022) ‘What Is Transfer Learning? Definition, Methods, and Applications [Online] Spiceworks’, *AI Researcher* [online]. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/articles-what-is-transfer-learning/ (Accessed: 5 January 2024)

kanaries.net. (2023) ‘RATH: The Future of Automated Data Analysis and Visualization’, *Kanaries.net* [online]. Available at: https://docs.kanaries.net/blog/rath-future-automated-data-analysis-visualization (Accessed: 5 January 2024).

Kaelbling, LP, Littman ML, Moore AW. (1996) ‘Reinforcement learning: a survey’, *J Artif Intell Res*; 4:237–85 [online]. Available at: https://arxiv.org/pdf/cs/9605103.pdf

Katsov, I. (2018) *INTRODUCTION TO ALGORITHMIC MARKETING*: Grid Dynamics.

Kelleher, J.D., Namee, B.M., D’Arcy (2015) *Fundamentals of Machine Learning for Data Analytics:* MIT Press.

Kerzel, U. (2021) ‘Enterprise AI Canvas Integrating Artificial Intelligence into Business’, *Applied Artificial Intelligence*, 35:1, 1-12, DOI: 10.1080/08839514.2020.1826146 [online]. Available at: https://doi.org/10.1080/08839514.2020.1826146

Khan, R. (2017) ‘Standardized Architecture for Conversational Agents a.k.a. ChatBots’, *International Journal of Computer Trends and Technology*, 50(2), 114–121 [online]. Available at: https://doi.org/10.14445/22312803/ijctt-v50p120

Kim, W. C., & Mauborgne, R. (2015) *Blue Ocean strategy: How to create uncontested market space and make the competition irrelevant*: Harvard Business Review Press, Boston, Massachusetts.

Kinney, S. (2023) 'How is generative AI relevant to telecom operators?*', RCRWireless News* [online]. Available at: https://www.rcrwireless.com/20230724/5g/how-is-generative-ai-relevant-to-telecom-operators

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E. et al. (2023) 'UNDERSTANDING THE EFFECTS OF RLHF ON LLM GENERALISATION AND DIVERSITY' [online]. Available at: https://arxiv.org/abs/2310.06452

Kiwit, F.J., Marso, M., Ross, P., Riofr´ıo, C.A., Klepsch, J., Luckow, A. (2023) 'Application-Oriented Benchmarking of Quantum Generative Learning Using QUARK', *[quant-ph]* arXiv:2308.04082v1

Kokab, S.T., Asghar, S. and Naz, S. (2022) 'Transformer-based deep learning models for the sentiment analysis of social media data', *Elsevier Inc.* [online]. Available at: https://doi.org/10.1016/j.array.2022.100157

Kolajo, T., Daramola, O., Adebiyi, A. (2019) 'Big data stream analysis: a systematic literature review', *Survey Paper* [online]. Available at: https://doi.org/10.1186/s40537-019-0210-7

Kriegeskorte, N., Golan, T. (2019) 'Neural network models and deep learning – a primer for biologists' [online]. Available at: https://arxiv.org/abs/1902.04704

Krohn, J., Beyleveld, G., Bassens, A. (2020) *DEEP LEARNING ILLUSTRATED* – A *visual Interactive Guide to Artificial Intelligence*: Pearson Education, Inc.

Kubeflow. (2023) 'Kubeflow brings MLOps to the CNCF Incubator', *Cloud Native Computing Foundation*. July 25, 2023 [online]. Available at: https://www.cncf.io/blog/2023/07/25/kubeflow-brings-mlops-to-the-cncf--incubator/ (Accessed: 5 January 2024)

'Kubernetes' (2020). *Wikipedia* [online]. Available at: https://en.wikipedia.org/wiki/Kubernetes. (Accessed: 5 January 2024)

Kulkarni, Viraj & Kulkarni, Milind & Pant, Aniruddha. (2020) 'Quantum Computing Methods for Supervised Learning', *[quant-ph],* arXiv:2006.12025v1

'Large language model' (n.d.). *Wikipedia* [online] Available at: https://en.m.wikipedia.org/wiki/Large_language_model

Lee, Hoejoo & Cha, Jiwon & Kwon, Daeken & Jeong, Myeonggi. (2021) 'Hosting AI/ML Workflows on O-RAN RIC Platform', *IEEE Globecom Workshops (GC Wkshps)*, DOI:10.1109/GCWkshps50303.2020.9367572

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C. et al. (2023) 'RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback', *Google Research* [online]. Available at: https://arxiv.org/abs/2309.00267

Leppänen, T. (2019) 'Distributed Artificial Intelligence with Multi-Agent Systems for MEC'. *Journal Paper* · August 2019 DOI: 10.1109/ICCCN.2019.8846960 [online]. Available at: https://www.researchgate.net/publication/333039750_Distributed_Artificial_Intelligence_with_Multi-Agent_Systems_for_MEC

Leswing, K. (2024) 'Intel unveils latest AI chip as Nvidia competition heats up', *CNBC*. [online]. Available at: https://www.cnbc.com/2024/04/09/intel-unveils-gaudi-3-ai-chip-as-nvidia-competition-heats-up-.html (Accessed 13 Apr. 2024)

Letaief, Khaled B. & Chen, Wei & Shi, Yuanming & Zhang, Jun & Zhang, Ying-Jun Angela. (2019) 'The Roadmap to 6G: AI Empowered Wireless Networks'**,** *IEEE Communications Magazine*.

Leventi-Peetz, A-M. (2022) 'Deep Learning Reproducibility and Explainable AI (XAI)', arXiv:2202.11452v3 [cs.LG]

Lewin J. D. (1990) *PARTNERSHIPS FOR PROFIT: Structuring and Managing Strategic Alliances:* Free Press.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N. et al. (2021) 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', [online]. Available at: https://arxiv.org/abs/2005.11401

Levy, S. (2021) 'How Y Combinator Changed the World', *wired.com* [online]. Available at: https://www.wired.com/story/how-y-combinator-changed-the-world/

Li, Q., He, B., Song, D. (2021) 'Practical One-Shot Federated Learning for Cross-Silo Setting', *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)* [online]. Available at: https://arxiv.org/abs/2010.01017

Littman, M. L. (1994) 'Markov games as a framework for multi-agent reinforcement learning', DOI: 10.1016/b978-1-55860-335-6.50027-1 [online]. Available at: https://wwwsemanticscholar.org/paper/Markov-Games-as-alFramework-for-Multi-Agent-Littman/7fbf55baccbc5fdc7ded1ba18330605909aef5e5

Liu, M., Ovhal, P., Dwivedi, K. (2023) ‘Azure private multi-access edge compute partner solutions’, Microsoft Azure [online]. Available at: https://learn.microsoft.com/en-us/azure/private-multi-access-edge-compute-mec/partner-programs (Accessed: 5 January 2024).

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I. (2020) ‘Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments’, arXiv:1706.02275v4 [cs.LG]

Luo, S., Zhang, L., and Fan, Y. (2021b) ‘Real-time scheduling for dynamic partialno-wait Mult objective flexible job shop by deep reinforcement learning’, *IEEE Trans. Autom. Sci. Eng*. 19, 3020–3038. doi:10.1109/tase.2021.3104716

Macrometa, (2024) ‘Apache Spark and Flink’, *macromerta.com* [online]. Available at: https://www.macrometa.com/event-stream-processing/spark-vs-flink

Mahony, N. O., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L. Riordan, D., Walsh, J. (2019) ‘Deep Learning vs. Traditional Computer Vision’ [online]. Available at: https://arxiv.org/abs/1910.13796.pdf

Mao, Hongzi & Alizadeh, Mohammad & Menache, Ishai & Kandula, Srikanth. (2016) ‘Resource Management with Deep Reinforcement Learning’, *MIT & Microsoft Research* [online]. Available at: https://people.csail.mit.edu/alizadeh/papers/deeprm-hotnets16.pdf

Mann, T. (2024) ‘Now OpenAI CEO Sam Altman wants billions for AI chip fabs’, *The Register* [online]. Available at: https://www.theregister.com/AMP/2024/01/20/altman_chip_fabs/ (Accessed 5 Jan. 2024)

Marr, B. (2024) ‘Generative AI Sucks: Meta’s Chief AI Scientist Calls for A Shift to Objective-Driven AI’, *Forbes.com* [online]. Available at: https://www.forbes.com/sites/bernardmarr/2024/04/12/generative-ai-sucks-metas-chief-ai-scientist-calls-for-a-shift-to-objective-driven-ai/?sh=696c2017b82b [Accessed 13 Apr. 2024].

MathWorks Inc. (2019) ‘Deep Learning or Machine Learning’, *The MathWorks, Inc*, 18 09 2019 [online]. Available: https://explore.mathworks.com/machinelearning-vs-deep-learning/chapter-1-129M100NU.html (Accessed 18 09 2019)

Maurya, A. (2012) 'Why Lean Canvas vs Business Model Canvas', *Leanstack*. Feb 27, 2012 . [online]. Available at: https://blog.leanstack.com/why-lean-canvas-vs-business-model-canvas/ (Accessed 10 Jan. 2024).

McEnroe, P., Wang, S. and Liyanage, M. (2022) A Survey on the Convergence of Edge Computing and AI for UAVs: Opportunities and Challenges. IEEE INTERNET OF THINGS JOURNAL, VOL. 9, NO. 17, [online]. Available at: https://doi.org/10.1109/JIOT.2022.3176400

McKinsey & Company. (2023) 'Economic potential of generative AI', *McKinsey* [online]. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/ (Accessed 5 Jan. 2024).

McMahan, B. (2019) 'Federated Learning, from Research to Practice', *CMU* 2019.09.05. [online]. Available at: https://www.pdl.cmu.edu/SDI/2019/slides/2019-09-05Federated%20Learning.pdf

Meriem, T.B., Chaparadza, R., Radier, B., Soulhi, S., LozanoLópez, J-A., Prakash, A. (2016) 'GANA - Generic Autonomic Networking Architecture Reference Model for Autonomic Networking, Cognitive Networking and Self-Management of Networks and Services', *ETSI White Paper* No. 16. ISBN No. 979-10-92620-10-8

Meyer, R and Volberda, H.W. (1997) 'Porter on Corporate Strategy', *Article,* DOI: 10.1007/978-1-4615-6179-8_4 [online]. Available at: https://www.researchgate.net/publication/254804823

Mismar, F. B., Evans, B. L., Ahmed A. (2020) 'Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination', DOI: 10.1109/TCOMM.2019.2961332 [Online]. Available at: https://www.researchgate.net/publication/337962607

Mitola, Joseph. (2000) 'Cognitive Radio - An Integrated Radio architecture for Software Defined Radio*', PhD Thesis, Royal Institute of Technology (KTH) - Teleinformatics.*

MLOps. (201x) 'Machine Learning Operations' [online]. Available at: https://ml-ops.org/

MLOps Platforms. (2024) ‘thoughtworks/mlops-platforms’, *Thoughtworks.* [online] Available at: https://github.com/thoughtworks/mlops-platforms (Accessed: 5 January 2024).

Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Harley, T.; Lillicrap, T.P.; Silver, D.; Kavukcuoglu, K. (2016) ‘Asynchronous Methods for Deep Reinforcement Learning’, *In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA*; Volume 48, pp. 1928–1937 [online]. Available at: https://arxiv.org/abs/1602.01783

MobiledgeX. (2022) ‘Multi-Access Edge Computing Solution’, *Fiercewireless.com* [online]. Available at: *https://www.fiercewireless.com/wireless/google-acquires-edge-software-provider-mobiledgex*

Modi, M. (2023) ‘Top 10 Kubernetes Alternatives in 2023’, *knowledgehut.com*. 27th Dec, 2023 [online]. Available at: https://www.knowledgehut.com/blog/devops/kubernetes-alternatives. (Accessed: 5 January 2024)

Mohammad, Umair & Sorour, Sameh. (2018) ‘Adaptive task allocation for mobile edge learning’, arXiv:1811.03748 [cs.DC]

MSV, J. (2023) ‘Generative AI Cloud Platforms: AWS, Azure, or Google?’, *thenewstack.io* [online]. Available at: https://thenewstack.io/generative-ai-cloud-services-aws-azure-or-google-cloud/ . (Accessed: 5 January 2024)

MSV, J. (2019) ‘How Kubernetes Has Changed the Face of Hybrid Cloud’, Forbes.com. [Online]. Available at: https://www.forbes.com/sites/janakirammsv/2019/12/16/how-kubernetes-has-changed-the-face-of-hybrid-cloud/?sh=4e61d516228d

Naveed, H., Khan, A-U., Qiu, S., Saqib, M., Anwar, S., Usman, M. et al. (2023) ‘A Comprehensive Overview of Large Language Models’, *PREPRINT* [online]. Available at: https://arxiv.org/abs/2307.06435

Negnevitsky M. (2011) *Artificial Intelligence- A guide to Intelligent Systems*: 3rd ed., Pearson Education Ltd.

Nguyen, G., Dlugolinsky, S., Bobak M. and Tran, V. (2019) ‘Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey’, *Article in Artificial Intelligence*. DOI: 10.1007/s10462-018-09679-z [online]. Available at: https://www.researchgate.net/publication/329990977

Nguyen, T. T., Nguyen, N. D. Vamplew, P., Nahavandi, S., Dazeley, R., Lim, C. P. (2020) 'A Multi-Objective Deep Reinforcement Learning Framework' [online]. Available at: https://arxiv.org/abs/1803.02965

Nikaein, Navid & Bonnet, Christian & Knopp, Raymond & Ksentini, Adlen & Kaltenberger, Florian & Gupta, Rohit. (2017) 'Towards building Cloud-Native Radio Access Network using OpenAirInterface', *Eurecom* [online]. Available at: https://openairinterface.org/community/whitepapers/towards-building-cloud-native-radio-access-network-using-openairinterface/

Nolan, B. (2021) 'Sam Altman sought billions to fund a chip company that could rival Nvidia before he was fired, report says', *Business Insider* Nov 21 2023 [online] Available at: https://www.businessinsider.in/tech/news/sam-altman-sought-billions-to-fund-a-chip-company-that-could-rival-nvidia-before-he-was-fired-report-says/articleshow/105389485.cms (Accessed: 5 January 2024)

Nolan, B. (2023) 'Microsoft is trying to reduce its reliance on OpenAI by developing a cheaper, less powerful AI model, report says', *Business Insider* [online]. Available at: https://www.businessinsider.in/tech/news/microsoft-is-trying-to-reduce-its-reliance-on-openai-by-developing-a-cheaper-less-powerful-ai-model-report-says/articleshow/104017974.cms (Accessed 5 Jan. 2024)

O'Brien, Matt. (2024) 'AI 'gold rush' for chatbot training data could run out of human written text', *The Washington Times*. [online]. Available at: https://www.washingtontimes.com/news/2024/jun/6/ai-gold-rush-for-chatbot-training-data-may-run-out/

Oderanti, Festus Oluseyi. (2013) 'Fuzzy inference game approach to uncertainty in business decisions and market competitions', Springer Open Journal, SpringerPlus.

O-RAN Alliance, (2018) 'O-RAN: Towards an Open and Smart RAN' [Online]. Available at: https://static1.squarespace.com/static/5ad774cce74940d7115044b0/t/5bc79b371905f4197055e8c6/1539808057078/O-RAN+WP+FInal+181017.pdf

O-RAN Alliance, (2020) 'O-RAN Use Cases and Deployment Scenarios Towards Open and Smart RAN*'* [online]. Available at: https://static1.squarespace.com/static/5ad774cce74940d7115044b0/t/5e95a0a306c6ab2d1cbca4d3/1586864301196/O-RAN+Use+Cases+and+Deployment+Scenarios+Whitepaper+February+2020.pdf

Ogbuachi, M.C., Reale, A., Suskovics, P. and Kovacs, B. (2020) ‘Context-Aware Kubernetes Scheduler for Edge-native Applications on 5G’, *JOURNAL OF COMMUNICATIONS SOFTWARE AND SYSTEMS*, VOL. 16, NO. 1
DOI: 10.24138/jcomss.v16i1.1027

Okaibedi, D. (2023) “ChatGPT and the Rise of Generative AI: Threat to Academic Integrity?”, *Journal of Responsible Technology* [online]. Available at: https://doi.org/10.1016/j.jrt.2023.100060

OpenAI. (2023) ‘Our structure’, [online]. Available at: https://openai.com/our-structure

OpenAI. (2024) ‘Learning to Reason with LLMs’, [online]. Available at: https://openai.com/index/learning-to-reason-with-LLMs/

Osterwalder, A., Pigneur, Y. (2010) *Business Model Generation*: John Wiley & Sons. Inc.

Owczarek, D. (2023) ‘Lambda vs. Kappa Architecture. A Guide to Choosing the Right Data Processing Architecture for Your Needs’, *Nexocode*, December 30, 2022 – 233 *UPDATED ON* APRIL 24, 2023 [online]. Available at: https://nexocode.com/blog/posts/lambda-vs-kappa-architecture/ (Accessed: 5 January 2024).

Padgham, L (20xx) ‘Introduction to Belief Desire Intention Agents’, *RMIT University, Melbourne, Australia*. [online]. Available online: http://goanna.cs.rmit.edu.au/linpa/Presentation/BdiIntro.pdf

Park, S. (2022) ‘Introduction to Federated Learning’, *Machine Learning and Vision Lab, UNIST* [online]. Available at: https://namhoonlee.github.io/courses/optml/s14-fl.pdf

Pascual, A.L. (2021) ‘Comparison of AutoML solutions 2021’, *Analytics Vidhya* [online]. Available at: https://medium.com/analytics-vidhya/comparison-of-automl-solutions-2021-6625d494b695 (Accessed: 5 January 2024)

PATEROMICHELAKIS, E., MOGGIO, F., MANNWEILER, C., ARNOLD, P., SHARIAT, M., EINHAUS, M., WEI, Q., BULAKCI, Ö., AND DOMENICO, A.D. (2019) ‘End-to-End Data Analytics Framework for 5G Architecture’, *SPECIAL SECTION ON ROADMAP TO 5G: RISING TO THE CHALLENGE*. Digital Object Identifier 10.1109/ACCESS.2019.2902984

Pedersen, Magnus. (2016) ‘Artificial Intelligence for Long-Term Investing’, *Hvass Laboratories* [online]. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740218

Pfarrer, M. D. and Smith, K. G. (2015) ‘Creative Destruction’ [Online]. Available at: https://www.researchgate.net/publication/319587682

Piper, K. (2024) ‘Inside OpenAI’s multibillion-dollar gambit to become a for-profit company’, Vox [online]. Available at: https://www.vox.com/future-perfect/380117/openai-microsoft-sam-altman-nonprofit-for-profit-foundation-artificial-intelligence

Ponomarev, S., Voronkov A. E. (2017) ‘*multi-agent systems and decentralized artificial superintelligence’,* [online]. Available at: https://www.arxiv.org/abs/1702.08529

Poulton, N. and Joglekar, P. (2020*) THE KUBERNETES BOOK:* Shroff Publishers and Distributors Pvt. Ltd, India

Perdomo-Ortiz, A., Benedetti, M., Realpe-Gómez, J. and Biswas, R. (2018) ‘Opportunities and challenges for quantum assisted machine learning in near-term quantum computers’, *Quantum Sci. Technol*. 3. *IOP Publishing*. https://doi.org/10.1088/2058-9565/aab859

Pettersson, Linus. (2020) ‘Convolutional Neural Networks on FPGA and GPU on the Edge: A Comparison’, *Upssala University*, ISSN: 1401-5757, UPTEC F 20028 [online]. Available at: https://www.diva-portal.org/smash/get/diva2:1447576/FULLTEXT01.pdf

POUYANFAR, S., SADIQ, S., Yan, Y. (2018) ‘A Survey on Deep Learning: Algorithms, Techniques, and Applications’, *Association for Computing Machinery*. 0360-0300/2018/09-ART92 $15.00 https://doi.org/10.1145/3234150 *ACM Computing Surveys*, Vol. 51, No. 5, Article 92 [online]. Available at:
Online: https://www.semanticscholar.org/paper/A-Survey-on-Deep-Learning-Pouyanfar-Sadiq/cb8a1b8d87a3fef15635eb4a32173f9c6f966055

Promwongsa, Nattakorn & Ebrahimzadeh, Amin & Naboulsi, Diala & Kianpisheh, Somayeh & Belqasmi, Fatna & Glitho, Roch & Crespi, Noel & Alfandi, Omar. (2020) ‘A Comprehensive Survey of the Tactile Internet: State-of-the-art and Research Directions’, *IEEE Communications Surveys & Tutorials. PP*. 10.1109/COMST.2020.3025995.

PwC. (2020) ‘Transforming telecoms’ internal ecosystems How to rethink business support systems and operational support systems in the age of 5G’, [online]. Available at:

https://www.pwc.com/gx/en/industries/tmt/5g/pwc-transforming-telecoms-internal-ecosystems.pdf

Qi, J., Zhou, Q., Lei, L., Zheng, K. (2021) ‘Federated Reinforcement Learning: Techniques, Applications, and Open Challenges’, [online]. Available at: https://arxiv.org/abs/2108.11887v2

Radcom. (2020) ‘5G Architecture with Distributed NWDAF’, Radcom [online]. Available at: https://www.radcom.com/uploads/pdf/RADCOM's%20NWDAF%20Solution_ACE.pdf

Rao, A., Yigit, G. and Nagy, W. (2020) ‘BUILDING AUTONOMOUS NETWORKS FOR THE 5G ERA: A REFERENCE FRAMEWORK TO DELIVER BUSINESS OUTCOMES’, *[WP], analysysmason.com* [online]. Available at: https://www.analysysmason.com/contentassets/32729f9ab3554d4ab100a7b1028ae808/analysys_mason_5g_autonomous_networks__oct2020_rma07_rma01_rma02.pdf

Rao, A. S., and Georgeff, M. P. (1995) ‘BDI Agents: From Theory to Practice’, *Proceedings of the Final International Conference on MultiAgent Systems* [online] Available at: https://aaai.org/papers/ICMAS95-042-bdi-agents-from-theory-to-practice/

RAND, W., Evanston, IL. (2006) ‘MACHINE LEARNING MEETS AGENT-BASED MODELING: WHEN NOT TO GO TO A BAR’ [online]. Available at: https://www.semanticscholar.org/paper/MACHINE-LEARNING-MEETS-AGENT-BASED-MODELING-:-WHEN-Rand/65d565d5186345efc73457cf71cea61f6cfc8645

RCR Wireless News. (2017) ‘How Jio digitally transformed India and built a 4G LTE Network in 170 days’, *RCR Wireless* [online]. Available at: https://rcrwireless.com/20170418/sponsored/learn-reliance-jio-built-4g-lte-network-digitally-transformed-india-170-days

Read, A.P., Chapman, B.J., Lei, C.U., Curtis, J.C., Ganjam, S., Krayzman, L. et al. (2023) ‘Precision Measurement of the Microwave Dielectric Loss of Sapphire in the Quantum Regime with Parts-per-Billion Sensitivity’ [online] Available at: http://arxiv.org/abs/2206.14334v2

Rest API. (20xx) ‘Restful Application Programming Interface’, *blog.dreamfactory.com* [online]. Available at: *https://blog.dreamfactory.com/the-importance-of-loose-coupling-in-rest-api-design/#:~:text=What%20is%20Loose%20Coupling%2D%20Its,affect%20the%20operation%20of%20others*

Ribas-Fernandes, J., Solway, A., Diuk, C., McGuire, J. T., Barto A. G. (2011) ‘A Neural Signature of Hierarchical Reinforcement Learning’, *Neuron 71, 370–379, Elsevier Inc.* DOI: 10.1016/j.neuron.2011.05.042

Rogers, D. L. (2016) *The Digital Transformation Playbook: Rethink Your Business for the Digital Age*: Columbia Business School Publishing.

Roth, S., Leydesdorff, L., Kaivo-oja, J.R.L. and Sales, A. (2019) ‘Open coopetition: when multiple players and rivals team up’, *Article in Journal of Business Strategy*, DOI: 10.1108/JBS-11-2018-0192
https://www.researchgate.net/publication/333211141

Rowley, Jason D. (2023) *‘LLM Benchmarks: Guide to Evaluating Language Models’, Deepgram* [online]. Available at: https://deepgram.com/learn/llm-benchmarks-guide-to-evaluating-language-models

Ruffatti, G. (2009) ‘SpagoWorld, the Open-Source Initiative’, *Engineering Article* [online]. Available at: https://www.researchgate.net/publication/242266716

Rusko, R. (2012) ‘Perspectives on value creation and coopetition’, *Problems and Perspectives in Managemen*t, Volume 10, Issue 2 [online]. Available at: https://www.businessperspectives.org/images/pdf/applications/publishing/templates/article/assets/4601/PPM_2012_02_Rusko.pdf

Russel, Stuart Jonathan & Norvig, Peter. (2003). *Artificial Intelligence, A modern Approach:* Prentice Hall, Upper Saddle River, NJ.

Sagemaker Partners. (2024) ‘Amazon Sagemaker Partners’, *Amazon Sagemaker* [online]. Available at: https://aws.amazon.com/sagemaker/partners/ (Accessed 5 Jan. 2024).

Sagemaker Developer Guide. (2024) ‘Amazon Sagemaker: Developer Guide’, *Amazon Web Services* [online]. Available at: https://aws.amazon.com/sagemaker/ (Accessed 5 Jan. 2024).

‘Apps Run the World’ (2024) *List of Amazon SageMaker Customers.* [Online]. Available at: https://www.appsruntheworld.com/customers-database/products/view/amazon-sagemaker (Accessed 5 Jan. 2024).

Sam-Solutions (2022) ‘AWS vs. Azure vs. Google Cloud: Which is Better?’ [online]. Available at: https://sam-solutions.us/aws-vs-azure-v-s-google-cloud-which-is-better/

Santiso, J. (2013) ‘Startup Europe: The Accelerator & Incubator Ecosystem’, *Source: Telefónica Global Affairs & New Ventures* [online]. Available at: https://lisboncouncil.net/wp-content/uploads/2020/08/Javier-Santiso-Presentation.pdf

SAS NWDAF (2022) ‘5G NWDAF NETWORK DATA ANALYTICS FUNCTION’, *SAS Institute Inc*. [online]. Available at: https://www.sas.com/offices/pdf/mx/20221103-nwdaf.pdf

Schetakis, N., Aghamalyan, D., Griffin, P. and Boguslavsky, M. (2022) ‘Review of some existing QML frameworks and novel hybrid classical–quantum neural networks realizing binary classification for the noisy datasets.’ *Scientific Reports*. [online]. Available at: https://doi.org/10.1038/s41598-022-14876-6.

Schmarzo, B. (2020) *The Economics of Data, Analytics, and Digital Transformation*: Packt Publishing.

Scikit-multiflow. (2019) *‘A machine learning package for streaming data in Python’* [online]. Available at: https://scikit-multiflow.github.io/

Sekaran, U. and Bougie, R. (2016) *Research Methods for Business: A Skill-Building Approach:* 7th Edition, Wiley & Sons, West Sussex.

Seppo, H., Henning, S. and Sartori, C. (2011) *LTE Self-Organizing Networks (SON), Network Management Automation for Operational Efficiency:* John Wiley & Sons.

Seshadri, K., Akin, B., Laudon, J., Narayanaswami, R., Yazdanbakhsh, A. (2022) ‘An Evaluation of Edge TPU Accelerators for Convolutional Neural Networks’, [online]. Available at: https://arxiv.org/abs/2102.10423

Sevgican, S., Turan, M., Gökarslan, K., Yilmaz, H.B., and Tugcu, T. (2020) ‘Intelligent Network Data Analytics Function in 5G Cellular Networks using Machine Learning’, 326 *JOURNAL OF COMMUNICATIONS AND NETWORKS*, VOL. 22, NO. 3., Digital Object Identifier: 10.1109/JCN.2020.000019

Sewak, M. (2019) ‘Deep Q Network (DQN), Double DQN, and Dueling DQN: A Step Towards General Artificial Intelligence’, *Chapter · June 2019*
DOI: 10.1007/978-981-13-8285-7_8 CITATIONS
https://www.researchgate.net/publication/334070121

Shahroudnejad, A. (2021) ‘A Survey on Understanding, Visualizations, and Explanation of Deep Neural Networks’ [online]. Available at: https://arxiv.org/abs/2102.01792v1

Shankland, S. (2021) ‘Quantum computer maker D-Wave embraces its rivals’ approach’, *CNET* [online] Available at: https://www.cnet.com/tech/tech-industry/quantum-computer-maker-d-wave-embraces-its-rivals-approach/ (Accessed: 5 January 2024)

Sherrer, K. (2021) ‘Kubernetes Alternatives & Competitors’, *Technology Advice* [online]. Available at: https://technologyadvice.com/blog/information-technology/kubernetes-alternatives-competitors/ (Accessed: 5 January 2024)

Shihab, S.A.M & Wei, Peng. (2021) ‘*A Deep Reinforcement Learning Approach to Dynamic Pricing for Airline Revenue Management’*, DOI:10.13140/RG.2.2.18185.57443/1

Shilov, A. (2024) ‘OpenAI CEO Sam Altman will be at Intel's next foundry event and he's currently looking for chip partners’, *Tom's Hardware* [online]. Available at: https://www.tomshardware.com/tech-industry/artificial-intelligence-openai-ceo-sam-altman-will-be-at-intels-next-foundry-event-and-hes-currently-looking-for-chip-partners (Accessed 5 Jan. 2024).

Shinde, S.K., More, J.S., Chaudhari, C.P. (2025) Mastering Drone Technology with AI: BPB Publications, India

Shivakumar, Shailesh Kumar. (2018) ‘Digital Experience Platforms - An Overview: One Platform to manage all Customer Interactions’. Infosys.

Shlezinger, N., Whang, J., Eldar, Y. C., and Dimakis, A. G. (2022) ‘Model-Based Deep Learning’ [online]. Available at: https://arxiv.org/abs/2012.08405

Shoham, Y. (1993) ‘Agent-oriented programming’, *Artificial Intelligence*, 60(1):51–92. [online]. Available at: https://www.semanticscholar.org/paper/Agent-Oriented-Programming-Shoham/552b8a82ca06d73254958cbe16d8dd994e5b5f99

Shor, P.W. (1996) ‘Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer’, *[quant-ph]* [online]. Available at: http://arxiv.org/abs/quant-ph/9508027v2

Shuyang, L., Hu, X. & Yongwen, D. (2016) ‘Deep Reinforcement Learning and Game Theory for Computation Offloading in Dynamic Edge Computing Markets’, *IEEE Access* PP(99):1-1 DOI:10.1109/ACCESS.2021.3109132

Siavoshi, M. (2024)
Simpson’s Paradox: When Aggregated Data Tells a Different Story [online]. Available at: https://www.statology.org/simpsons-paradox-when-aggregated-data-tells-a-different-story/

Silva D. L., Meneguzzi, F., Logan, B. (2020) 'BDI Agent Architectures: A Survey', *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).* [online]. Available at: https://doi.org/10.24963/ijcai.2020/684

Simeone, O. (2022) 'An Introduction to Quantum Machine Learning for Engineers', *Foundations and Trends® in Signal Processing*: Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX
arXiv:2205.09510v4

Sivakumar, N., Mura, C., Peirce, S.M. (2022) 'Innovations in Integrating Machine Learning and Agent-Based Modeling of Biomedical' [online]. Available at: https://arxiv.org/abs/2206.01092

Skansi, S. (2018) *Introduction to Deep Learning – From Logical Calculus to Artificial Intelligence:* Springer International Publishing, AG

www.slideteam.net. (2021) 'Coopetition strategy showing implementation outcomes and process', *Slide Team* [online]. Available at: https://www.slideteam.net/coopetition-strategy-showing-implementation-outcomes-and-process.html. (Accessed 10 Jan. 2024).

Sniezynski, B. (2008) 'An Architecture for Learning Agents', *In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds) Computational Science* – ICCS 2008. ICCS 2008. *Lecture Notes in Computer Science book series*, vol 5103. *Springer, Berlin, Heidelberg* [online]. Available at: https://doi.org/10.1007/978-3-540-69389-5_80

Sorger, S. (2013) *Marketing Analytics.* CreateSpace Publishing, California, US.

sourceforge.net. (2024) 'Best RATH Alternatives & Competitors', *SourceForge.net* [online]. Available at: https://sourceforge.net/software/product/RATH/alternatives (Accessed: 5 January 2024).

Ssengonzi, C., Kogeda, O. P., Olwal, T. O. (2022) 'A survey of deep reinforcement learning application in 5G and beyond network slicing and virtualization' [online]. Available at: https://doi.org/10.1016/j.array.2022.100142

Stewart, D., Loucks, J., Casey, M., Wigginton, C. (2020) 'Bringing AI to the device: Edge AI chips come into their own' [online]. Available at: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2020/ai-chips.html#endnote-45 . (Accessed: 5 January 2024)

Storbacka, K., Nenonen, S. (2009) 'Customer relationships and the heterogeneity of firm performance', *Article in Journal of Business and Industrial Marketing*. DOI: 10.1108/08858620910966246
https://www.researchgate.net/publication/236622007

Stradling, C. (2023) 'Microsoft to make Copilot less reliant on ChatGPT', *Windows Central* [online]. Available at: https://www.windowscentral.com/software-apps/microsoft-looks-to-make-copilot-invulnerable-by-lessening-its-reliance-on-openais-chatgpt (Accessed 5 Jan. 2024).

Sung, Tien-Wen & Tsai, Pei-Wei & Gaber, Tarek & Lee, Chao-Yang Lee. (2021) 'Artificial Intelligence of Things (AIoT) Technologies and Applications', *Editorial, Hindawi Publications*. John Wiley & Sons.

Superannotate.com. (2024) 'Small Language Models (SLMs): Complete Overview', *SuperAnnotate.com* [online]. Available at: https://www.superannotate.com/blog/small-language-models (Accessed 7 Sep. 2024)

Sutton, R. S., Barto, A. G. (2018) *Reinforcement Learning: An Introduction*: MIT Press, Cambridge, MA. Open-Access [online]. Available at: http://incompleteideas.net/book/the-book-2nd.html

Swoop Arrow (2023) 'Swoop Aero Leverages AWS to Deploy a Fleet of Aircraft Across the Globe', AWS [online]. Available at: https://aws.amazon.com/solutions/case-studies/swoop-aero-case-study/

Symeonidis, G., Nerantzis, E., Kazakis, A., Papakostas, G.A. (2022) 'MLOps - Definitions, Tools and Challenges', arXiv:2201.00162v1 [cs.LG]

Sze, V., Chen, YH., Yang, T-J., Emer, J. (2017) 'Efficient Processing of Deep Neural Networks: A Tutorial and Survey', arXiv:1703.09039v2 [cs.CV] [online]. Available at: https://www.arxiv.org/abs/1703.09039v2

Tacchino, T., Macchiavello, C., Gerace D. and Bajoni, D. (2018) 'An Artificial Neuron Implemented on an Actual Quantum Processor', arXiv:1811.02266v1 [quant-ph] 6 Nov 2018

Taherdoost, H., Madanchian, M. (2021) Determination of Business Strategies Using SWOT Analysis; Planning and Managing the Organizational Resources to Enhance Growth and Profitability [online]. DOI: https://doi.org/10.30564/mmpp.v3i1.2748

Tan P.N, Steinbach M., Kumar V. (2012) *Introduction to Data Mining*: 7thed. Pearson Education India.

Tanenbaum, A.S. and Steen, M. V. (2002) *Distributed Systems: Principles and Paradigms:* 1st ed. New York, USA: Prentice Hall.

Takyar, A. (2023) 'How to implement AI in your Business?', *LeewayHertz* [online]. Available at: https://www.leewayhertz.com/how-to-implement-adaptive-ai/ (Accessed: 5 January 2024)

Tarka, Piotr & Maciej Łobiński. (2014) 'Decision Making in Reference to Model of Marketing Predictive Analytics – Theory and Practice', *Management and Business Administration* 22(1):60-69 DOI:10.7206/mba.ce.2084-3356.90

The Hindu. (5th September, 2014). AI key reason for labour income dip, says ILO. Pg 1. Vol. 55, Bo. 213 [online]. Available at: https://newsth.live/fb

TMForum Insight. (2020) 'NWDAF: Automating the 5G network with machine learning and data analytics', *Data Management. inform.tmforum.org* [online]. Available at: https://inform.tmforum.org/features-and-opinion/nwdaf-automating-the-5g-network-with-machine-learning-and-data-analytics
(Accessed: 5 January 2024)

Toloka Team. (2023) 'The history, timeline, and future of LLMs', *toloka.ai* [online]. Available at: https://toloka.ai/blog/history-of-llms/

Torreno, A., Onaindia, E., Komenda, A., Stolba, M. (2017) 'Cooperative Multi-Agent Planning: A Survey', arXiv:1711.09057v1 [cs.AI] 24 Nov 2017

'Transfer Learning' (n.d.). *Wikipedia* [online]. Available at: https://en.wikipedia.org/wiki/Transfer_lrarning

Trinh, H. D., Giupponi, L., Dini, P., (2018) 'Mobile Traffic Prediction from Raw Data Using LSTM Networks', [online]. Available at: https://www.researchgate.net/publication/326773789

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C.B., Farivar, R. (2019) 'Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools', arXiv:1908.05557v2 [cs.LG] 3 Sep 2019

Tsakiris, G., Papadopoulos, Christos., Patrikalos, G., Kollias K-F., Asimopoulos, N and Fragulis, G. F. (2022) 'The development of a chatbot using Convolutional Neural Networks', [online]. Available at: https://doi.org/10.1051/shsconf/202213903009

Upadhyay, A. (2023) 'Implementing RAG with Langchain and Hugging Face', *Medium* [online]. Available at: https://medium.com/@akriti.upadhyay/implementing-rag-with-langchain-and-hugging-face-28e3ea66c5f7

Vaswani, A. et al. (2017) 'Attention is all you need', [online]. Available at: https://arxiv.org/abs/1706.03762

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser L. (2023) 'Attention Is All You Need', [online] Available at: https://arxiv.org/abs/1706.03762

Vicente, J. and Vicente, B. (2019) 'Multi-Agent Systems' [online]. Available at: https://www.researchgate.net/publication/332199176_Multi_Agent_Systems

Visconti, Roberto Moro. (2019) '*THE VALUATION OF ARTIFICIAL INTELLIGENCE'*, DOI:10.13140/RG.2.2.28602.24002

Volpicelli, G. (2023) 'ChatGPT broke the EU plan to regulate AI', *Politico.eu* [online]. Available at: https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/ [Accessed: 23 Apr 2024].

Waehner, K. (2021) 'Kappa Architecture is Mainstream Replacing Lambda' [Online]. Available at: https://www.kai-waehner.de/blog/2021/09/23/real-time-kappa-architecture-mainstream-replacing-batch-lambda/#:~:text=Real%2Dtime%20data%20beats%20slow . (Accessed: 5 January 2024)

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam S. et al. (2023) 'Diffusion Model Alignment Using Direct Preference Optimization' [online]. Available at: https://arxiv.org/abs/2311.12908

wandb.ai. (202x) 'Understanding LLMOps: Large Language Model Operations' [online]. Available at: https://wandb.ai/site/articles/understanding-llmops-large-language-model-operations

Wang, Shiqiang & Tuor, Tiffany & Salonidis, Theodoros & Leung, Kin K. & Makaya, Christian & He, Ting & Chan, Kevin. (2019) 'Adaptive federated learning in resource constrained edge computing systems', *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205- 1221.

Wang, W-C, Lin, C-H, Chu, Y-C. (2011) ‘Types of Competitive Advantage and Analysis’, *International Journal of Business and Management* Vol. 6, No. 5; doi:10.5539/ijbm.v6n5p100

Warren, T. (2023) ‘Microsoft extends OpenAI partnership in a ‘multibillion dollar investment’, *The Verge* [online]. Available at: https://www.theverge.com/2023/1/23/23567448/microsoft-openai-partnership-extension-ai (Accessed 5 Jan. 2024).

WEF and Kearney. A. T. (2017) ‘Technology and Innovation for the Future of Production: Accelerating Value Creation’, *World Economic Forum* [online]. Available at: https://www.weforum.org/whitepapers/technology-and-innovation-for-the-future-of-production-accelerating-value-creation

Weiss, K., Khoshgoftaar, T. M., Wang, D-D. (2016) ‘A survey of transfer learning. Journal of Big Data’ [online]. Available at: https://doi.org/10.1186/s40537-016-0043-6 and at: https://journalofbigdata.springer.com/articles/10.1186/s40537-016-0043-6

Wiles, J. (2023) ‘Beyond ChatGPT: The Future of Generative AI for Enterprises’. *Gartner* [online]. Available at: https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises (Accessed: 5 January 2024)

Wu, S., Fei, H., Qu, L., Ji, W. and Chua, T-S. (2023) ‘NExT-GPT: Any-to-Any Multimodal LLM’, arXiv:2309.05519v2 [cs.AI] 13 Sep 2023

www.dwavesys.com. (2023) ‘Quantum Computing’, *dwavesys.com* [online] Available at: https://www.dwavesys.com/learn/quantum-computing/ (Accessed: 5 January 2024)

www.dwavesys.com. (n.d.) ‘The Advantage^{TM} Quantum Computer’, *D-Wave*. [online]. Available at: https://www.dwavesys.com/solutions-and-products/systems/ (Accessed: 5 January 2024)
www.fortunebusinessinsights.com. (2022) ‘Edge AI Market Size, Share & Forecast Report [2022-2029]’, [online]. Available at: https://www.fortunebusinessinsights.com/edge-ai-market-107023 (Accessed: 5 January 2024)

www.run.ai. (20xx) ‘Kubeflow Pipelines: The Basics and a Quick Tutorial’, *run.ai* [online]. Available at: https://www.run.ai/guides/kubernetes-architecture/kubeflow-pipelines-the-basics-and-a-quick-tutorial (Accessed: 5 January 2024)

Xie, J., Liu, C-C. (2017) ‘multi-agent systems and their applications’, *Journal of International Council on Electrical Engineering*, 7:1, 188-197, DOI: https://doi.org/10.1080/22348972.2017.1348890

Xu, Feiyu & Uszkoreit, Hans & Du, Yangzhou & Wei Fan. (2019) ‘Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges’, DOI:10.1007/978-3-030-32236-6_51 & In book: Natural Language Processing and Chinese Computing (pp.563-574)

Y-Combinator (2023) ‘Y Combinator’, *Ycombinator.com*. [online]. Available at: https://www.ycombinator.com/ (Accessed 5 Jan. 2024).

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y. (2023) ‘REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS’, [online]. Available at: https://arxiv.org/abs/2210.03629

Yang, Q., Liu, Y., Chen, T., Tong, Y. (2019) ‘Federated Machine Learning: Concept and Applications’ [online]. Available at: https://arxiv.org/abs/1902.04885v1

Yang, R., Sun, X. and Narasimhan, K. (2019) ‘A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation’, [online]. Available at: https://arxiv.org/abs/1908.08342

Yang, Y. and Wang, J. (2021) ‘An Overview of Multi-agent Reinforcement Learning from Game Theoretical Perspective’ [online]. Available at: https://arxiv.org/abs/2011.00583

Yannis, F-B. (2019) ‘The Promise of Hierarchical Reinforcement Learning’, *The Gradient* [online]. Available at: https://thegradient.pub/the-promise-of-hierarchical-reinforcement-learning/

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, X., Chen, E. (2023) ‘A Survey on Multimodal Large Language Models’ [online]. Available at: https://arxiv.org/abs/2306.13549

Yu, F., Xiu, X., Li, Y. (2022) ‘A Survey on Deep Transfer Learning and Beyond’, *Mathematics* 2022, 10, 3619 [online]. Available at: https://doi.org/10.3390/math10193619

Zelikman, E., Wu, Y., Mu, J., Goodman, N.D. (2022) ‘STaR: Self-Taught Reasoner Bootstrapping Reasoning with Reasoning’ [online]. Available at: https://arxiv.org/abs/2203.14465

Zhang, Hao Lan & Lau, Hoong Chuin. (2014) ‘Agent-based problem-solving methods in Big Data environment’, *Web Intelligence and Agent Systems* 12(4):343-345, DOI:10.3233/WIA-140300

Zhang, A., Lipton, Z.C., Li, M. and Smola, A.J. (2022) ‘Dive into Deep Learning’ [online]. Available at: https://arxiv.org/abs/2106.11342

Zhang, B., Chen, B., Yang, J., Yang, W. & Zhang, J. (2018) ‘A Unified Intelligence-Communication Model for Multi-Agent System PartI: Overview’, *Pre-print.* [online] Available at: https://www.researchgate.net/publication/329236521

Zhang, K., Yang, Z. and Basar, T. (2021) ‘*Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms’,* [online]. Available at: https://arxiv.org/abs/1911.10635v2

Zhang, Z, Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A. (2023). ‘Multimodal Chain-of-Thought Reasoning in Language Models'. [online]. Available at: https://arxiv.org/abs/2302.00923v4

Zhengxin, F., Yi, Y., Zhang, J. Liu, Y., Mu, Y., Lu, Q. et al. (2023) ‘MLOps Spanning Whole Machine Learning Life Cycle: A Survey’ [online]. Available at: https://arxiv.org/abs/2304.07296

Zhou, M., Luo, J., Villella, J., Yang, Y. Rusu, D., Miao, J. et al. (2020) ‘SMARTS: Scalable Multi-Agent Reinforcement Learning Training School for Autonomous Driving’ [online]. Available at:
https://arxiv.org/abs/2010.09776

Zhu, X., Xu, J., Ge, J., Wang, Y., Xie, Z. (2023) ‘Multi-Task Multi-Agent Reinforcement Learning for Real-Time Scheduling of a Dual-Resource Flexible Job Shop with Robots’, *Processes* [online]. Available at: https://doi.org/10.3390/pr11010267

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q. (2020) ‘A Comprehensive Survey on Transfer Learning’ [online]. Available at: https://arxiv.org/abs/1911.02685v3