Towards Pervasive and Trustworthy Artificial Intelligence

How standards can put a great technology at the service of humankind

By

Alessandro Artusi, Andrea Basso, Marina Bosi, Sergio Canazza, Leonardo Chiariglione, Miran Choi, Fabiano Columbano, Mert Burkay Çöteli, Nadir Dalla Pozza, Roberto Dini, Michelangelo Guarise, Hüseyin Hacıhabiboğlu, Roberto Iacoviello, Chuanmin Jia, Jisu Kang, Panos Kudumakis, Valeria Lazzaroli, Marco Mazzaglia, Guido Perboli, Niccolò Pretto, Paolo Ribeca, Mariangela Rosano, Mark Seligman

MPAI

18 December 2021

Preface

With the printing industry sparing no efforts publishing books on Artificial Intelligence (AI), why should there be another that, in its title and subtitle, combines the overused words AI and trustworthy, with the alien words standards and pervasive?

The answer is that the book describes a solution that covers all the elements of the title: to effectively combine the AI and trustworthy words, but also to make AI pervasive. How? By developing standards for AI-based data coding.

Many industries needed standards to run their business and used to have high respect for them. The MP3 standard put users in control of the content they wanted to enjoy, and the television – and now the video – experiences have little to do with how users approached audio-visual content some 30 years ago.

At that time, the media industry was loath to invest in open standards. The successful MPEG standards development model, however, changed its attitude. Similarly, the AI industry has been slow in developing AI-based data coding standards making proprietary solutions their preferred route.

This book provides the full description of mission, achievements and plans of the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) standards developing organisation. It describes how MPAI develops standards that can also be used, how standards can make AI pervasive and promote innovation, how MPAI gives users the means to make informed decisions about how to choose a standard implementation having the required level of trustworthiness.

MPAI is a unique adventure open to those who want to make the MPAI vision real.

Contents

1 Introduction

As it often happens in research, a technology that had attracted the interest of researchers decades ago and stayed at that level for a long time, suddenly comes into focus. This is the case of the collection of different technologies called Artificial Intelligence (AI). Although this moniker might suggest that machines are able to replicate the main human trait, in practice such techniques boil down to algorithmically sophisticated pattern matching enabled by training on large collections of input data. In this book we will consider Machine Learning (ML) as part of AI. Embedded today in a range of applications, AI has started affecting the life of millions of people and is expected to do so even more in the future.

AI provides tools to "get inside" the meaning of data to an extent not reached by previous technologies. In this book we use the word "data" to indicate anything that represents information in digital form ranging from the US Library of Congress to a sequenced DNA, to the output of a video camera or an array of microphones, to the data generated by a company. Through AI, the number of bits required to represent information can be reduced, "anomalies" in the data discovered, and a machine can spot patterns that might not be immediately evident to humans.

AI is already among us doing useful things. There is keen commercial interest in implementing more AI-centric processes unleashing its full potential. Unfortunately, the way a technology leaves the initial narrow scientific scope to become mainstream and pervasive for products, services and applications is usually not linear nor fast. However, exceptions exist. Looking back to the history of MPEG, we can see digital media *standards* not only accelerated the mass availability of products enabled by new technologies, but also generated new products never thought of before.

In fact, the MPEG phenomenon was revolutionary because its standards were conceived to be industry neutral, and the process unfolded successfully because it had been designed around this feature. The revolution, however, was kind of "limited" because MPEG was confined to "media" (even though it tried to escape from that walled garden).

This book concerns itself with AI-centric data coding standards, which do not have such limitations. AI tools are flexible and can reasonably be adapted to any type of data. Therefore, as digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar, if not stronger impact. Research shows that AI-based data coding is generally more efficient than existing technologies for, e.g., data compression and description.

These considerations have led a group of companies and institutions to establish the Moving Picture, Audio and Data Coding by AI – MPAI – as an international, unaffiliated not-for-profit Standards Developing Organisation (SDO).

However, standards are useful to people and industry if they enable open markets. Still, the industry might invest hundreds of millions into the development of a standard, only to find that it is not practically usable or it is only accessible to a lucky few. In this case rather than enabling markets, the standard itself causes market distortion. This is a rather new situation for official standards, caused by the industry's recent inability to cope with tectonic changes induced by technology and market. As a result, developing a standard today may appear like a laudable goal, but the current process can actually turn into a disappointment for industry. A standards development paradigm more attuned to the current situation is needed.

For this reason, the MPAI scope of activity goes beyond the development of standards for a technology area. It includes Intellectual Property Rights guidelines to compensate for some standards organisations' shortcomings in their handling of patents.

While in the rest of the book there will be opportunities to go more in depth into the nature of AI, it is appropriate for this introduction to briefly compare how the incumbent Data Processing (DP) technology and AI work. When they apply DP, humans study the nature of the data and design a priori methods to process it. When they apply AI, prior understanding of the data is not paramount – a suitably "prepared" machine is subjected to many possible inputs so that it can "learn" from the actual data what the data "means".

In a sense, the results of bad training are similar in humans and machines. As an education with "bad" examples can make "bad" humans, a "bad", i.e., insufficient, sectorial, biased etc. education makes machines do a "bad" job. The conclusion is that, when designing a standard for an AI-based application, the technical specification is not sufficient. So, MPAI's stated goal to make AI applications interoperable and hence pervasive through standards is laudable, but the result is possibly perverse if ungoverned "bad" AI applications pollute a society relying on them.

For these reasons, MPAI has been designed to operate beyond the typical remit of a standards-developing organisation – albeit it fulfills this mission

quite effectively, with five full-fledged standards developed in 15 months of operation. An essential part of the MPAI mission consists of providing the users with quantitative means to make informed decisions about which implementations should be preferred for a given task.

In conclusion, this book will talk about AI, what it is, which tools it offers, which applications it makes possible and how MPAI delivers AI-based standards. Thanks to MPAI, implementers have available standards that can be used to provide trustworthy products, applications and services, and users can make informed decisions as to which one is best suited to their needs. This will result in a more widespread acceptance of AI-based technology, paving the way for its benefits to be fully reaped by the society.

The book is organised in three sections:

Section 1 – "AI opportunities" describes the current state of the fields in which MPAI currently plays a role. It contains the following chapters:

Chapter 2	Introduces the notions of information, data, DP and AI.
Chapter 3	Describes the wide used of AI and some issues arising from it.
Chapter 4	Gives basic AI and ML notions referenced by the book.
Chapter 5	- Visual humans and machines
Chapter 6	- Humans conversing with machines
Chapter 7	- Conversing with machines
Chapter 8	- Audio for humans
Chapter 9	- Visual for humans
Chapter 10	- Non-media data
Chapter 11	Introduces regulation trends in AI

Section 2 – "Using AI for the better" describes how standards allow us to get the benefits of AI and avoid its pitfalls. It contains the following chapters:

Chapter 12	Introduces the notion of and benefits from an AI framework.
Chapter 13	Describes the first MPAI data coding standards developed.
Chapter 14	Presents components and structure of MPAI standards.
Chapter 15	Illustrates some technologies specified in MPAI standards.

Section 3 – "AI needs more than standards" reiterates the need to complement AI standards with additional measure. It contains the following chapters:

The authors of this book thank the MPAI members and the community for making the MPAI mission real, Philip Merrill for his assistance editing and improving the way this book conveys the value of MPAI and Renato Valentini for his unstinting support.

Section 1 - AI opportunities

2 Data and data processing

Datum (plural: data), often used in singular form as data, can be defined as information that is available for processing. Our body can feel the temperature of the ambient air, but it is not uncommon that two persons have different feelings regarding the ambient temperature. If we use a thermometer, however, we can get an "absolute" measure of the temperature taken at a given point and at a given time. We can say that the temperature measured by the thermometer is now "data" that can be used, e.g., to correlate with the temperature at other places and at different times. We can also say that time and geographical coordinates are "metadata" of the temperature "data", but we can also say that geographical coordinates are data, and time and temperature are metadata depending on the purpose of the processing.

Humans are good at processing data. Given a table or a stream of data, an experienced human may discover correlations that other humans do not discern. This type of processing happens at a large scale at a Stock Exchange where experienced traders make decisions to buy, hold or sell company shares, based on the data flowing in and the experience, i.e., data stored in their brains.

No matter how good certain humans can be at processing data, they have limits: they only work a certain amount of time with sufficient performance before they take a rest; their capability to ingest different streams of data is limited; their ability to "retrieve" data is constrained by their experience; their ability to correlate data is limited by the speed of the electro-chemical technology used in their brains (a few tens of kHz). While it is possible in certain cases to parallelise certain type of data processing (just think of the administrative offices of large companies 50 years ago), in other – especially time critical – cases, parallelisation is not possible unless the speed of the processing technology is increased.

The mainframe computers of the 1960s and 1970s, with their large storage and processing capabilities – as seen with the eyes of that time – essentially uprooted the way administrative offices processed data.

Figure 1 - Old style administrative office

Humans were required to teach, i.e., program, machines to process data and humans were still needed to process the data resulting from the processing of machines, again to discover correlations. Application programs such as VisiCalc and their successors helped spread the notion and consolidate the practice of machines processing data for further processing by humans.

Photography was a great tool to distribute information, as wase telephony, facsimile, radio, and television. Great ingenuity was required, however, to convert such information into data: as long as image, voice, audio and video signals were “analogue”, however, it was hard to get “data” from them. Some may still remember different “meters” attached to device clamps, providing data about voltage and current from which humans could infer something to their interest.

The great phenomenon started in the late 1920s by Harry Nyquist et al. at the Bell Labs progressed through a series of milestones that would be too long to recall here, and eventually gave rise to the world of digital media as we know it. If today we are inundated by data, however, it is because the enormous streams of data measured in kbit/s, then Mbit/s, today Gbit/s and tomorrow Tbit/s were, are and will be reduced in size by exploiting the “inner structure” of those data streams.

Human ingenuity achieved that. By resorting to different tools, little by little humans could dig into those data streams and discover ways to avoid sending useless parts, or if necessary, “sacrificing” some parts because their

removal did not seriously affect the intelligibility and usability of the resulting data.

As soon as media information could be converted into data, humans could apply their ingenuity to teach machines to extract meaningful information or data. This was done for speech, e.g., extracting what the speaking human is saying word by word, but also understanding the "meaning" of the words put together and even attempting to provide a reply to the sentence. Similar levels of human-like understanding of other types of data such as visual have also been achieved.

In the old-style administrative offices, humans used to perform tasks that were eventually taken over by mainframes to a large extent. The kind of minute investigations that have allowed the processing of data to achieve the present-day results are in line for a similar eventual fate. Humans are excellent at discovering correlations between data because they have a powerful "processor" and vast amounts of data resulting from years of experience. However, technology makes it possible to develop machines that have a similar learning and processing capability. They are not constrained to operate with a kHz clock but can work (today) with a GHz clock, the amount of "experience" they can ingest and process data without being constrained by working hours or by the opportunities that life gives to humans.

Unlike DP, i.e., the human programming of machines to perform operations that it would be too tedious or impractical or even impossible for humans to do, AI is the programming of machines to perform operations that require a human-like level of intelligence and discernment. ML, a subset of AI, is the programming of machines so that they can learn and adapt without following explicit instructions. A large amount of data is provided to an ML algorithm and let it explore the data and search for a model achieving what the programmers have set out to achieve.

AI is the technology that enables the transformation of data to suit the needs of an application. It can process financial data to provide Key Performance Indicators (KPI), it can convert data in the form of a stream of speech samples to words that a human can read, it can process characters expressing words and extract the meaning of those words, it can name and characterise the objects in a visual scene and can reduce the number of bits needed to transmit a high-definition video stream. It is doing this today and promises to do more in the future thanks to a large investment made by a sizeable share of worldwide research and academia stakeholders.

The same data used by a human is often needed by other humans who want them in a form to be understood by them. The role of MPAI standards is to do exactly that for machines: to define a *standard data representation* that is suitable for processing by *AI technologies*.

There is one proviso, though, because the data processing technologies that have had decades of track records are still in the process of passing their baton to their successor AI technologies. The transition is not going to happen in an instant because there are excellent data processing technologies that have been honed for decades whose life cycle is not over. MPAI focuses on data coding by AI, but its standards are meant to serve industry needs, not an abstract principle. Whenever desirable, traditional data processing-based standards will be included next to AI-based standards. MPAI standards will also support integration of data processing-based and AI-based data coding.

3 AI-potential and drawbacks

Some may not have realised it, but many AI applications already permeate everyday life and will be able to transform practically all aspects of our life and the economy in general. This is shown in Figure 2 and some of the fields indicated there are analysed in the following.

- *Health*: analyse large amounts of medical data and discover matches and patterns to improve diagnosis and prevention; develop programs to respond to emergency calls by recognising a cardiac arrest faster than a human operator; develop multilingual textual research tools that will make it easier to find more relevant medical information available.
- *Transport*: improving the safety, speed, and efficiency of rail traffic, and the initial use of AI for autonomous driving.
- *Industry*: using robots to bring factories back; sales channel planning and predictive maintenance; collaborative and augmented reality systems to increase worker satisfaction in smart factories
- *Agriculture and food supply chain*: building sustainable food systems through forms of analytics and monitoring to minimise fertiliser, pesticide, and irrigation use, helping productivity, and reducing environmental impact; monitoring movement, temperature and feeding of livestock.
- *Public administration and services*: efficiency of natural disaster alert systems to enable prevention, preparedness, and resilience.

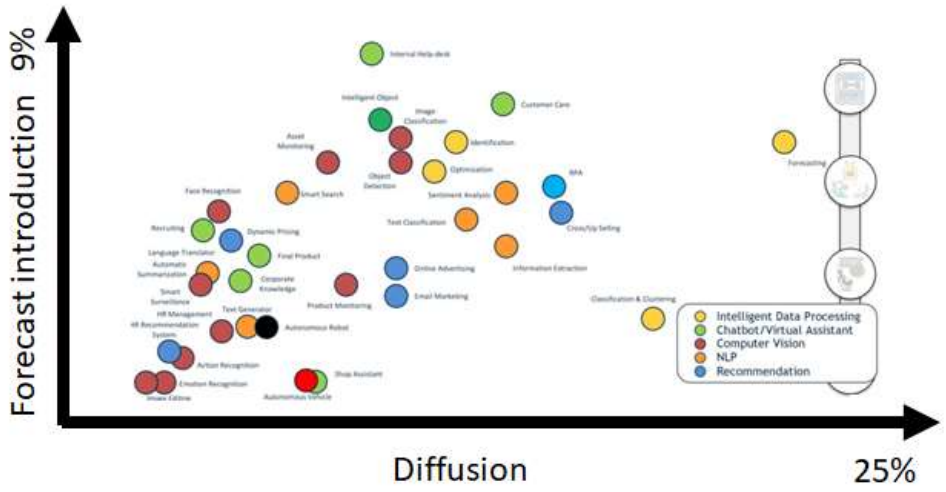

Figure 2 - AI application maturity (source: osservatori.net)

3.1 Application of AI to traditional IT

AI was born in the 1950s, but it is only today that technological advances in computing power, data availability, and the ability of data analysis to solve complex problems have triggered the creation and dissemination of AI applications. The basic technologies are mature and, through APIs and cloud services, available at an affordable cost. However, a design approach is needed to introduce AI into processes. If up to 10 years ago the barriers to AI introduction in companies were missing tools or inadequate analytical skills, most issues today are not technological, but cultural and the lack of specific skills. According to experts, today 70% of the effort related to an AI project is for process redesign, 10% to algorithms development and only 10% to technology.

AI and ML are therefore lightening the workloads of help desk, cybersecurity, and other typical IT tasks. In 12 out of 13 major vertical industries, the segment that makes the most use of AI is IT, with more than 46% of IT teams at large companies integrating AI into their applications. Therefore, AI technologies present significant opportunities for IT professionals. The ability to implement AI technology and integrate it with other tools and services to achieve maximum business value opens new career paths. But even at the most basic level, AI frees IT professionals from repetitive tasks by allowing them to focus on something of higher value.

But what are the distinctive capabilities of an AI system compared to traditional DP? A traditional system basically performs two functions: data storage and processing, is done by developed and preset programs operating on more or less complex deterministic algorithms and formulas acting on

structured databases. On the contrary, an AI system, contrary to the more traditional concept of programming, can process structured *and* non-structured data to formulate hypotheses based on a knowledge domain on which the system is trained. From the point of view of intellectual abilities, the functioning of an AI is mainly substantiated through four different functional levels:

- *Understanding*: through the simulation of cognitive capacities of correlation of data and events, AI can recognise texts, images, tables, videos, voice and extrapolate information.
- *Reasoning*: through logic the systems can connect data collected from multiple sources, through mathematical algorithms and in an automated way.
- *Learning*: systems with specific features for data input analysis and "correct" return in output. This is the classic example of ML systems.
- *Interaction*: the way AI works in relation to humans, e.g., Natural Language Processing (NLP), based on technologies that allow human to interact with machines using natural language enables virtual assistants and chatbots.

3.2 AI as a business opportunity

The introduction of AI as a tool for business development is a relatively recent topic for small/medium enterprises, although it has been outstanding for quite a few years. Apart from some specific realities that make this subject their core business (e.g., companies operating in high-tech or in specific sectors, such as those who produce machinery for diagnostics, etc.), most companies find themselves having to consider the adoption of AI by intuiting (correctly) the great potential underlying the technology but failing to identify the contours and, above all, the potential benefits.

Talking to entrepreneurs helps us understand that there are two main questions: *first*, how AI can help their business, and *second*, what they need to implement it correctly assessing time, costs, and benefits in the long run. As already stated above, the topic of AI and its impact on society and economic and industrial systems is lively and current, representing, without a doubt, one of the most promising technological facilitators of the digitalisation process. In the current phase of the industrial revolution, a process of digital transformation is taking place whose objective is to provide companies with organisational models and work organisations able to rapidly and continuously introduce new technologies to support process innovation of support entire sectors.

Manufacturing, assembly, and distribution systems have always produced enormous amounts of data and today, thanks to intelligent processing and

analysis systems, can be used productively, along the entire value chain and taking on an increasingly decisive guiding role in corporate decision-making systems. The introduction of AI systems in a business process can be compared with the impact that the introduction of mechanisation in industries had in the mid-eighteenth century. A decisive step in many productive sectors that has improved the conditions of workers engaged at the time in repetitive or exhausting work, replacing the human force, or animal, with the mechanical one, generated by combustion from a motor machine, originating a whole series of economic and social changes.

Therefore, AI has the objective, or rather the ambition, to emulate the cognitive abilities of the human being, increasing its effectiveness. A traditional computer system used to perform complex operations and record and store huge amounts of data, performing both tasks at levels that no human mind can approach, both in terms of speed and accuracy. Today, AI can express extraordinary new abilities (such as emulating reasoning, analysing unstructured data, interpreting language, reading images, reading text, being able to reason in probabilistic terms...) that can give the individual, worker and citizen new capabilities and enormous advantages in work or personal life.

AI will change the relationship between human and machine for the better in a perspective of "collaborative intelligence" and this will offer many opportunities. In this sense, as information technology and the processing capabilities of computers and information analysis systems continuously progress, a world in which machines will replace some of the activities now performed by humans is easy to imagine.

For a long time, economists have been wondering what tools to activate to prevent society from evolving towards an increasingly labour-intensive economy – whose evolution is today accelerated by AI – resulting in an impoverishment of the population. Alongside these social issues, there are ethical questions about the development and evolution of AI and new technologies. The fears might seem excessive but underestimating the impact of AI could be the number one risk. In fact, despite the rapid ongoing evolution, machines as well as expert systems will continue to be at the humans' side, as assistants with respect to the tasks to be performed, whether they are in a factory or integrated into our daily lives.

3.3 AI: a very wide field

Today, there are several noteworthy use cases in many fields and industries such as healthcare (Intelligent Health), finance (Financial Trading), automotive (Connected Cars, Autonomous Vehicles), as well as scenarios that are now commonly used such as security and emergency services (Image Processing, Computer Vision, Object Detection and Recognition) and

also with regard to marketing personalisation (Natural Language Processing for Sentiment Analysis).

If we move on to AI in the industry, for some time now automation has brought robots to perform complex jobs at very high efficiency in assembly lines, but these robots are often made to perform a single task and the cost of reprogramming is very high, if even possible. In fact, this is automation, not AI, but we are already thinking about and designing the first adaptive and collaborative robots, equipped with AI that can learn different tasks through learning by demonstration and with hardware that is better suited for task reconfigurability. Adaptive manufacturing is also strategic for industry and requires AI solutions that can adapt to different human scenarios and flexible IT infrastructures that can adapt accordingly with the people.

In manufacturing plants, ML is increasingly being used for anomaly management and predictive maintenance. The next step is to have production systems able to intervene autonomously based on experience and that are not limited to intervening on the system when thresholds are exceeded, and according to pre-established rules, but that learn from previous analyses and create the representation of the production process according to "non-programmed" variables. This is the type of intelligence required for the interaction of complex integrated automated systems, for example to ensure the operation of a highly automated extended supply chain and is one of the main areas of investment for many companies.

Indeed, there are use cases and industry scenarios where AI brings tangible and quantifiable benefits such as, for example, workforce support (skills rotation), process automation (working capital), customer management (customer loyalty) and product innovation (servicing of products). Now we should realise that over the next few years, it will be fundamental to clarifying and regulate the mission and areas of use, especially from an ethical perspective.

Trust toward AI will be earned over time as it is with human relationships. It will need to be demonstrated that it is not a "humans vs. machines" struggle by moving beyond fears and clarifying doubts about adoption through actual accomplishments. AI can complement the work of humans, increasing judgment and analysis and skills, allowing resources to be focused so as to significantly expand ingenuity, creative effort and experience, and improve speed, scope and efficiency.

The correlation with AI seems to be clear if we start from 6 technological areas, widely reported in the literature:

A. *Automation*: in this context, this term is used to indicate any solution that allows humans to be relieved of repetitive actions, even complex ones, but which are easy to be implemented by a computer or mechanical system. Often, we talk about "industrial automation" referring to technolo-

gies that allow the automation of part or of full production processes. We can associate the term automation also in non-industrial contexts, as in service delivery processes, logistics and asset tracking. The introduction of AI opens new frontiers in the field of automation, making it possible to address processes even with a complex decision-making component.

B. *Computerisation*: the introduction of software applications and necessary infrastructure, computers, servers, and networks within a company to automate or make processes more efficient. Computerisation is a basic requirement for exploiting AI potential, because it leads to producing and collecting structured and unstructured data from which AI can extract useful information to improve processes or define new business models.
C. *Dematerialisation*: the replacement of paper with digital documents, not only because of the computerisation process, but as a driver to revisit business processes and make them more efficient, effective, and secure. Tracking tools and the protection of access rights to digital documents, also in relation to the new European General Data Protection Regulation (GDPR) regulation, are now an essential tool.
D. *Virtualisation*: in general, it can be defined as the set of technologies that makes the most of the processing capabilities of a hardware system. In a more technical way, the ability to abstract the hardware resources (CPU, RAM, and Storage) makes them available to the software in a virtual way. With virtualisation technologies, a single physical server can simulate the execution of multiple virtual servers that work simultaneously using hardware more efficiently.
E. *Cloud Computing & Big Data*: by Cloud Computing we mean the ability to use hardware and software resources with a "pay-per-use" logic. The potential of cloud computing goes beyond mere economic efficiency because it can deal with sudden increases in processing needs, mechanisms to increase the availability of services (Disaster Recovery and Business Continuity) and guaranteed levels of support for the application context of reference. Big Data is the set of technologies and methodologies able to process huge amounts of structured and unstructured data and to extract useful information for business decision-making processes.
F. *Mobile*: one of the most disruptive changes of recent years is the spread of mobile devices in daily life and business. The evolution of wireless telecommunications networks and the upcoming introduction of 5G in mobile telephony represent some of the most important factors in the digital transformation processes of companies. The impact in the business world is extended to all areas, e.g., smart working and the Internet of Things. AI is already a reality, e.g., virtual assistants that are now on all smartphones and the facial recognition techniques to protect access to devices.

3.4 Dealing with AI: high level of expertise

Introducing AI in a company does not necessarily mean using technology accessible only to a few experts or burdening the existing infrastructure (which typically manages the day-by-day business) with onerous changes, but rather relying on existing data and infrastructure to maximise its value. One of the general initiatives in AI is the simple virtual assistant or chatbot, which can be used in the company without IT department support. Complete predictive maintenance systems collect real-time data on the operation of machines and can predict, for example, production breaks or irregularities, thus enabling timely actions across the entire chain of material, spare parts, movements etc. Therefore, the system understands the features to be introduced and how to manage over time this new way of governing business processes.

AI is an interdisciplinary phenomenon where, alongside technical figures who are experts in specific disciplines such as data science, in general transversal figures such as psychologists, anthropologists, sociologists, linguists, and humanists play fundamental roles. They should be able to improve the interaction between AI and its users, which will become increasingly complex because they can handle language, and emotions from face, gestures, and body.

The world of work is already affected and will be more so in the future by a profound transformation and, in the short term, we will see the emergence of new professions, while the existing ones will be extensively modified by the introduction of new processes and methodologies. Above all, manufacturing companies are called upon to develop technical skills to allow work on appropriate manufacturing processes and products by implementing pilot initiatives, in a short time and with limited resources. In addition, each worker needs to develop attitudes related to their ability to work in a context where people and machines are connected, and continuous learning throughout their working life is a priority.

This is accompanied by the ability to experiment, find, and learn independently to carry out one's own activities and experiments, both in a team context and independently, as one of the particularities of AI is precisely democratising access to technology for all individuals. Among the new professions that will be created by the increasing adoption of AI within production processes, we can identify three: Trainers, Explainers and Sustainers [1].

- *Trainers*: are in charge of correcting and addressing AI-based services when interacting with humans in complex and frustrating situations, to add understanding and empathy to the conversation.
- *Explainers*: fill the gap between technologists and business leaders in the understanding of highly complex systems considered as "black boxes"

because they hide the logic with which they suggest actions. The ability to analyse the rationale that led to a potentially harmful suggestion will be required.
- *Sustainers*: ensuring that the systems behave according to the specifications based on which they were designed and trained, taking immediate corrective action in case of abnormal activities.

3.5 Defining standards for AI use

To enable a single digital market, standards for AI uses that are ethical, democratic, and inclusive are required. These standards should include rules for a safe and reliable AI where obligations, requirements and parameters are defined and universally shared. The goal is to arrive at the determination of the most transparent uses possible, validated in advance for market access according to requirements. Guidelines with mandatory steps must determine *AI systems* according to the following parameters:

- Risk assessment and mitigation systems.
- High quality data sets to train the AI system.
- Recording of AI systems activities to ensure traceability of results.
- Detailed documentation, providing all necessary information about the system and its purpose, so that authorities can assess its compliance and provide information to users that is clear and to the point.
- Adequate human surveillance to minimise risk is provided.

3.6 The right to know it is AI

The development of AI and automated decision-making does present challenges to consumer trust and well-being: when interacting with these systems, consumers should be properly informed so that they can decide about their use. Relying on AI carries risks, especially when it has the power to make decisions without human supervision. ML is based on training that relies on specific data sets. However, the data sets can reflect social biases, and in that case, AI incorporates those same biases into its own decision-making process. AI is increasingly being used in the design of decision-making algorithms. Decisions made by algorithms can have a significant impact on people's lives: from granting credit, getting a job or medical care to influencing the outcome of court judgments.

In some cases, automated decision-making processes risk perpetuating the social gap. Some algorithms, for example, have been shown to discriminate against women: these are AI systems used in the human resources departments of some companies that give priority to male or promotion over female employees because of historical biases in the data they use to decide.

Therefore, authorities of many countries have acted, providing legislative guidelines aimed at ensuring that:

- Consumers are protected from unfair and/or discriminatory business practices, or risks arising from commercial AI services.
- AI-based decision-making processes are transparent.
- Only non-discriminatory, high-quality data are used in automated decision-making systems.

3.7 Is AI a new speculative bubble?

When a phenomenon like AI grows so much, there is always fear of a "new winter" with the accompanying bursting of the bubble, i.e., a period in which funds and interest in the sector vanish.

Since the 1950s, research in the sector has followed a regular pattern: moments of enthusiasm followed by periods of mistrust. The term "AI winter" appeared for the first time in 1984 and used to explain the 1970s decline in funding. A few years later, in fact, the industry began to collapse, making the 1980s a long "winter".

Today, however, conditions have totally changed and the spectre of a decline in the sector seems almost impossible. There are those maintaining that all the promises made so far are in fact illusions, next to those who reel off a series of data proving the contrary. If we gather top experts' opinion in the field, though, this winter seems far from coming. This doesn't mean that researchers around the world ignore the current debate, but rather look at it considering data. The World Economic Forum (WEF) mentions that AI specialist and ML specialist are among the leading jobs of the next five years. The WEF further estimates that public funding from now until 2025 in the United States will be over six billion USD, while in China it already exceeds ten billion USD in annual investment. All of this bodes well for the future and should remove fears of another AI winter, where the only real risk is underestimating AI technology.

4 Machine Learning and Neural Networks

ML is an area of AI which has emerged as a leading tool for data analysis because of its ability to learn directly from raw data, with minimal human intervention [2]. The main scope of a ML application is to automatically detect meaningful patterns in data, to make subsequent predictions about new data [3]. ML comprehends several strategies to reach this goal, which are called *learning models*. In this chapter, after a short introduction on how the brain works, we will introduce recent Deep Learning (DL) approaches that are used throughout the book and the basic concepts of one of the most common learning models, the traditional Artificial Neural Networks (ANNs).

The human brain could be simply described as a network of specialised cells called *neurons*, connected with each other by "cables" called *axons*. Through them, an electrical impulse can travel to other neurons at the speed of ~100 m/s. At the termination of an axon – called *synapsis* – the passing of information between the synapsis and the "receivers" of the next neuron – called *dendrites* (or cell bodies) – is enabled by a chemical process taking place in a 20-40 nanometre-wide gap between the synapsis and the dendrites.
Humans can conduct intellectual activities because their brain is made of ~100 billion neurons interconnected by trillions of synapses. Interestingly, and confirming the value of our intellectual activities, the operation of the brain is costly in terms of energy as the brain uses up to 20% of the energy consumed by the body in normal activities.
A machine that displayed human capabilities has been dreamed by many, but it was only with the advent of electronic computers that the research could take the road that has led us this far.

4.1 Learning paradigms

Many "learning paradigms" have been proposed and are in actual use for specific purposes.
Supervised Learning is the task of identifying a function that maps certain input values to the corresponding output values. This paradigm requires the learning model to be "trained" by using known input and output sets. An example application is found on the ANNs considered in this book.
Unsupervised Learning identifies patterns in datasets whose data are neither classified nor labelled. It is a convenient process whenever the dataset is large, since the annotation process would be costly. An example can be found in the Principal Component Analysis (PCA), which is a well-known method used to identify patterns in data sets by exploiting a linear transformation of the axes in the principal directions, and therefore it can help reducing the number of redundant features. Being it a linear method, it may not capture the full extent of data complexity, but it can be used before the training of an ANN to reduce its required complexity.
Reinforcement Learning is a ML method based on rewarding desired behaviours and/or punishing undesired ones. The main element in a reinforcement learning process is the so-called agent, which can perceive and interpret its environment, taking actions and learning through trial and error.
Imitation Learning is a framework for learning a behaviour policy from demonstrations.
Few-shot learning is a learning method whose predictions are based on a limited number of samples.
Transfer learning is ML where a model developed for a task is reused as the starting point for a model on a different but related task. It is based on the re-

use of the model weights from pre-trained models. It can be thought as a type of weight initialization scheme.

4.2 Traditional Artificial Neural Networks

Artificial neural networks, introduced in the 1940s, have as main component the artificial neuron or node which tries to model the biological neurons in the brain. In the following, artificial neural networks will also be called neural networks, when there is no risk of confusion, and shortened to ANN or NN.

ANNs are composed of connected artificial neurons capable of transmitting signals to each other: the artificial neuron that receives the signal elaborates it, and then transmits it to the near neurons connected to it. Each connection has a weight, which has the scope to increase or decrease the strength of the signal. In this way, the signals coming into an artificial neuron from other artificial neurons are multiplied by a weight, then they are summed by the neuron together with an added bias, and finally they are passed through an "activation function" before they can be retransmitted (Figure 3). The hyperbolic tangent and the step functions are examples of those used for this purpose.

The non-linear properties introduced by the activation functions into the NN help it learn complex relationships between input and output.

The neurons of a NN are typically organised in layers, where the neurons of a layer are connected to the ones of the preceding layer and to the ones of the subsequent layer.

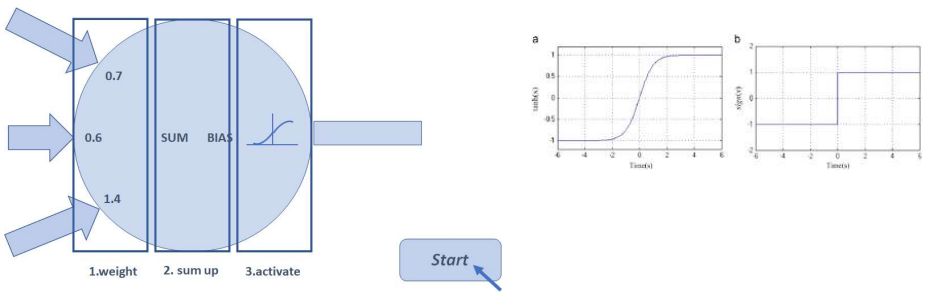

Figure 3 - A neuron of an ANN and two activation functions

The layer with no preceding layer receives input data and is called "input layer", while the layer with no subsequent layer produces output data (also referred as *labels*) and is called "output layer". The layer(s) in between are called "hidden layers", since they are generally not exposed outside the NN (Figure 4).

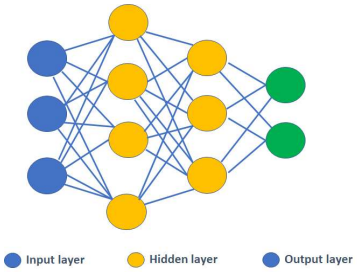

Figure 4 - Layers in a NN

The neurons of a layer can connect to the neurons of the subsequent layer in multiple ways. When all neurons of a layer connect to all neurons of the subsequent layer, the layer is called *Fully Connected* (Figure 5).

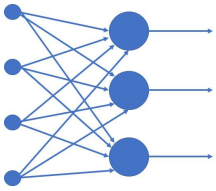

Figure 5 - Fully Connected Layer

As computation may become an issue when the network grows, a group of neurons in one layer may connect only to a single neuron in the subsequent layer.

The general structure of a NN (i.e., the number of hidden layers, the number of neurons for each layer, the activation function and other parameters) can be decided by the designer but, before it can perform the task for which it has been designed, a NN must be trained.

The *training* is the process where the learning model (the NN in our case) adjusts its parameters to minimize the observed errors when fed with a *training set*, a dataset completely known to the designer consisting of tuples which associate values taken from a domain space (those coming at the input of the model) to values taken from a label space (the results that should be outputted by the model). For example, in an image-to-text converter application, the domain space could be the set of all possible images containing some text, while the label space could be the set of all strings.

In case of a NN, the training can be done through the so-called *learning process*, which is an adaptation process where the NN adjusts its weights to minimize the observed error on the training set (called *training error*) calculated through a cost function. The learning process takes advantage of the *back-propagation* algorithm, which calculates the gradient of the cost function as-

sociated to the weights of the network, starting from the output layer and moving backward to the input layer.
Hyperparameters are used to control the learning process. Of particular importance is the "learning rate", which defines the size of the corrective step that the model takes to adjust for errors in each observation: a high learning rate makes the training time short, but the ultimate accuracy is low, while a low learning rate takes longer, but the ultimate accuracy is greater. As NNs are highly non-linear systems, adaptively changing learning rates avoids oscillations inside the network (e.g., when connection weights go up and down), and improves the convergence rate.
Vanishing and Exploding Gradient is a common problem in NN training associated with the *backpropagation* algorithm (Figure 6). When the NN is "deep", i.e., with many hidden layers, the gradient may vanish or explode as it propagates backward.

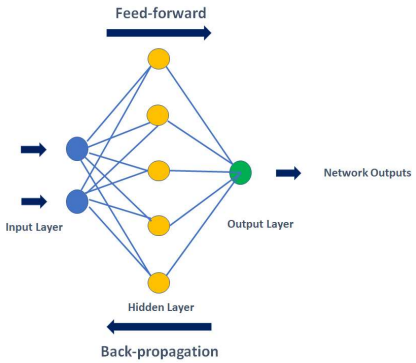

Figure 6 - Backpropagation in a NN

Together with the training set, it is possible to use a *test set* to validate the performance of a model when the learning process is concluded. A test set is a dataset of tuples taken from the same domain and label spaces of the training set but containing different values. Its importance resides in the possibility to compute the *generalisation error*, which is the probability that our model does not predict the correct label on a random data point. It is possible to have an estimate of this error by giving to the already trained model the input values of the test set, and then comparing the output of the model with the labels of the test set.
There are two cases of particular importance which can give hints to optimise the learning process. In fact, it may happen that the model presents a high training error (or high training times) and a high generalisation error. This eventuality is called *underfitting*, and it shows that the model is too simple to describe the complexity of the data. On the other hand, it may happen that the model presents a very small training error, but at the same time a big generalisation error. In this case, we are talking of *overfitting*, and it shows that

the model has adapted too much to the specific input data, with the consequence of no longer being a generalisation of the data model.

4.3 A question

After this quick tour of algorithmic data structures comparable to biological neurons, intriguing questions arise. Can the natural operation of a biological neuron be simulated with an artificial neural network or are NN-like data structures just a mathematical abstraction? How does a biological neuron process information from other neurons' inputs? The answer is that the mechanisms operating in a biological neuron are well-studied, and indeed realistic models are available to simulate the chemical reactions taking place in, and the neurobiology of, several different classes of neurons. However, the way neurons respond to signals is inherently complex, and each neuron is equivalent to a deep network of computational neurons [4].

5 Speaking humans and machines

Communication, whether among humans or between humans and machines, can proceed along multiple channels – hence the importance of multimodal interfaces. Still, language will remain the principal communication channel. While linguistic communication can be text-based or voice-based, our concern here is vocal communication. The state of the art in artificial voice-based communication as it relates to MPAI's wider goals will be sketched: to foster the creation and increasing use of standard AI-based modules (AIMs) that facilitate implementation of varied multi-module use cases, called AI Workflows (AIWs).

With this approach in mind, the state of the art in automatic speech recognition (ASR) and text-to-speech (TTS) is considered with an eye toward current and future workflows and their benefits and dangers.

5.1 Automatic Speech Recognition

Classical ASR

ASR has made particularly dramatic progress in the last two decades. Throughout the 2000s, speaker-dependent ASR remained dominant: to achieve acceptable accuracy using commercially available ASR, each speaker had to provide speech samples, initially twenty minutes or more. In most systems, the speech signal to be converted into text was sliced into short segments, so that the system could estimate the probability of certain text sequences given a sequence of sound slices, generally using Hidden Markov Models (HMMs). These estimates yielded possible words or word fragments and their probability rankings; and one could go on to estimate which word sequences were most likely, using compilation of word sequence

probabilities called language model. The search through the associated set of possibilities – the associated space of possible words and word sequences – was usually managed through some variant of Viterbi search techniques.

By means of these techniques, and with sufficient speaker-specific and domain-specific recordings and accurate transcripts as training material, accuracies well above 90% became feasible. Necessary recording time dropped in a few years from twenty-plus minutes to less than a minute as processing power steadily increased according to Moore's Law and as usable recording databases became much larger. As a result, speaker-independent training had finally arrived by the early 2010s: that is, training time per new speaker had dropped to zero!

Neural ASR

Then neural speech recognition appeared on the scene: by the late 2010s, Deep Neural Networks (DNNs) had essentially replaced HMM-based systems. Fundamentally, NN models learn input-to-output relationships: when given certain patterns as input, they learn to yield certain patterns as output. For ASR, they can learn to deliver the most probable text transcripts when given suitably pre-processed speech signals. However, since speech recognition involves mediating between sequential patterns for both input (sequences of sounds) and output (sequences of graphemes – that is, letters or characters – and words), neural architectures specialized for sequences are essential. Until recently, Recurrent NNs and Convolutional NNs were preferred – the first, designed, when computing sound-to-text probabilities for the next step along a sequence in progress, to accumulate the output of all prior steps and include these as input; and the second, designed to exploit a window moving across the sequence. These have now made room for transformer-based setups, which can efficiently shift attention throughout an entire sequence, thus providing superior consideration of audio and textual contexts.

5.2 ASR issues and directions

ASR Issues

Numerous problems remain. Much speech, whether collected in real time or from recordings, is spontaneous rather than from written materials, and consequently contains hesitations, stutters, repetitions, fragments, and other features unfriendly to recognition. Speech often occurs in noisy environments. It often involves multiparty conversations, with several voices that often overlap. The voices may be speaking different dialects and may even mix languages.

To address these and other issues, continued ASR development beyond neural network techniques themselves is under way. Numerous possible ar-

chitectural variations and component interactions can be tried according to the use case. Noise reduction modules can deliver cleaner audio input. Language, dialect, and/or domain recognition modules can pre-select optimally trained variant ASR modules.

Integration of knowledge sources will also be a fruitful ongoing research direction. Presently, ASR still usually operates with little knowledge of the language structure other than sequence relations. Also usually lacking is any attempt to understand the objects and relationships in the speech situation.

ASR Directions

Considerations of understanding raise the question of future use cases for ASR. As one example for now – we'll see several more below – consider self-driving cars: the car will "know" about its dynamic environment, having acquired from "experience" (multiple instances) visual "concepts" like CAR, TRUCK, STREET, and their spatial and causative relations. And so, when recognizing user questions or commands concerning cars, trucks, streets, etc., the car will be able to use knowledge about the referents – and not only the audio and the prior text – to raise or lower probabilities of currently recognized text. But a car's concepts could include not only visual percepts but also a wide range of sensor data, such as sounds, vibrations, lidar or radar. In coming years, this incorporation of perceptually grounded knowledge is likely to transform all areas of AI, speech recognition not least. The results will affect speech translation; transcription of all audio and video (real-time and otherwise); and in fact, every use case demanding ASR – roughly, every use case involving speech.

Speech Analysis

While considering speech recognition, we should not overlook speech analysis to extract extra-textual information, such as sentiment or other social factors: what are the speaker's emotions, styles, backgrounds, or attitudes? That vocal analysis can complement textual analysis of the language. If carried out by ML, it must depend heavily on the amount and quality of available data – for instance, on collections of recordings reliably labelled, or otherwise identified, for the emotion or other relevant factors.

5.3 Text-to-Speech (TTS)

Synthetic speech reached an acceptable quality level – understandable if colourless and unmistakably artificial – in the 1990s. The problem was considered largely solved; and, partly for that reason, remained relatively static while ASR was rapidly and noticeably improving. We'll look at "classical" text-to-speech first, then move on to the current neural era.

Classical TTS

Concatenative TTS

The most widely used classic technology – still in use for some purposes – was concatenative: that is, short, recorded audio segments associated with speech sounds (phonemes and their sub-parts or groupings) were stitched together (concatenated) to compose words and larger units.

The segments in question were collected from large databases of recorded speech. Utterances were segmented into individual phones, syllables, words, etc., usually using a specially modified speech recognition system yielding an alignment between sound elements and those linguistic units. An index of the units was compiled, based on the segmentation and on acoustic parameters including pitch, duration, and position among other units. And then, to build a target utterance given a text, one selected the best chain of candidate units, typically using a decision tree while extending the chain. Good results could be achieved, but maximum naturalness required large recording databases, up to dozens of hours. (Alternatives to such concatenative text-to-speech could synthesize utterances from scratch, by artificially generating waveforms. The resulting speech was less natural, but waveform methods had advantages e.g., in size, so that they lent themselves to implementations in small devices, even toys.)

General TTS Issues

Concatenative or otherwise, any speech synthesis system confronts several issues.

Allophones and Co-articulation. Phonemes are generally pronounced differently (as allophones, or phoneme variants) according to their place in words or phrases. For instance, in US English, phoneme /t/ may be pronounced with or without a puff of air (called aspiration, present in top but absent in pot). Moreover, even those variants – and all other speech sounds – will vary further in context according to the neighbouring sounds (i.e., to co-articulation effects): for instance, the puffed /t/ sounds different before different vowels. (For this reason, diphones, or pairs of phonemes, are frequently used as speech sound groupings.) Co-articulation changes arising from some sound sequences can be dramatic in given styles or registers, as when /t/+/y/ in don't you becomes the /ch/ of doncha. If classical TTS handled such cases – they usually didn't – it was through dedicated spellings ("doncha") or through programs implementing hand-written combinatory rules.

Disambiguation. Then there's the problem posed by text sequences that can be pronounced entirely differently according to their use in a sentence, like "record" in "For the record, …" vs. "We need to record this meeting." Some analysis of sentences is needed to select the appropriate variant and re-

solve the ambiguity – that is, to perform disambiguation. In classical text-to-speech, this need was often met by symbolic (hand-written) parsing programs.

Normalization. Yet another challenge is presented by text elements whose pronunciation isn't specified in text at all but is instead left to the knowledge of the reader-out-loud. Numbers and dates are typical examples: 7/2/21 might be pronounced as "July second, twenty twenty-one" in the US – though variants are many, even leaving aside the matter of European writing conventions. Some ways must be found to convert symbols etc. to pronounceable text – to normalize the text.

Pronunciation problems. Foreign or unfamiliar words ("Just hang a uey on El Camino.") pose obvious difficulties for text-to-speech. They're normally addressed either through compilation of specialized or custom dictionaries or through use of a guesser – a program that uses rules (then) or AI (now) to guess the most likely pronunciation.

Prosody. Some treatment is needed of prosody – movement of pitch (melody), duration (rhythm), and volume (loudness). In the classical era, the prosody of a sentence was superimposed on speech units via various digital signal processing techniques. For instance, via the Pitch Synchronous Overlap and Add (PSOLA) technique, the speech waveform is divided into small overlapping segments that can be moved further apart to decrease the pitch, or closer together to increase it. Segments could be repeated multiple times to increase the duration of a section or eliminated to decrease it. The final segments were combined by overlapping them and smoothing the overlap. The means of predicting the appropriate prosody were relatively simple – e.g., by reference to punctuation – so the results were often repetitive and lacking in expression.

Extra-prosodic speech features. Extra-prosodic speech features like breathiness, vocal tension, creakiness, etc. were only occasionally treated in research, e.g., by simulating the physics of the voice tract. Using models of vocal frequency jitter and tremor, airflow noise and laryngeal asymmetries, one system was used to mimic the timbre of vocally challenged speakers, giving controlled levels of roughness, breathiness, and strain.

Neural TTS

As mentioned, neural technology learns input-to-output functions – usually from corpora of input-output examples. For neural speech synthesis, the job is now usually divided into two input-to-output problems: (1) given text, what should be the corresponding acoustic features (numbers indicating factors like segment pitch, duration, etc.) – call this acoustic feature generation; and (2) given acoustic features, what actual waveforms should be generated – call this waveform generation, the function of a vocoder.

For (1), the acoustic features are represented as spectrograms, which show frequency changes over time: in an X/Y graph, the vertical (Y) axis shows frequency, and the horizontal (X) axis shows time. (A modified frequency scale is often substituted for raw frequency: the mel frequency scale – mel for "melody" – which takes account of human perception.)

Neural text-to-speech began as recently as 2016, when DeepMind demonstrated networks able to model raw waveforms and thus to generate speech from acoustic features. In 2017, the technology was used by others to produce such raw waveforms directly from text – and neural text-to-speech was born. At the same time, Google and Facebook offered Tacotron and VoiceLoop, which could generate acoustic features, as opposed to waveforms, from input text. Then Google proposed Tacotron2, combining a revised acoustic feature generator with the WaveNet vocoder. The entire sequence – text to waveform – is termed *end-to-end speech synthesis*. Now that current end-to-end systems can generate speech whose quality approaches that of humans, this methodology has been widely adopted.

End-to-end speech synthesis models are indeed attractive. Good models for given speakers or languages, or for new data, can be created with little engineering. They're robust, since there are no components that can fail. Unlike classical concatenative models, they require no large databases at run time.

Neural TTS Issues

But of course, challenges remain.

1. *Learning of models takes much time and computation*. Resolution efforts have emphasized architectural variation for handling NN-based prediction of acoustic sequences. Transformer-based architecture (which, as mentioned, can scan back and forth throughout an entire sequence) is substituted for auto-regressive models, which make predictions about future sequences by reference to a limited number of past elements, or for Recurrent NNs, which refer to an accumulation of all past elements. Transformer-based sequence prediction is enhanced by also modelling the duration of speech sounds.
2. *If training data is insufficient or low in quality, speech quality suffers*. The problem turns out to be strongly related to text alignment failures; so, focus has been on improving alignment by leveraging the known relations between text and speech sounds: their respective sequences march forward in tandem, and nearby text and sounds are more helpful for prediction than distant ones.
3. *Control points are absent*: what you hear is what you get. Research has stressed variational auto-encoders – methods of learning representations of certain speech features as embeddings, or points in multi-dimensional

(vector) space. For example, the points can represent emotions (like anger or sadness) as expressed through speech features like pitch or rhythm. That representation remains separate from, e.g., the pronunciation, and thus can be combined with it. Moreover, the emotions themselves can be blended or combined. Another control tactic is to break up the speech synthesis problem into several stages or aspects, so that each aspect can be separately programmed or trained, and thus controlled. For instance, a separate pre-processing stage can handle co-articulation combinations like don't + you to yield the pronunciation of doncha. Any such combinatory or divide-and-control methods can be become elements of automated or semi-automated workflows.

4. *Prosody and pronunciation tend to be flat*, since they're averaged over large collections of training data. Intervention is possible at or after synthesis time: users can interactively post-tune preliminary flat (emotionless, bland, boring) renderings, either through demonstrations (via microphones or recordings) or via manual user interfaces. In addition, a single text-to-speech model can be made to generate speech with various speaker styles and characteristics. The trick is to create embeddings representing speakers and/or speaking styles, as opposed to emotions in our previous example.
5. *And more*. The challenges surveyed above in relation to classical speech synthesis are still with us in the neural era: normalization ("Call 521-4553 after 6pm for a good time."); disambiguation ("Chuck Berry wanted to record a new record."); pronunciation of foreign or unfamiliar words ("Just hang a uey on El Camino."); and so on. However, each such problem also provides an opportunity to propose a dedicated AI module (for MPAI, an AIM) as a solution.

Neural Vocoders

We mentioned that neural speech synthesis can be handled or conceptualized in two stages, where the second is sound generation (acoustic-features-to-waveforms), as performed by a vocoder. That vocoder can exploit neural networks, as do the popular Wavenet and WiFi-GAN vocoders.

5.4 Some final considerations

TTS Evaluation

How can we judge the quality or adequacy of a speech synthesis system?

1. Human judgements are unavoidable at the state of the art; but, once elicited, these judgments could also become input for ML, leading in time to automatic judgments approaching human ones.
2. Establishment of common test sets will become increasingly important.

3. Also significant will be development of automatic assessment of styles, emotions, attitudes, etc. In discussing ASR, we mentioned use of speech analysis to extract such extra-textual information. When these techniques become reliable, they can be applied to speech synthesis evaluation.

Speech Technology: Dangers

Since language is so central to human experience, linguistic technology can only be hugely influential, the more so as it grows more powerful. Like medical technology; like energy technology; like computational technology; like communications technology – linguistic technology promises to be hugely beneficial – but, inevitably, also dangerous. For example, speech recognition magnifies the danger of ubiquitous surveillance. And speech synthesis, as an element of technology's growing capacity to simulate every aspect of perception, threatens a world of deep fakes, in which we can never be sure who said what – bad enough for celebrities and personal enemies, but worse for the powerful and entrusted. We can hope, however, that laws and norms will ultimately combine with technological fixes to ward off the most dystopian dangers.

Speech Technology: Benefits

But as for the potential benefits: MPAI's aim is to promote the creation of standard modules that can be assembled in endless configurations, so that myriad beneficial systems can be created without endless reinvention of the wheel.

6 Visual humans and machines

6.1 Introduction

Actions and facial expressions play an important role in communication between humans and machines. Even if the exact words are the same, the information contained in words will change if the actions and facial expressions are different. Therefore, for perfect human-machine communication, changes in visual behavior and facial expressions are required. Also, people can have a comfortable conversation when they feel like they are human and not machines, which means we need to create the same facial expressions as humans.

There is a need for a standard for AI that can facilitate the implementation of AI technology using multiple modules. One of these goals of MPAI is to address the latest AI technology for the visual part.

6.2 Facial attributes estimation

For a human generator to create a face and body according to the change of human attributes, it is necessary to understand human attributes. It can do thst in two ways: explicit and implicit.

Explicit Way

- **Key Point Detection** is a method of extracting a key point of a person from a video in which a person appears, such as a landmark detection or 3D angle detection. In general, the key point of a person is extracted using the AI classifier, and for this, key point information of the person in the image and the annotated image is required. Based on this person's supervision, the AI module learns the person's attribute information.
- **Facial Emotion Detection.** In general, humans can read emotions based on the facial expressions of the other party. Therefore, people can label a person's emotion in the video by looking at the expression and behavior of the person in the video. Learning the AI classifier based on the labeled videos makes learning the AI classifier that detects the person's emotion. Again, the AI module explicitly learns human emotions based on human supervision.

Implicit Way

- **Key Point Detection**. In addition to learning key points based on human supervision, the AI module can learn key points indirectly. By learning several people's movements, the AI module can extract key points of people that can be generally applied. The key point of the person that the AI module thinks is different from the landmark that humans annotate, but since the AI module extracts a key point that is easier to understand, it shows a higher performance when creating a person.
- **Latent Space in Human Generator**. To create a person, we need a generator for this. The most used generator for synthesizing virtual humans is a Generative Adversarial Network (GAN). GAN enables the generator and discriminator to be trained competitively and generate images. The image generated by the generator is determined whether it is an actual image or a generated image through the discriminator. And then, the generator is trained to deceive the discriminator. In this process, the generator finds a low-dimensional latent space that can express the image to be created. Since the low-dimensional latent space found in this way implicitly contains information about the image, it contains attribute information about the person that the AI model thinks, like the attribute about the person that a person thinks. Generators will be discussed in more detail in the next section.

- Voice Information Voice information is needed to implement the human visual part perfectly. When a person speaks, it will have a mouth shape that is in sync with it. In addition, depending on the person's words, their actions and expressions change. Therefore, it is necessary to extract information about mouth shapes and emotions contained in human speech and also be able to use this information. This suggests the need for a multi-modal AI module because it generates a visible part based on voice information.

6.3 Synthesising humans visually

As explained in the chapter above, a constructor is needed to synthesize a virtual human visually. The generative adversarial network (GAN) is the most used, and the most popular technology among GANs is Style GAN.
Style GAN has the advantage of creating a high-quality image of 1024 x 1024 pixels. It is trained to adjust the attributes of the image created by disentangling the low-dimensional latent space. Suppose we use a non-disentangled low-dimensional latent space when we want to change the characteristics of the generated image (e.g., face angle). In that case, the characteristics we want to change and other characteristics (ex. facial expressions) also change simultaneously. This causes difficulties in synthesizing Humans. On the other hand, if we use the low-dimensional latent space of the disentangle style GAN, it is possible to synthesize humans by changing only the characteristics we want to change.
In addition to GAN, a technology called NeRF has recently been attracting attention. GAN is a technology that generates images in two-dimensional space, whereas NeRF is a technology that creates images in three-dimensional space. Since the human face and body are three-dimensional data, there is a limit to generating an image in a two-dimensional space. On the other hand, because NeRF represents a person in a three-dimensional space, more accurate human creation is possible. The NeRF projects the camera ray in the direction of the image pixel to be generated from the observer's viewpoint. The pixel's color (r, g, b) information is calculated by adding up the color (r, g, b) information and volume density information of each point where this camera ray passes through the object.
The way GAN and NeRF utilize the Human attribute is different. Because GAN controls the image generated by moving in the low-dimensional latent space derived from the learning process, the human analyzes the human attribute understood by deep learning, creating the desired human image. On the other hand, in the case of NeRF, an image is generated based on three-dimensional coordinates and angle information, so a desired human image is created using explicit human attributes.

6.4 Applications & potential dangers

Applications

Virtual humans can be used in many ways. First, the part that is receiving the most attention is marketing. Since a virtual human has fewer temporal and spatial constraints than a real person, it is unnecessary to consider the temporal and spatial aspects when shooting an advertisement video or appearing content. Next is a virtual chatbot. Human-computer interfaces that use only voice and text, such as kiosks or navigation, can utilize visual information due to the advent of virtual humans. This increases friendliness and can further lead to improving the company's image. In addition, it is directly related to the creation of characters in the metaverse world, which has recently been in the spotlight. Even within the metaverse, a virtual world, virtual humans are required for activities, and AI technology will be utilized in many ways to create such virtual humans.

Potential Dangers

People lose their jobs as virtual humans replace their jobs. Virtual humans can replace influencers who earn income from marketing filming, and virtual humans can replace various occupations such as announcers and cafe staff. In addition, if a virtual human imitating a real famous person appears, the virtual person may appear in negative images such as pornography and violent images, thereby damaging the famous person's image. Furthermore, it can be abused to manipulate public opinion.

7 Humans conversing with machines

In recent years, the development of AI-based systems for providing human-friendly services based on human understanding has been steadily progressing. The main technology of such AI systems can be called *language intelligence* technology. This allows users to communicate desired knowledge expressed in words while accessing services, or to acquire and communicate the same knowledge without being confined to a specific language. The realization of language intelligence technology for communication and knowledge acquisition is expected to make human life more convenient.

Lexical and sentence grammar analysis used in ML/deep learning-based Natural Language Processing (NLP) and Question Answering (QA) are making AI-based natural language understanding successful in commercial services. Today, various language intelligence-based systems and services are employed in intelligent information services applications.

Speech interface technology fundamentally expands the way humans interact with machines. For example, the field of machine-based professional counselling services aims to reach the level of counselling that an expert hu-

man can provide using speech recognition, conversation processing and knowledge mining. Through this, full-fledged interactive AI-based services (e.g., unmanned AI-based call centre, AI shopping host, etc.) are possible, and they can also be used for public services such as 24-hour counselling for the socially marginalized elderly and youth, and for soldiers needing help.

The speech interface can be expanded to other knowledge processing areas by recognising variously acquired emotions and situations, merging them with speech and linguistic intelligence, and mapping visual and other types of information to a unified knowledge space.

Speech and language are the most natural communication means for humans and language intelligence is the AI technology that enables machines to recognize, understand, and generate meaning out of speech. AI-based language intelligence is making it possible for machines to understand content, learn, think, and reason much like a human.

7.1 Question Answering (QA)

QA is an intelligent function that generates answers to a user's question in a natural language. In the future, more and more systems will be equipped with QA functions for a better user experience. Currently, the most widely used QA system is an intelligent agent provided in smart phones. This is not a completely intelligent agent able to answer all questions that users may ask, and more domain-specific QA services would be useful.

Figure 6 presents an example of the functional blocks of a QA system architecture. A traditional QA system consists of several functional blocks: natural language processing (NLP), question analysis, candidate answer generation, answer inference and answer generation.

Natural language QA is a technology that proposes correct answer candidates to a user's natural language question and selects and presents an optimal answer among them. Natural language QA technology consists of the following core technologies.

Structure/semantic analysis understands the contents of a natural language text by means of syntax analysis, and semantic analysis such as named entity recognition.

Question analysis and understanding classifies and recognizes questions.

Open knowledge extraction generates pre-defined data structures filled with the results of syntax analysis or semantic role assignment for Big Data processing. It includes extracting knowledge by identifying subjects and objects of sentences using only text.

Knowledge to KB linking technology compares the knowledge extracted from the text with the knowledge existing in the KB, analyses the truth, redundancy, and plausibility of knowledge, and finally decides whether to delete or keep it.

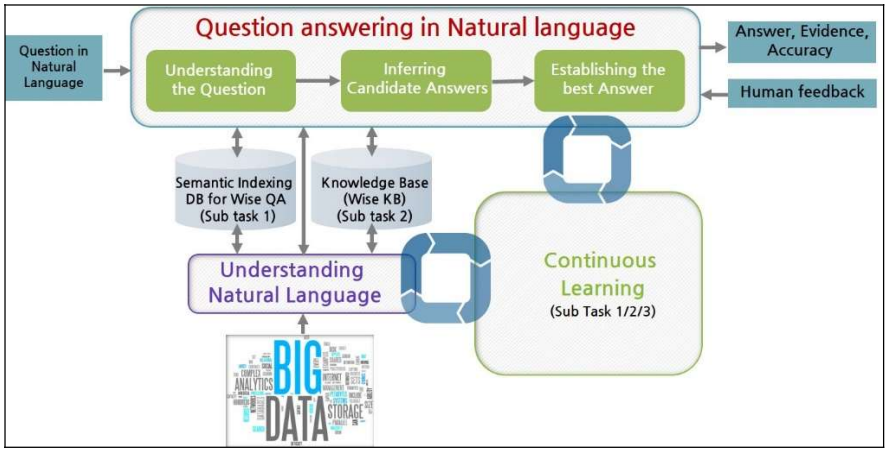

Figure 6 – Example of functional blocks in QA architecture

Candidate answer generation searches for effective documents/paragraphs, extracts correct answer candidates from related paragraphs, and calculates reliability for each candidate answer.

Best answer selection performs answer inference based on feature normalization and ranking of candidate answers and best answer generation on the terminal.

Neural network approach for QA

The neural network approach to QA systems is gradually introduced replacing some modules of the traditional data processing-based QA systems.

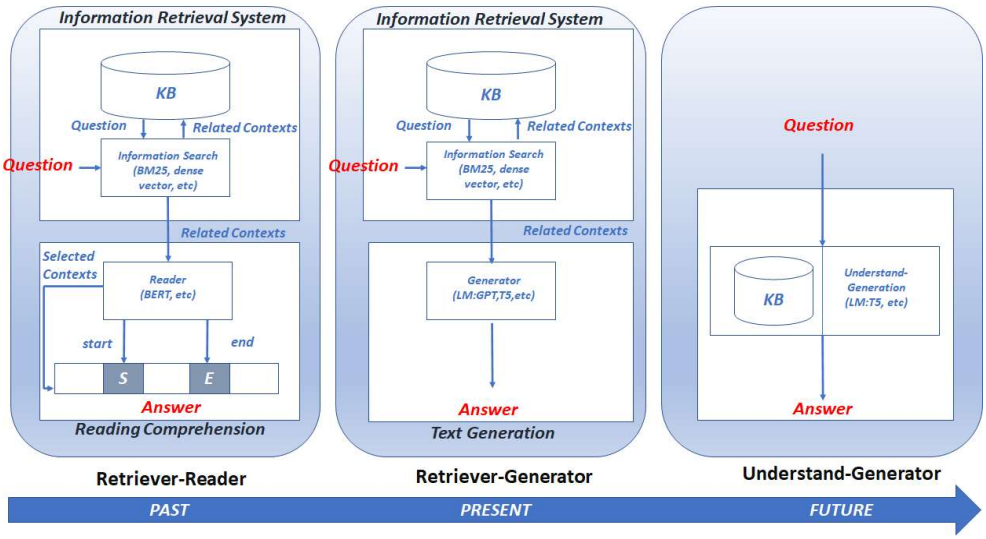

Figure 7 – Three types of open-domain question answering[1]

[1] https://lilianweng.github.io/lil-log/2020/10/29/open-domain-question-answering.html

The classification introduced by [5] shows different types of open-domain questions in increasing order of difficulty that a NN can answer:

1. The question and answer have been seen at training time.
2. The question was not seen at training time, but the answer was.
3. Neither the question nor the answer was seen at training time.

7.2 Dialog Processing

Natural language dialogue processing includes *conversation understanding* technology to determine the intention of the questioner, *dialogue modelling* technology for generating natural dialogues, and *dialogue error correction.* The last is relevant to foreign language learning applications.

Traditional dialogue systems are based on a modular data processing architecture and their workflow is as follows: the *natural language processing* module identifies user intention and extracts task-specific knowledge, and the *dialogue state tracking* module tracks the user goal considering dialogue history. The system can obtain the belief state for comparison with the database, such as finding the number of matching entities. Based on the information, the *dialogue policy* module determines the next system action and then the *language generation* module generates an appropriate response matching the system action.

With the progress of NNs, recent work on dialog systems handles individual modules in a unified way. The approach *end-to-end dialog processing* is to train a model with users' utterance sentences, so that the system generates a response suitable for given purposes, contexts, and situations. This has shown very promising results. Such models are typically developed by fine-tuning the large pre-trained models.

Another approach that leverages knowledge transfer is *multi-task learning*. The goal is to learn common knowledge representations between related tasks. It has been shown that multi-task learning not only improves NN performance, but also mitigates the problem of overfitting, i.e., the case when the training was made to match too closely a particular set of data and may therefore be inadequate to make reliable predictions. Both approaches are complementary and combing them improves the performance of natural language understanding.

End-to-End Dialog Processing

For the dialog processing subsystem to generate a system response suitable for its purpose, various learning methods are required such as transfer learning, semi-supervised dialog knowledge extraction, longitudinal supervised learning, self-conversational reinforcement learning, text processing, query generation, and text understanding. As a result of this learning, NN DBs for end-to-end dialogue model knowledge and domain dialogue know-

ledge are built. A system response is made by the end-to-end conversational processing produced using these NN DBs.

The end-to-end dialog processing decoder produces an appropriate system response according to the current situation. It uses structured or unstructured domain knowledge and the dialog history of the prior system and user utterances up to the current turn. In addition to generating relevant system responses, the dialog processing decoder can process out-of-domain utterances. That is, it can generate a natural system utterance in response to the user's non-domain utterance, and then return to the original goal-oriented dialog.

7.3 Deep Learning Language Model

Language models are those that assign probabilities to sequences of words for word prediction. Statistical Language Models use traditional statistical techniques like N-grams, Hidden Markov Models (HMM) and certain linguistic rules to learn the probability distribution of words. Neural Language Models use different kinds of Neural Networks to model language and have better performance and effectiveness than statistical language models. The Deep Learning language model is a pre-trained NN built from a large text data set for general-purpose semantic expressions that can be used for various tasks. To learn universal semantic expressions from text Big Data, self-supervised learning tasks such as predicting a blank or the next word are used. In these cases, a human can find a correct answer without a separate correct answer being previously labelled. The deep learning language NN has recently been used as a base NN for various language processing tasks such as language understanding and language generation.

Transformer Model

The deep learning language model mainly uses the transformer model that has advanced the state of the art in many Natural Language Processing (NLP) tasks. The structure of the transformer model is shown in Figure 8. It consists of an encoder part (left), and a decoder part (right) with stacked self-attention and pointwise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 8, respectively.

The transformer model, originally proposed in the field of machine translation, consists of an encoder part (left) that expresses the input text in the deep learning space, and a decoder part (right) that predicts the next word using the input text and previous output results. The encoder and decoder consist of 2 main modules: first module is to select a nearby word important for blank word prediction or to select previous words important for next word prediction using a self-attention mechanism. The second module is to calculate a deep expression using a NN using a two-layer feed-forward neural network (FFNN). Compared with CNNs and RNNs, the *transformer model* can

reflect information by directly calculating the relationship between a plurality of words that have an important relationship with each other and can be easily parallelized. Therefore, the technology is gradually expanding not only in the field of language processing but also in the field of vision.

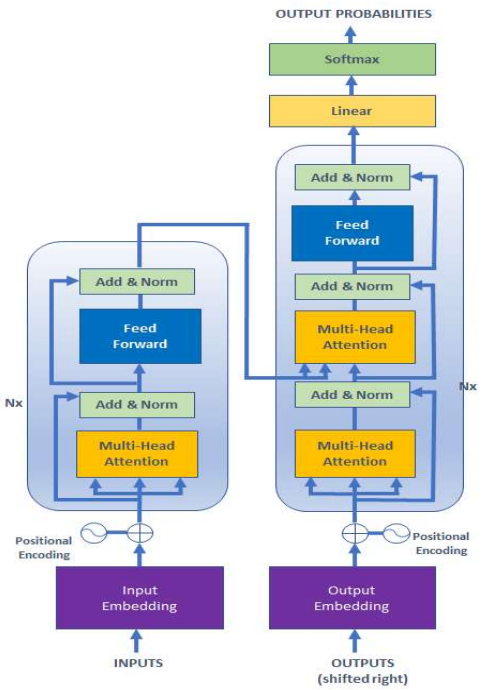

Figure 8 – Structure of Transformer Model [6]

There are 3 types of deep learning language models: encoder-based models, decoder-based models, and encoder-decoder-based models. The language model that pre-trained the encoder of the transformer using bulk text is a BERT-series language model, and the model that pre-trained the decoder is a Generative Pre-trained Transformer (GPT)-series language model. It includes the raised GPT-3 model. T5 and BART are examples of pre-trained models using the transformer's encoder and decoder together.

Encoder-based Language Model

The BERT series language model shows excellent performance in various language understanding tasks such as machine reading, document classification, language analysis, and search result ranking. This model has the most follow-up and application studies among pre-learning language models. However, it has several disadvantages: it is not easy to use for language generation type tasks such as summary and dialogue processing, and it is necessary to add a task-specific layer when applied to a specific application task.

The addition of task-specific layers makes it difficult to learn multi-tasks. Even if a very large BERT language model is built, if a new task-specific layer cannot be learned with a small number of learning examples, few-shot learning like GPT-3 will be difficult.

Decoder-based Language Model

The GPT model uses the decoder structure of the transformer and predicts the next word by using the previous word information for each word. GPT-3 increases the size of the GPT model to 175 billion. It shows that a sufficiently large and large-capacity language model is capable of few-shot learning. To prove that few-shot learning is possible, GPT-3 tested various fields such as news article generation, QA, machine reading, translation, and virtual tasks, and measured the few-shot learning performance by model size in each experiment.

GPT-3 has shown that the 175 billion-scale model can be quickly generalized like the few-shot learning. Especially in the case of news article generation, it is possible to create an article that humans can hardly distinguish from an article written by a human.

However, it is not clear whether the language model has solved the few-shot learning through actual reasoning, or the task was solved by the pattern recognition result of the pre-learning stage. GPT-3 still repeats or contradicts similar words when generating articles. There are limitations in the fact that it generates texts that show low performance in semantic comparison tasks, does not reflect video and real-world physical interactions, and requires high costs in the learning and application stages. In addition, the GPT series model performs only one-way operation on input sentences, and the fact that it shows low performance compared to the model size in the machine reading comprehension task is a limitation to its utilization.

Encoder-Decoder-based Language Model

There is a pre-learning language model that uses both a transformer encoder and a decoder, the *T5 model*, which shows a universal input/output framework that receives text as input and outputs text as a result. Structurally, the T5 model includes both encoder and decoder, and has the advantage of showing excellent performance in language comprehension tasks such as machine reading and sentence classification, and language generation tasks such as summary and translation. And, unlike BERT, using a text-to-text general-purpose input/output framework, separate task-specific layer learning for each application task is unnecessary and natural multi-task learning is possible. The T5 model requires 10% additional computation compared to the BERT and GPT single models because it requires intensive computation with the encoder result in the decoder structure. When a specific task such as

machine reading is the goal, it has a disadvantage that the performance is not excellent compared to the BERT series model with the same amount of computation.

Issues on Language Model

Although the deep learning pre-trained language model is being used as a base model in various language processing tasks such as language understanding and language generation, it still has many limitations. For example, the current pre-learning technology requires tens to hundreds of times more learning data than text that humans see throughout their life. Another issue is that all knowledge is implicitly stored in real values in the deep learning model, and there is no learning about real-world physical interactions. It is expected that these limitations can be overcome through future studies such as improvement of sample efficiency in the learning process, model structure that can utilize external knowledge, and learning that reflects other modalities such as video.

8 Audio for humans

We are surrounded by sound. The perception of different sounds is an important part of the human experience and can trigger a wide range of reactions. Think of the human response to a dangerous sound such as a lion's roar in the wild, or to the noise of a fast truck coming our way in a trafficked road, or to the soothing tone of a gentle lullaby. Audio gives us an electrical representation of sounds, and sometimes we are tempted to think that modern sensors may handle sound better than humans. So far, however, they lack the ability to manage sound the way humans do. AI algorithms can come to our rescue as they are powerful tools in understanding audio patterns.

8.1 Predictive maintenance

Besides human speech, we can characterize a wide range of audio signals, including: (1) *music*; (2) *environmental audio* such as a car passing by, a door closing, a gun shooting or a human screaming; (3) *machine audio* originated from electronic, electrical, or mechanical machines. Both environmental and machine audio are the result of physical events of interest that can be interpreted for practical use. Did a security door latch properly when it was closed? Does the fan on a cooling unit in a nuclear power plant sound like it is about to malfunction? Does a person scream in terror? The ability to detect panic in a person's voice or a cry for help could make the difference in an emergency. In these two classes of audio signals, AI could be integrated as a value-added feature in building technology solutions, particularly in the context of physical safety and security, e.g., by detecting and localising "in-

teresting" events. *Predictive maintenance in industrial settings* is an example where AI can be impactful by augmenting existing sensing capabilities. For instance, AI could analyse a motor's sound and predict a malfunction before it occurs – learning from subtle deviations in noise signatures. So, it could be an additional layer of monitoring solution for early warning systems that offer incredible value for industry by reducing downtime and saving both human lives and extensive repair costs.

There are audio AI use cases in *healthcare* as well. The human body generates sounds with clinically relevant information. AI could contribute to data-driven real-time healthcare decisions, giving alarms when a person requires immediate assistance, such as in the case of the elderly or in hospitals.

8.2 Music production and artistic industries

In general, AI is having a transformative effect on a broad array of industries, including *the music production and artistic industries* which are drawn to AI as an aid to the creative process. It is worth underscoring that AI machines don't replace humans, rather AI provides tools to render complex processes more intuitive and reduce human time spent on tedious, uncreative tasks. For example, in the music recording/production/mastering industry, there are at least three main areas where AI demonstrates its impact: assisted mastering, assisted mixing and assisted composition. Today, one of the main goals of modern mastering is to make the listening experience consistent across all formats and platforms (from low rate mp3 files rendered by inexpensive earphones to high-resolution audio reproduced with sophisticated sound systems), with a wide range of loudness constraints. All these options make mastering extremely challenging and potentially costly. AI is proving to be a viable and egalitarian choice for many musicians. By analysing data and learning from previous tracks, AI-powered tools for assisted mastering enable musicians with a small budget to easily achieve professional-level results (albeit, ultimately, without the finesse of a human expert).

8.2.1 Post-production

With so much content being created for Over-The-Top (OTT) services such as Netflix and Amazon Prime, the number of audio files to work with in post-production is dramatically increasing. Facilities are therefore looking for ways to work faster and in a more cost-efficient manner when it comes to mixing audio material. AI tools can help engineers and audio teams make basic decisions and complete the more routine tasks, thereby saving valuable pre-mixing time and enabling humans to focus on the more complex and creative elements. For example, some mastering plugins contain built-in intelligence that analyses source material (such as guitars or vocals) and considers its placement in the context of the rest of the mix to suggest mixing de-

cisions. By taking on much of the initial heavy lifting, such tools can be hugely beneficial for less experienced users.

In the commercial world, ML applications in products already exist: LANDR[2] [7], an automated audio mastering service which relies on AI to set parameters for digital audio processing and refinement; Neutron 3 released by iZotope[3], an audio mixing tool that features a "track assistant" which utilizes AI to detect instruments and suggest fitting presets to the user. In more direct processing of audio by means AI, iZotope also features a utility for isolating dialogue in their audio restoration suite RX 9[4].

8.2.2 Audio effects

Audio effects design for games and movies is another area where environmental and machine sounds are accurately recorded, catalogued, and applied. The concept of *procedural audio design* brings a partial solution to this function in that the process of sound recording is replaced by manually designed algorithms that can synthetically generate such audio. Procedural audio design is an intermediate step in automatizing audio effect design and more exciting developments can be expected via the combination of natural language processing and generative networks for AI-supported automatic sound design. A related development can be envisaged where silent movies (e.g., Man with a Movie Camera by Dziga Vertov) can be enhanced with AI-generated sound effects.

8.2.3 Assisted composition

Assisted composition is another area of music production that is quickly realizing the value of AI. More and more tools are using deep learning algorhythms to identify patterns in huge amounts of source material and then utilising the insights generated to compose basic tunes and melodies.

8.3 Immersive audio experience

If we want to *mimic and reproduce auditory scenes we hear in real life*, we utilise a set of techniques known as immersive audio. Immersive audio provides a "life-like" sound experience to end-users, different from the traditional stereo methods. This new audio experience envelops the listener, and it produces the perception on the audience of being surrounded by different audio universes by simulating credible auditory soundscapes. Disruptive innovations with AI in recording, encoding, and transcoding between immersive audio formats has gained importance for a broad industry thanks to the ever-increasing capacity of communication networks. The *holy grail* of im-

[2] https://www.landr.com/
[3] Izotope; https://www.izotope.com/en/products/neutron.html
[4] Izotope, https://www.izotope.com/en/products/rx.html

mersive audio has always been to create a sonic reality that supplants the listener's real acoustic environment by providing an emulated or synthetic auditory reality that is indistinguishable from the listener's actual reality. 3D and immersive sound, for a long time not at the forefront of multimedia applications, are now an essential part of immersive games, extended reality applications, audio-visual arts, teleconferencing applications, and advanced broadcast applications.

8.3.1 3D and immersive audio

Several key technologies for 3D and immersive audio have been proposed. These techniques can be broadly classified into synthetic and recorded 3D audio in terms of how content is created, and headphone-based and loudspeaker-based in terms of how audio is reproduced. Three approaches are especially relevant and form the basis for existing (e.g., MPEG-H 3D Audio [8] and upcoming media coding standards (e.g., MPEG-I): Higher-order Ambisonics (HOA), Object-Based Audio (OBA), and binaural synthesis. Audio signals in one of these representations can sometimes, but not always, be transcoded into the others. HOA involves the representation of the sound field in the spherical Fourier domain as a series of spherical harmonic functions. Apart from allowing straightforward operations such as the 3D rotation of a sound field, this representation offers a theoretical framework that makes it possible to synthesise physically (as opposed to perceptually) accurate sound fields generated by simple sources such as a plane waves, point sources, and/or a combination thereof.

For a long time, HOA was constrained to synthetic 3D audio, where complex sound fields could be created through what is called Ambisonics panning. Although such an approach is beneficial in synthetic and virtual environments, real sound scenes, such as those from a real concert are better suited for real recordings. Special microphone arrays that can capture HOA are now commercially available and HOA recordings are becoming more commonplace. Such microphone arrays typically comprise pressure sensors on a rigid spherical baffle and require a pre-processing stage that converts the microphone signals (also known as the *A-format* representation) into spherical harmonic de-composition (also known as the *B-format* representation). The B-format signals can then be decoded for playback from a loudspeaker rig. As such, HOA provides a large listening area, and does not require tracking the listener position to reproduce an immersive audio field. HOA, by virtue of its capability to represent a sound field that is amenable to perfect reconstruction (limited by the maximum HOA order), acts as the basic format from which other approaches can be derived. For example, perceptual sound field reconstruction (PSR) signals can be derived from HOA

signals. This important advantage resulted in HOA being selected as the *scene-based format* for MPEG-H 3D Audio.

8.3.2 Object-based audio

OBA is more of a concept than a well-defined immersive audio approach. OBA involves the storage, transmission, and processing of audio sources as distinct *audio objects*, in a way like how audio stems are used in audio production. The audio stems or objects can be positioned and repositioned to compose a 3D auditory scene; enhanced, faded, or eliminated if necessary, and embellished with reverberation. Such flexibility is essential in providing the listener with a fully personalised listening experience, one where the listener can redesign the reproduced acoustic scene within the design space that they are given.

Despite its obvious advantages, OBA is not the first choice – at least today – for representing recorded sound fields. This because OBA requires the availability of audio sources as separate audio objects in addition to the definition of the reverberation characteristics of the intended acoustic scene using a representation that is either parametric or non-parametric. The extraction of the audio objects and the reverberation characteristics from real recordings is not a trivial task and requires among other things, source localization, source separation, dereverberation and optionally the extraction of the geometry of the acoustic scene.

8.3.3 Binaural audio

Binaural audio involves the presentation of appropriate binaural cues to listeners over a pair of headphones so that they have the illusion of virtual sources in the 3D space surrounding them. The advent of mobile phones made binaural audio the de facto immersive audio approach for audio-on-the-go applications. The recent roll-out of spatial audio delivery services indicate the readiness of the market for such applications.

Binaural audio can be recorded by using anthropomorphic microphones also known as *dummy head* microphones that comprise a manikin shaped as a human head with realistic ears having microphones at the entrance of the ear. Dummy head microphones such as Neumann KU-100 physically capture binaural cues that are essential for the perception of sound sources in 3D. However, binaural recordings do not provide any means of interactivity and the listeners are presented with a high-quality immersive experience if their head is stationary. When the listener moves the head, the auditory scene also rotates drastically reducing the realism and the immersion. This renders binaural audio recordings useless in interactive 3D audio applications unless a head-tracking mechanism is applied in conjunction with personalized *head-related transfer function* (HRTF) filters.

Binaural audio can also be re-synthesised using appropriate digital HRTF filters. These filters mimic the acoustic path from a predefined sound source to the ears of the listener. Each distinct sound source is processed with a pair of HRTF filters (one each for the left and one for the right ear). Binaural audio synthesis should also respond to the movement of the listener's head, which is typically achieved using hardware-based solutions called *head trackers*.

8.3.4 Virtual reality

AI can also solve previously unsolvable problems in immersive audio and greatly improve the end-user experience in games, Virtual Reality (VR) and six degree of freedom (6DoF), navigable audio-visual content applicable in many domains including entertainment, broadcast, gaming, and cultural heritage. The generation of appropriate room reverberation to improve auditory immersion is a problem that potentially would benefit considerably from an AI-based approach and more specifically using the concept of differentiable digital signal processing (DDSP) which combines elements of deep learning with DSP.

8.3.5 Rendering immersive audio

Often, the end-user rendering capabilities dictate whether they can play back the available immersive audio content. In the early days of multichannel audio coding, this problem was addressed by designing coding algorithms (see for example [9]) that were backwards compatible, meaning, for example, that 5.1 multi-channel content could be downmixed and transmitted for reproduction over two channels only (i.e., the original 5.1 multichannel audio information would contain the transmitted stereo signal). When more complex representations such as HOA and binaural audio are considered, simple downmixing will not be sufficient. Transcoding from HOA to binaural audio is possible and is widely used since such transcoding also provides distinct computational advantages [10]. Similarly, binaural content and synthetic HOA can be obtained from OBA representations. However, three key conversions are currently missing: binaural to OBA, binaural to HOA and finally, HOA to OBA. Recent developments in data processing resulted in algorithms for high-quality audio object extraction. There also exist direction estimation, reverberation time estimation and dereverberation methods that rely on AI-based approaches.

Such AI-based approaches could make it possible to create immersive audio content that can be repurposed, recomposed, and/or remixed and pave the way for expedited and flexible AI-based 3D audio content creation

8.4 Audio preservation and preparing for the (AI) future

Preservation of audio assets recorded on a variety of media (vinyl, tapes, cassettes etc.) is an important activity for a variety of application domains, in particular cultural heritage. Audio archives are an important part of this heritage, but require relevant resources in term of people, time, and funding, since preservation requires more that "neutral" transfer of audio information from the analogue to the digital domain. In general, it is necessary to recover and preserve a lot of information in addition to the audio signals, e. g., annotations by the composer, by the technicians, etc. AI can drastically change the way we preserve, access, and add value to heritage, making its safeguarding sustainable.

The introduction of electronic and information technology into art present new challenges for archives and for the preservation of multimedia interactive installation, an important part of contemporary art. New multimedia artworks show a complex nature leading to a radical upheaval of the practice of preservation. The deep interconnection with technology is taking its toll in terms of fast obsolescence of hardware and software, which may soon become an irreversible loss. They exist only for a limited time inside an exposition (often less than a month). A computational AI-based model for preserving new multimedia art forms could be a very interesting medium-term aim.

Because of its immaterial nature, music was one of the earliest types of art to explore the creative use of new technologies: new musical forms have assumed increasing artistic importance since the second half of the last century. In the medium term, AI could be used to design and control complex installations (networks of computers and software), by means of audio-over-IP. .

8.5 Possible risks to plan for: Audio AI needs high quality data

To design robust, audio-data-driven AI-based applications to a given audio scenario, high-quality data sets are needed to train the AI components. A good data set for supervised training must be large enough to cover the different circumstances that may occur. In addition, class imbalance should be minimal, that is the number of elements in each class must be similarly balanced. Data sets for audio recordings preservation, for example, should be built from hundreds of thousands of documents obtained from several different archives.

Some use cases require data sets that comprise audio, visual and textual content. For example, the sonic "inpainting" of a silent movie could benefit from visual analytics and codified knowledge that characterises typical sounding objects identified from the movie.

Other critical factors include well-defined performance metrics and testing procedures. MPAI addresses conformance and performance attributes and rules both at the level of performance specifications as well as at the level of organization of its ecosystem governance. While only few, high-level examples are described in this section, MPAI delivers AI-based data coding standards looking at the full spectrum of applications, as will become more apparent in later chapters.

9 Video for humans and machines

9.1 DP-based video coding

Because of the limited bandwidth available for transmission and storage, we would not have reached the present state of digital television development without the series of efficient MPEG video codecs. In the last 30 years, they have been universally adopted by different product sectors. Standard codecs offer a stable environment for broadcasters and manufactures to develop their systems and services.

Why do we need compression? Digital HD 50 frame per second for studio requires a bitrate of 3 Gbit/s for, and 4K 50 frame per second requires 12 Gbit/s. Storing a 2-hour movie requires a capacity of 21 TBytes and 84 TBytes, respectively. Capacity of practically used terrestrial channel is in the range of 13 to 30 Mbit/s, and 40 Mbit/s for satellite and cable. In Italy, where digital standard definition television is typically transmitted at 2-5 Mbit/s on 12 Mbit/s channels, some four to five programmes can be broadcasted within an 8 MHz UHF channel. To achieve this, a bitrate, reduction between the high-quality studio and broadcasted video, is required.

Fortunately, video frames contain lots of redundant information, which can be exploited to reduce the bitrate. Since video coding studies began in the 1960s, various coding techniques have been explored and incorporated into compression standards starting from the early 1980s. The principal technique is the subdivision of a picture into square blocks of pixels. When the average number of bits/pixel is low, blocks may appear in a coded picture. Over the years, many algorithms have been developed to compensate for this effect.

The MPEG-2 standard, approved in 1994 after extensive research and participant market testing, continues to be one of the most widely used compression technologies in digital video.

Since then, video compression has used a block-based hybrid video codec, which basically uses the following processing elements (Figure 9):

- Exploit the spatial and temporal redundancies (Intra and Inter coding)
- Create a residual signal between the current video frame and a predicted frame

- Transform the residual signal into the frequency domain using the discrete cosine transform (Residual Coding)
- Quantise the frequency domain coefficients to maximise the zero run-lengths.
- Encode the quantised transform coefficients with an entropy encoder (Entropy Coding)

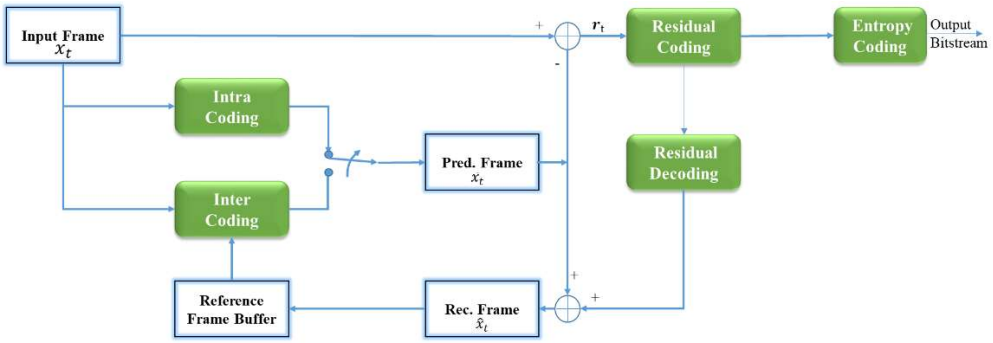

Figure 9 – Hybrid video coding scheme

In 1998 and 1999, two video coding standards were published: H.263 and MPEG-4 Visual. The MPEG-4 Advanced Video Coding (AVC) standard includes various video compression algorithms, e.g., coding structure and intra prediction. Such coding tools are still being used today. AVC is perhaps best known as the video coding format for Blu-ray discs, streaming Internet sources, HDTV broadcasts over terrestrial (ATSC, ISDB-T, DVB-T or DVB-T2), cable (DVB-C), and satellite (DVB-S and DVB-S2) systems.

Increased precision or adaptation of existing coding tools and the introduction of new coding tools has led to the High-Efficiency Video Coding (HEVC) standard, approved in 2013. The standard could rely the increased computational resources made available by the increased capabilities of semiconductor technologies, as a consequence of the Moore's law (a doubling of processing power every 18 months).

The same applies to the latest Versatile Video Coding (VVC) standard that offers a significant compression improvement over HEVC, although it uses a similar design methodology (i.e., by increasing precision and adaptation relative to HEVC).

To grasp the ideas behind the evolution of codecs it is necessary to understand the video coding standardisation process. Compression algorithms have undergone continuous improvements or refinements. These have been incorporated in MPEG standards, after competitive assessment of their visual quality merits. However, MPEG only standardised the bitstream format and the decoding process. This enables the industry to introduce a sequence of refinements at the encoder side, without altering the bitstream syntax. This

enables service providers to offer a better service without replacing end-users' equipment, i.e., set-top boxes and digital television sets.

Usually, the improvement rate declines as a negative exponential: the most gains are usually made in the first years, whereas later improvements do not significantly reduce the bitrate, approaching an asymptote. Before the asymptote is reached, a new codec is standardised and introduced into the market. At this point, there is a quantum-leap gain in compression efficiency because the new codec can take advantage of all the existing improvements as well as incorporating new concepts into its design from research.

Some codecs are application-specific, and some are designed to have wide use. MPEG-1 was intended for interactive video storage on CD-ROM. MPEG-2 was developed for video broadcast and DVD and eventually used for television studio processing. H.263 was intended for video conferencing. MPEG-4 Visual addressed a broad range of applications, including video animation and coding of multiple objects within a video frame. With its higher compression, HEVC addresses Ultra High Definition (4K), High Dynamic Range and Wide Colour Gamut. Versatile Video Coding was developed for 8K, 360 videos, screen content coding, adaptive resolution changes, and independent sub-pictures.

The principal reason for retaining a video codec is the need to sustain the economic life of consumer devices and video services. In the case of national television broadcasting, socio-political factors come into play, along with backwards-compatibility constraints.

The *market* today is still dominated by the AVC codec, approved in 2003, due to its broad decoder support and accessible licensing scheme. HEVC, approved in 2013, was expected to supersede its predecessor AVC, as it provides an improved compression performance of ~50% for comparable perceptual video quality. Its adoption, however, has been slow because of unclear royalty schemes that some say are outdated. Besides three patent pools (MPEG LA, HEVC Advance, Velos Media), there are several other Intellectual Property (IP) holders who are not members of any of them.

The MPEG-5 Essential Video Coding (EVC) standard tried to mitigate the problem. The standard has two profiles, one making use of 20+ year-old technologies. The performance of the other profile is close to the performance of HEVC. A reduced number of entities are reported to hold patents in MPEG-5 EVC and 3 have declared they would publish their licences within 2 years after approval of the standard.

This complex situation has undermined a widespread adoption of HEVC that is now being challenged by other technologies promising better performance and/or royalty free access to the compression technology. The fate of VVC is still unclear; a licence is yet to be announced.

9.2 AI-based video coding

The first video coding standard (H.120) came into being 40 years ago. Since then, a long series of video coding standards have been investigated and released. In addition to the MPEG standards mentioned, AVS, AVS2, and AVS3 continue to improve coding efficiency by introducing more adaptive coding tools, as well as offering more flexible features to enlarge the selection of rate-distortion candidates. However, almost all the designs come from suitably improved prior models.

Recently, deep learning has consistently made breakthroughs in the fields of computer vision, language processing and multimedia signal processing. There is also a large amount of deep network-based explorations in the context of video coding, especially end-to-end learned approaches.

Figure 10 shows the historical development of neural NN-based image and video compression approaches. The NN-based image roadmap is given by upper part and the video coding roadmap in the lower part, respectively.

Apparently, NN-based image compression synchronises with neural network research trends. A NN-based image compression algorithm was proposed in the late 1980s, corresponding to the time when NN back-propagation learning was proposed. Generative models for video coding were first used for high compression. Since 2016, however, a flurry of neural models for the hybrid video coding framework can be observed, where each module can also be optimised by data-trained deep networks. In addition, the end-to-end video compression solutions followed after the NN-based image compression methods began to surpass the coding efficiency of conventional codecs.

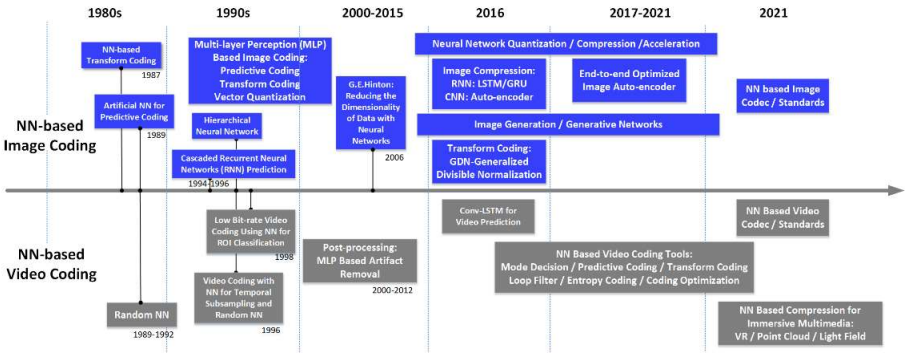

Figure 10 – Evolution of AI-based image and video coding

It has been shown that significant coding gains can be obtained with deep learning-based models in the hybrid coding framework. If every single module were realised with deep networks, then a fully deep neural network-based coding framework, called End-to-End Video Coding (EEV), can be realised.

A typical EEV model is depicted in Figure 11, which is very similar to the well-known hybrid model. The main difference is that such a framework is an end-to-end trainable and fully parametrised NN. Some expect that the EEV equipped with neural networks will signal the beginning of a new age for video compression. Moreover, the end-to-end optimisation can overcome the problem of local optimisation in hybrid coding frameworks.

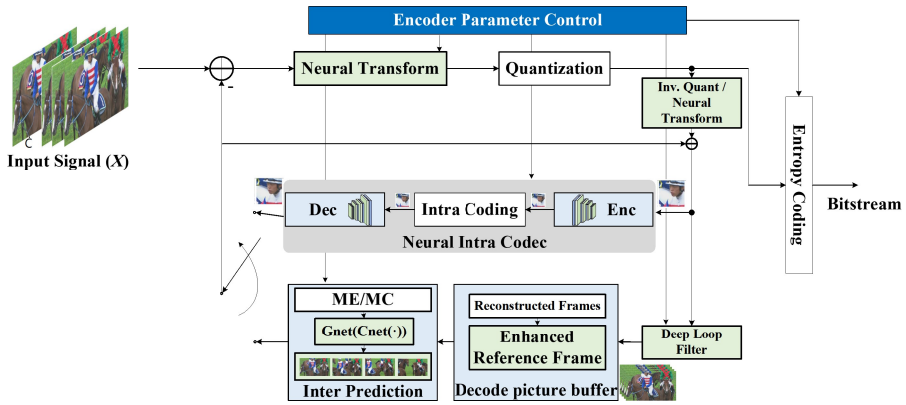

Figure 11 – A typical EEV framework

Figure 12 plots published works in this area from recent years showing that EEV is a fresh research area.

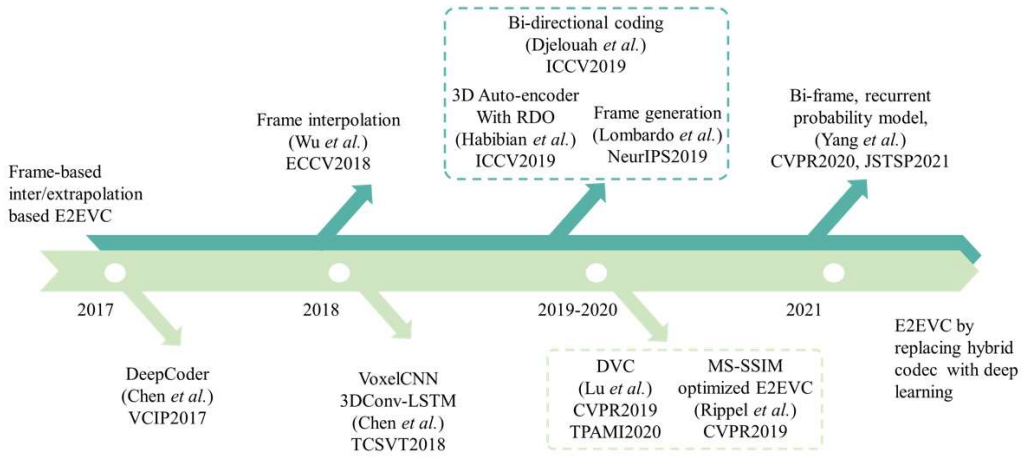

Figure 12 – Recent progress of EEV

In 2017, the first fully neural network based EEV model was proposed by Chen et al. [11]. Much like had happened for deep image coding, in the following years many optimisation and refinement models were proposed. Currently, state-of-the-art EEV models have similar or slightly better perform-

ance than the x265 codec, an HEVC open-source implementation. Based on this coding performance, however, it seems that much effort needs to be allocated to further boost EEV's coding efficiency, if the conventional video coding standards is to be surpassed. Interested readers might refer to the papers identified in the figure where two main categories of works are indicated: frame-based inter/extrapolation and hybrid codec replacement with deep networks.

9.3 Point clouds

Lidars provide the distance between a sensor and a point on a surface by measuring the time taken by the light reflected by the surface to return to the receiver. The operation wavelenghth is in the μm range – ultraviolet, visible, or near infrared light.

A Point Cloud is a set of individual 3D points, each point having a 3D position but also being able to contain some other attributes such as RGB attributes, surface normal, etc. 3D point cloud data can be applied to many fields, such as cultural heritage, immersive videos, navigation, virtual reality (VR) /augmented reality (AR) etc. Because applications are different, two different standards exists. If the points are dense, Video-based Point Cloud Compression (V-PCC) is recommended. If less so Graphic-based Point Cloud Compression (G-PCC) is recommended. The algorithms in both standards are lossy, scalable, progressive and support random access to subsets of the point cloud.

V-PCC projects the 3D space into a set of 2D planes. The algorithm generates 3D surface segments by dividing the point cloud into a number of connected regions, called 3D patches. Each 3D patch is projected independently into a 2D patch. The resulting patches are grouped into 2D frames and encoded by using traditional video technologies.

G-PCC addresses directly the 3D space to create the predictors (with algorithms that resemble the intra prediction in video coding). To achieve that, G-PCC utilizes data structures that describe the point locations in a 3D space. Moreover, the points have an internal integer-based value, converted from a floating point value representation.

Currently V-PCC offers a compression of ~125:1, a dynamic point cloud of 1 million points can be encoded at 8 Mbit/s with good perceptual quality. G-PCC provides a compression ratio up to 10:1 and acceptable quality lossy coding of ratio up to 35:1

9.4 Video for machines

After MPEG-2 was completed and when MPEG-4 was being developed, in 1996, MPEG started addressing the problem of video coding for uses other than efficient compression for transmission and storage. The scope of

the MPEG-7 standard is well represented by Figure 13 where a human looking for a content item – image, video, but audio as well – queries a machine which knows about content items through their features.

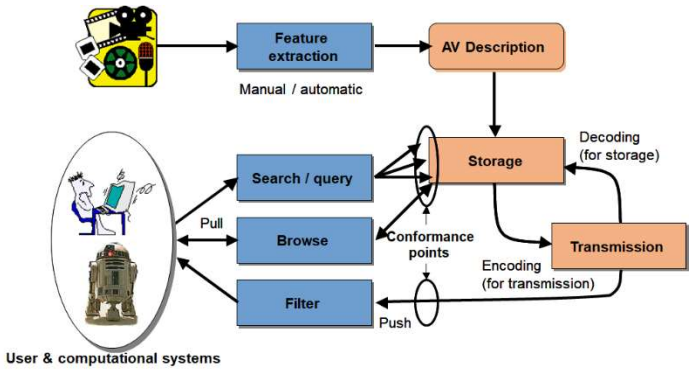

Figure 13 – The MPEG-7 standard

MPEG-7 developed a large number of Descriptors. A short list is: Color-Structure Descriptor, Texture Descriptor, Edge Histogram, Shape Descriptors, Camera Motion, Motion Trajectory, Motion Activity, Region Locator, Spatio-Temporal Locator. The descriptors have a human understandable semantics and a precise syntax.

Video key information, i.e., visual features, can be extracted, represented and compressed in a compact form. In particular, it is possible to transmit the feature stream in lieu of the video signal stream using significantly less data than the compressed video itself. Compact descriptors for visual search (CDVS) were standardised in Sep. 2015 and compact descriptors for video analysis (CDVA) in July 2019. The standardized bitstream syntax of compact feature descriptors enable interoperability for efficient image/video retrieval and analysis. CDVS is based on hand-crafted local and global descriptors, designed to represent the visual characteristics of images. CDVA is based on the deep learning features, adopted to further improve the video analysis performance.

In 2018 MPEG launched the *Video Coding for Machines* investigation. The main motivation was the large number of video content generated that is not expected to be watched by humans, if not occasionally, but is monitored by machines. The new type of video coding seeks to enable a machine to process the features received in compressed form without significant degradation.

10 Data for machines

The preceding chapters have dealt with the Moving Pictures and Audio part of the MPAI mission. However, MPAI does think that the benefits of AI can

be extended to all types of data, not just those that can be directly consumed by machines. The purpose of this chapter is to present the state of the art in some of the data types being considered by MPAI, namely, financial data, online gaming, autonomous vehicles, and genomics.

10.1 Financial data

One of the most important requests from the financial field is the ability to monitor the health of companies and detect evidence of anomalies to reduce the risk of future defaults and thus preserve business continuity. An extensive scientific literature about insolvency predictions was formulated in the late 1960s [12, 13] and during the first half of the 1980s [14, 15]. The purpose of the models reported in the literature is the identification of some indicators able to predict the level of risk and the possible default of the company by using appropriate econometric techniques.

Financial institutions, governments and in general the various market players have sought methods that are as efficient as possible and numerous researchers have pursued that goal by developing various quantitative methods, most of which are based on the statistical approach. Traditional models are accurate for about 12 months and in some cases, where the forecast has a sufficient accuracy, they are 70% precise in about 24 months [16, 17, 18, 19, 20]. This represents a severe limitation considering that the ideal forecasting model should allow medium-term predictions because the symptoms of a failure can be traced back to 5-8 years prior to failure. To overcome the characteristic limitations of statistical models, research work was carried out on pattern recognition methods developed in the field of ML. These studies have shown how ML models can offer better performance than traditional methods. Despite the greater accuracy, however, the ability to forecast in the medium term (over 24 months) remains, the main problem.

Therefore, research has focused on improving the forecasts' accuracy and on extending the time horizon. Most of the literature has focused its efforts on selecting the most appropriate financial indicators and neglected non-financial information. On the other hand, the introduction of non-financial data could improve performance in terms of accuracy and the forecasting horizon, for both traditional and ML-based models [21].

Even though improvements have been recorded over the years, there is a gap between market demands and the best available practices. The current state of the art, in fact, does not offer models that are accurate for both short and medium term, versatile in relation to different markets with an agreed tuning method, possibly as an automated process and inclusive of the capability to analyse the effects on financial and non-financial variables.

10.2 Online gaming

In the history of online gaming, developers have always been confronted with the problems arising when information moves between the clients involved in the game and the server. They approached the problems with several strategies: by formally defining strategies and protocols that could solve the most obvious problems such as the unsynchronised and smooth display of actions between clients and server; by making protocols available to clients residing behind complex networks, designed for home connections and not in a local network with direct exposure of the player's machine. After a series of evolutions and solutions such as Microsoft's Direct Play and Gamespy, to date, they are still confronted with two problems that have not found a consolidated solution: the delay or absence of data packets, and cheating players.

There have long been attempts to solve the problem of missing data in a visual way from the client's viewpoint by predicting the data that had not reached the system, based on the information available to the client at that moment. The prediction, however, was based on the data of the current game. So far, this situation sometimes still generates problems and inconsistencies.

Over time, different methods have been developed for different types of games in the case of cheating. These range from modifying the behaviour of the client by placing BOTs in charge of the more complex skills of the game to be managed (targeting opponents or making moves with very demanding timing) instead of the human player creating artificial delays in the delivery of packages to gain advantage during certain phases of the game. With the advent of authoritative online servers, however, the approach to cheating has changed. Since this type of architectural choice requires that the server always chooses and validates the actual game state, one possible way to cheat is to add visual aids to the client that give an advantage to the player not shared by other opponents. An example is provided by indicators that allow a player to immediately understand where the ball will end up in a football game.

Neural networks have already been used in several titles. The first experiments were made on classic games that could not find adequate AI models of computer opponents due to the very complexity of the game. One of these first scenarios was backgammon with TD-Gammon. For commercial video games, we find the use of neural networks in video games of different genres such as Electronic Arts' Black & White (strategy game) and racing games.

In particular, the experience of Drivatar for the game "Forza Motorsport" made people realise how much data acquisition from the styles of community players could be used to build a computerized adversary that properly competes with the player. Starting in 2019, Milestone with its MotoGP fran-

chise used the A.N.N.A. (Artificial Neural Network Agent) neural network. This technological solution has worked on the characters and skills of the riders so that they can be like their real-life counterparts and, at the same time, can adapt over time to the skill of the player.

10.3 Autonomous vehicles

An Autonomous Vehicle can move itself in the physical environment on the basis of high-level instructions received by humans or a machine and by processing data acquired from the environment. Connected Autonomous Vehicle (CAV) can send and receive data to/from other entities such as other CAVs and other devices, e.g., a traffic light of a roadside unit.

Some of the data types have an electromagnetic nature, namely:

1. Global Navigation Satellite System (GNSS).
2. Radio data from various sources and frequencies.
3. Visual data in the human visible range (400-800 THz)
4. Lidar data in the 200 THz range.
5. Radar data in the 25 and 75 GHz range.

Other data types have an acoustic nature, namely:

1. Ultrasound data in the 20 kHz range.
2. Environmental audio in the audible range (16Hz-16 kHz).

Still other data have a heterogeneous nature, namely:

1. Weather, air pressure, humidity, road conditions, etc.
2. Position, Velocity and Acceleration.

The challenge for a CAV is the creation of an internal representation of the external world that is sufficiently robust to allow it to move itself to reach the instructed destination while satisfying a number of conditions that human drivers are assumed to know, for example, by passing appropriate examinations.

Considering that the automotive market is worth ~3.6 T$ in 2021[5] and the inevitable shift toward electric and eventually autonomous vehicles, it should not surprise us that many CAVs have been designed, built, and tested, and a few are being intensely trialled[6].

10.4 Genomics

The DNA of living beings is one of the most notable examples of natural data coding, evolved spontaneously to support life on Earth. Most cells of any living organism contain instructions that guide birth, growth, life, reproduction and interaction of an astronomical number of individuals belonging

[5] Global Car & Automobile Sales - Market Size 2005–2027 https://www.ibisworld.com/global/market-size/global-car-automobile-sales/

[6] Smart Mobility Projects and Trials Across the World, https://imoveaustralia.com/smart-mobility-projects-trials-list/

to a vast number of different species. The programs lying at the core of each living cell are expressed as long polymers of 4 different basic molecules (called "nucleotides") sequentially attached at the side of very long strings of sugars. Interestingly, the cell's programming is self-interpreting – it encodes the very same tools that will be used to decode it – and is specified in terms of a very abstract structure involving several different levels of regulation. So, at the most basic level, the cellular program specifies the production of nanomolecules and can read itself, extract energy from the environment and reproduce; however, the genomes of more complex life forms also encode information about how different cells interact among themselves to form complex organs and organisms, and implicit guidelines dictating how different organisms interact among themselves to form complex ecosystems. The programs can be remarkably complex – the DNA of each human cell, which is by no means the most complex of the genomes, has roughly 3.2 billion "letters" or nucleotides, equivalent to ~0.8 GBytes of binary encoded information. Considering that there are ~40 trillion cells in the human body, the total amount of data stored in each individual can be estimated at a staggering 32 EBytes (million TBytes). Each copy of the human genome stores blueprints for ~25,000 different types of nanomachines, plus a yet not well quantified or understood number of developmental programs that regulate the translation of the program into functional, well-adapted living creatures.

Humans may think that the DNA is very "human" and "personal", but the data is not easily accessible for inspection. Costly equipment called sequencers can "read" DNA but doing so is technically challenging. The machines currently available output "noisy", i.e., unreliable reads that are short fragments randomly extracted from much longer molecules without indication of their original placement. So, a sequencer must read the same genome many times, and provide many reads, to be reasonably sure of the value of a particular nucleotide – and reconstructing the original genome out of the short fragments is a challenging operation requiring huge computational resources. While human DNA has 3.2 billion nucleotides, the output of a sequencer must have a file size in the order of hundreds of GBytes to be reliably used. In addition, some parts of the genome are hard to sequence due to biological reasons – the Human Genome[7] sequencing project is continuously updated since the production of the first draft human genome (or "reference") in 2003.

Hundreds of algorithms and computer programs performing operations on the reads obtained from sequencers have been developed over the years. One vital problem is that of establishing how the genome of each human individual differs from the idealised human reference genome established by the Human Genome sequencing project, which does not correspond to any spe-

[7] The Human Genome Project; https://www.genome.gov/human-genome-project

cific individual. While all human genomes share a very high level of sequence similarity, any two genomes differ by millions of small changes, which can be substitutions of a single DNA "letter" or more complex differences. Changes with respect to the human reference can confer desirable or deleterious traits – for instance, a single-nucleotide change in the blueprint of an essential protein can sometimes cause severe genetic disease. The ability to identify the specific characteristics of an individual's DNA is making it possible to achieve what is called "personalised medicine" – genomic data can help discover whether an individual has an ongoing disease or risks developing one. By pinpointing the causes of an individual's clinical status, genomic data processing can lead to more personalised treatments.

However, achieving such results is possible only if a suitable infrastructure is put in place. In addition to tools able to recognise the genomic variants characterising each individual and determine their clinical relevance, one typically needs strategies to encode, compress, store and access the genetic data output by sequencing machines. By using a zip-like compression function, a genomic file can be reduced by a factor of ~3-4. By applying smarter compression algorithms, the size of the file can be reduced by a much larger factor.

In general, making sense of genomic data is not easy, due to the number of hierarchically nested levels of regulation and the very abstract way information is encoded. As a result, progress in some critical areas can be very slow – and AI, with its ability to discover among vast amounts of data patterns otherwise hidden to humans, is being increasingly helpful. Recent progress with the problem of protein folding – i.e., the ability to computationally predict the 3D-structure of cellular components out of their linear blueprints stored in DNA – has made the headlines of newspapers and captured universal attention. Other typical applications are recommendation systems able to optimise therapies or treatment for individuals based on their genetic makeup.

11 Towards a responsible AI

AI has generated easy enthusiasms but also fears. As a result, a narrative has developed that sees in the development of AI more the potential for a machine-ruled dystopian future than the possibility to solve long-standing problems afflicting humanity.

It goes without saying that some of the technologies deriving from AI do hold the potential for disruptive transformation of our society – and such a potential must be kept in check if potentially serious problems are to be avoided. It is recent news that the successfully Brexit campaign might have been unlawfully influenced by the ability of the Yes campaign to offer on social media targeted political messages to a restricted demographics of voters

liable to be convinced – and whose propensity to accept the desired political message had been determined through an AI recommendation system. Another increasing concerning example are deep fakes for video and textual contents – in particular, the sometimes uncanny generative ability of advanced linguistic models such as GPT-3 holds the potential for ethically questionable outcomes. Early proposals for use cases of GPT-3 have prompted intense soul searching at OpenAI, resulting in more than a year of restricted availability and a long list of restrictions on permitted applications when the technology has finally been made available to a larger set of developers[8]. And being AI a new technology, its problems and limitations are sometimes difficult to understand and evaluate; this is illustrated by the frequent identification of training bias or vulnerabilities such as typographic attacks, which would be concerning if found in system that are mission-critical or used to make sensitive decisions.

So it is only natural that governments in several countries have placed the ability of AI systems to be trustworthy or reliable at the heart of governance models. This has been particularly relevant in Europe, USA, and China. Autonomy from intelligent systems, damage prevention, justice and explainability of algorithms are some of the pillars of the debate.

With its "White Paper on Artificial Intelligence" [22], the European Commission has developed a framework reflecting the spirit and contents of ethical guidelines. The direction appears to be based in a risk-based approach to the development of intelligent systems that focuses on the central role of human beings and respect for their dignity.

It is worth asking how it is possible to "translate" this ethical framework into investment choices aimed at the growth of the reference economic sector and the generation of benefits for the data-fuelled economy while respecting such fundamental principles. The directives that are gradually taking shape underline the need to protect privacy, algorithm transparency, workers' rights, social and gender inclusion, system interoperability between AI systems and the importance of assessing AI technique reliability.

With reference to interventions of a public nature, the guidelines identify the best strategy to innovate in a responsible manner in the field of AI, considering the centrality of the human being and the role of literacy in this area. The recommendation is to substantially increase the funds dedicated to the implementation of this strategy and support these policies.

The European model is characterised both by the centrality of fundamental rights and by the possibility of regulatory interventions in the presence of con-crete risks for European citizens. Not surprisingly, the European Commission stressed that: "international cooperation on AI issues must be based on an ap-proach that promotes respect for fundamental rights, includ-

[8] https://beta.openai.com/docs/usage-guidelines/safety-requirements

ing human dignity, pluralism, inclusion, non-discrimination and the protection of privacy and personal data" [22] and that it will strive to export its values to the world.

With technological developments, this approach will force European institutions to constantly assess the risks of emerging AI technologies, even when it is necessary to use technological infrastructures located under foreign jurisdictions, and to decide on their use. It remains to be understood whether the institutional screening will be timely enough not to undermine the public trust necessary for industry and citizens to reap the opportunities offered by AI.

When compared with other governance models such as those developed by US and China, the European framework has clearly identifiable peculiarities. In the US model, the regulation of new technologies is traditionally entrusted to private forms of self-regulation. Government intervention is therefore mild and limited to the enforcement of existing regulations, especially those that protect competition. The US strategy outlined by [23] is no exception. It prioritises AI research and mentions reliability and technical security requirements [24]. At the same time, it delimits the scope of regulatory interventions solely for the purpose of protecting civil liberties, privacy, security, and the economic interests of the country.

On the contrary, the Chinese model is focused on the exploitation, by the state, of the potential of AI [25]. Although both an ethical framework and some privacy protection standards can be found, the non-binding nature of these precepts leaves room for legislative drifts that legitimise the prevalence of the public interest over the rights of the individual. An example of this complex intertwining is represented by the Social Credit System, the mechanism for attributing an "individual social score", based on facial recognition technologies and automated data processing. The collaboration of the giant Alibaba in the realisation of the complementary Sesame project also highlights how the collaboration between public and private in the diffusion of AI systems is oriented towards the pursuit of the interests of the State.

The presence of at least three models – European Union, United States and China – however, highlights a plurality of approaches to the governance of AI and its development. The role that MPAI intends to play to give users fact-based indications of the Performance of AI systems has thus to be located at an appropriate level.

Section 2 – Using AI for the better

12 Divide and conquer

The development of a market of AI solutions has to take into account the following basic components:

- All players (e.g., Microsoft, Amazon, Google) provide environments supporting the AI application life cycle with their frameworks (e.g., Azure AI Platform, Sagemaker, Greengrass, TensorFlow etc.) and offer a store.
- In general, migration of an application to a different environment is complicated because exporting models is not easy.
- Typically, an AI application requires large and scarce multidisciplinary competences that span from data science to domain knowledge and consequently require large capital investments in both human and computational resources.
- Models and guidelines for the development of explainable AI applications are in their infancy. Current applications are monolithic and opaque making their adoption at scale problematic.

There is no overriding reason, however, for AI applications to be monolithic. If we look at the human brain, for example, we do not find a single network but a collection of connected specialised subnetworks (and sub-subnetworks).

Even at the current state of brain research, it is possible to identify and characterise the functions of subsystems of the human brain.

Whether inspired or not by this reference, in its standards MPAI divides AI applications serving identified use cases into units called AI Modules (AIM) that are defined by their functions, interfaces and input/output data. Thanks to the way they are defined, AIMs can be combined in a way that is agnostic of the AIM provider, on condition that the component has been developed according to the standard. Additionally, AIMs can be combined to address scenarios that the AIM provider may not even have foreseen. Because AI is quite a different technology than those of the past, verifying standard conformance requires steps that enable MPAI implementation users to make informed decision about their applicability. Central to this are the notions of Conformance and Performance, the latter defined as a set of attributes characterising a reliable and trustworthy implementation.

Thus, a full AI application is obtained by interconnecting the selected AIMs to create a "workflow" of AIMs, that MPAI calls AI Workflow (AIW). Again, an AIW is defined by its function as identified by the use case, its interfaces, and the format of the input/output data.

The MPAI-specified framework – called AI Framework (AIF) – allows an application implementer to develop ready-to-use systems that combine AIMs. These may:

1. Have been developed in any of the mentioned environments using any of proprietary frameworks for any operating system.
2. Be AI-based and non-AI-based.
3. Be implemented in hardware or software or in a hybrid hardware and software combination.
4. Execute in Microcontroller Units (MCU) to High Performance Computers (HPC) in local and distributed environments, and in proximity with other AIFs.

irrespective of the AIM provider. Of course, appropriate profiles of the AIF standard may need to be defined to enable such a wide range of application contexts.

Figure 14 depicts the framework components of the MPAI approach to AI data coding standardisation:

1. *AIF, AIW* and *AIM*, as described above.
2. *Controller,* a computing component that exposes Application Programming Interfaces (API) to the AIMs and to the
3. *User Agent,* the means by which a user acts with the system.
4. Two *data storage* components, one specific to an AIM and another accessible by all AIMs).
5. *Access* to external slowly varying data.
6. *MPAI Store*, a platform containing and managing a repository of AIM, AIW and AIF Implementations managed by a non-profit commercial organisation established and controlled by MPAI.

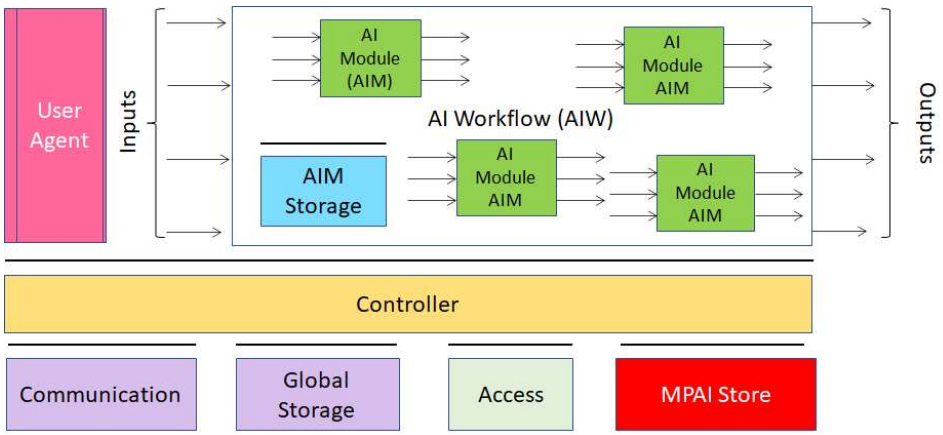

Figure 14 – The AI Framework (AIF) reference model and its components

The MPAI AI Framework offers several key advantages:

1. *Component providers* can offer conforming AIMs to an open competitive market.
2. *Application developers* can find the AIMs they need on the open competitive market.
3. *Consumers* have a wider choice of better AI applications from competing developers.
4. *Innovation* is fuelled by the demand for novel/more performing AIMs because constraints imposed by proprietary interfaces are removed and the end user's choice exclusively depends on the quality of the implementation.
5. *Society* can lift the veil of opacity from large, monolithic AI-based applications.

The operation of the Store is well described by the role played by the following actors:

1. *Implementers*:
    a. Upload their implementations to the Store.
    b. Submit their implementations to a Performance Assessor (if required).
2. *MPAI Store*:
    a. Verifies the implementation for security.
    b. Tests the implementation for conformance, i.e., that the it is a correct implementation of the standard providing a minimum level of functionality.
    c. If required, checks that the submitted implementation has been assessed by a Performance Assessors.
    d. Makes implementations available declaring their Performance grade.
3. *Users*:
    a. Pay fees on a cost-recovery basis.
    b. Download implementations.
    c. Report scores of user experience to the MPAI Store.

The MPAI Store has several desirable features for both implementers and end users:

1. It supports a market for both components (AIM) and applications (AIW).
2. It has an implementer-friendly business model like today's app market.
3. It promotes competition because different AIMs and AIWs with the same function and interfaces can be posted to the Store.
4. It offers implementations with different levels of AIW interoperability:
    a. *Level 1* – The AIW is implementer-specific but conforms with the MPAI-AIF Standard.
    b. *Level 2* – The AIW conforms with a use case specified by an MPAI application standard.

c. *Level 3* – The AIW conforms with a use case specified by an MPAI application standard and its AIMs are certified by Performance Assessors.

The ecosystem created by MPAI has several desirable features:

1. *Application developers* can build diverse applications (AIW) because AIMs can be integrated in creative ways.
2. *AIMs* can individually be of high quality because implementers can post implementations issued from a specialised field to the Store.
3. *AI technologies* can develop faster and better because the market is competitive at the level of its basic units (AIM).

The ecosystem described above will need governance. Chapter 17 defines how the MPAI ecosystem will be governed.

13 Some MPAI data coding standards

Many experts have participated in the development of MPAI data coding standards. In 15 months, MPAI has been able to develop three data coding standards:

Multimodal Conversation (MPAI-MMC): 5 use cases
Context-based Audio Enhancement (MPAI-CAE): 4 use cases
Compression and Understanding of Industrial Data (MPAI-CUI): 1 use case

Notwithstanding its short existence (established in September 2020), MPAI has already been able to publish results on its standard work at major conferences [26, 27, 28].

In the following, their work will be condensed in a few sentences and one figure. Those wishing to have a more in-depth understanding of the work done should study the standards publicly available for download (https://mpai.community/standards/resources/).

13.1 Conversation with emotion

Humans use a variety of modalities based on social conventions to communicate with other humans. The same words of a written sentence can place different emphasis, if put in different order, or even mean different things. A verbal utterance can be substantially complemented by intonation, colour, emotional charge etc. The receiver of the message may even get an utterance having widely different meanings compared to what the actual words without emphasis would mean otherwise. In some cases what words say can even be at odds with what eyes, mouth, face, and hands express over and beyond what words say.

Conversation with emotion (CWE) is a use case of the MPAI Multimodal Conversation (MPAI-MMC) standard providing a comprehensive human-machine conversation solution. Different media used by a human help the

machine fine-tune its response to a human vocal utterance thanks to its ability to understand – from speech and face – human's intention and meaning. The response of the machine can then be suitably expressed with different complementary media.

1. The machine's *Speech Recognition* AIM, *Language Understanding* AIM and *Video Analysis* AIM recognise the emotion embedded in speech and video.
2. The *Emotion Fusion* AIM fuses all Emotions into the Final Emotion.
3. The *Dialog Processing* AIM produces a reply based on the Final Emotion and Meaning from the text and video analysis.
4. The *Speech Synthesis* (Emotion) AIM produces Output Speech from Text with Emotion.
5. The *Lips Animation* AIM animates the lips of a Face drawn from the Video of Faces Knowledge Base consistently with the Output Speech.

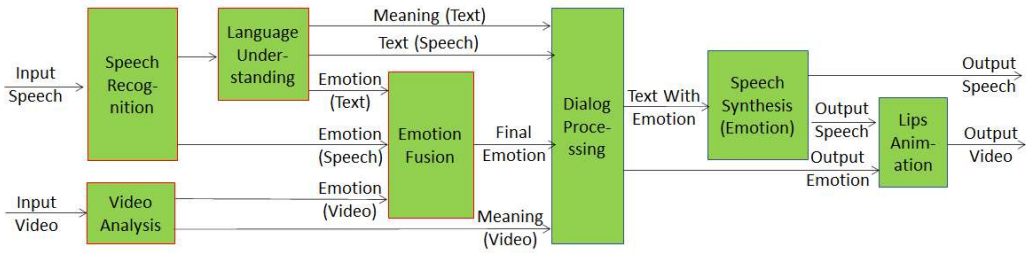

Figure 15 – Conversation with Emotion

This use case specifies function and interfaces of several reusable AIMs:

1. *Speech Recognition* where the output is text and emotion.
2. *Video Analysis* that extracts emotion and meaning from a human face.
3. *Language Understanding* that extracts meaning from text.
4. *Emotion Fusion* that fuses emotions extracted from speech and video.
5. *Dialogue Processing* that provides an emotion-enhanced text in response to text, emotion and meanings from text and video.
6. *Speech Synthesis* producing emotion enhanced speech from text and emotion.
7. *Lips Animation* producing video from speech and emotion.

Future work will open the scope of the technologies implied by the current selection of data formats and will address conversation about a scene and conversation with an autonomous vehicle. These are shortly described in Chapter 19.

13.2 Conversation about an object

Multimodal Question Answering (MQA) is a use case of the MPAI-MMC standard whose goal is to enable a machine to provide a response to a human

asking a question using natural language about the object held in their hand and the machine answer to the question with synthesised speech (Figure 16). Question and image are recognised and analysed in the following way and answers are produced in the output speech:

1. The machine's *Video Analysis* AIM analyses the input video and identifies the object in the vide producing the name of the object in focus.
2. The *Speech Recognition* AIM analyses the input speech and generates text output.
3. The *Language Understanding* AIM analyses natural language expressed as text using a language model to produce the meaning of the text.
4. The *Question Analysis* AIM Analyses the meaning of the sentence and determines the Intention
5. The *Question Answering* AIM analyses user's question and produces a reply based on user Intention.
6. The *Speech Synthesis (Text)* AIM produces Output Speech from Text in the reply.

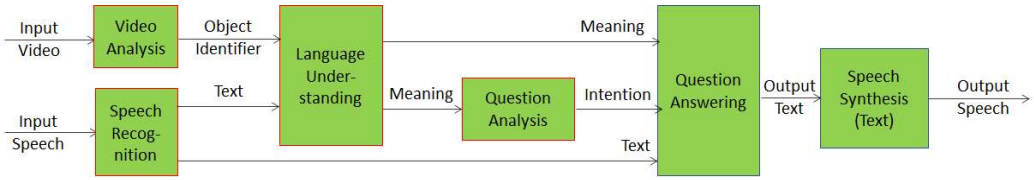

Figure 16 – Conversation about an object

This use case specifies function and interfaces of several reusable AIMs:

1. *Video Analysis* that extracts object identifier from the input video.
2. *Speech Recognition* where the output is text.
3. *Language Understanding* that extracts meaning from text integrated with the object name.
4. *Question Analysis* that extracts intention from the text with meaning.
5. *Question Answering* that provides a reply in text in response to text, intention and meanings from text and video.
6. *Speech Synthesis (Text)* producing speech from text.

13.3 Feature-preserving speech translation

Unidirectional Speech Translation (UST) is a use case of the MPAI-MMC standard enabling a user to obtain a spoken translation of their utterances to a specified language preserving the speaker's vocal features.

This workflow is designed to

1. Accept speech as input.
2. Convert that speech to text (using a *Speech Recognition* AIM).

3. Translate that text into text of another language (using the *Speech Translation* AIM).
4. Analyse the input speech (using the *Speech Feature Extraction* AIM).
5. Synthesize the translated text as speech that retains specified aspects of the input voice, e.g., timbre (colour) (using the *Speech Synthesis (Features)* AIM.
6.

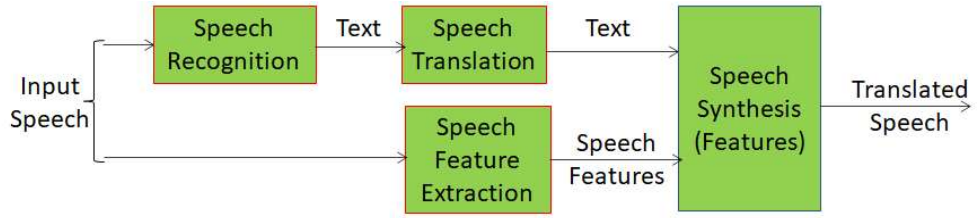

Figure 17 – Feature-preserving speech translation

For the purpose of point 4., the workflow utilizes, in addition to the *Speech Recognition* module, two more specialized speech-related modules: one – *Speech Feature Extraction* – extracts relevant speech features from the input speech signal; and another – *Speech Synthesis (Features)* – can incorporate those features, along with specification of the text to be spoken aloud, when synthesizing. All these modules can be reused in other workflows, and thus exemplify MPAI's standard module philosophy.

The MPAI-MMC standard supports two additional use cases. *Bidirectional Speech Translation* combines two flows to enable a bidirectional conversation and *One-to-many Speech Translation* enables a single speech to be simultaneously translated into a set of languages for distribution to appropriate recipients.

Several other useful configurations are currently subject to research and development and may become additional AIFs before long.

13.4 Emotion enhanced speech

Speech carries information not only about its lexical content, but also about several other aspects including age, gender, identity, and emotional state of the speaker. Speech synthesis is evolving towards support of these aspects.

In many use cases, emotional force can usefully be added to speech which by default would be neutral or emotionless, possibly with grades of a particular emotion. For instance, in a human-machine dialogue, messages conveyed by the machine can be more effective if they carry emotions appropriately related to the emotions detected in the human speaker.

Emotion-Enhanced Speech (EES) is a use case of the Context-based Audio Enhancement (MPAI-CAE) standard that enables a user to indicate a model utterance or an Emotion to obtain an emotionally charged version of a given utterance.

CAE-EES implementation can be used to create virtual agents communicating as naturally as possible, and thus improve the quality of human-machine interaction by bringing it closer to human-human interchange.

The CAE-EES Reference Model depicted in Figure 18 supports two modes implemented as pathways enabling addition of emotional charge to an emotionless or neutral input utterance (Emotion-less speech).

1. Along *Pathway 1* (upper and middle left in the Figure), a Model Utterance is input together with the neutral Emotionless Speech utterance into the *Speech Feature Analyser1*, so that features of the former can be captured and inserted into the latter by the *Emotion Inserter*.
2. Along *Pathway 2* (middle and lower left in the Figure), neutral Emotionless Speech utterance is input along with an identifier of the desired Emotion(s). *Speech Feature Analyser2* extracts Emotionless Speech Features that describe its initial state from Emotionless Speech and sends them to *Emotion Feature Inserter* that produces the Speech Features that specify the same utterance as Emotionless Speech, but now with the desired emotional charge. Speech Features are sent to *Emotion Inserter*, which uses the Speech Features set to synthesize Speech with Emotion.

Emotion-Enhanced Speech, designed for use in entertainment and communication, exploits the technology for creation of embeddings within vector spaces that can represent speech features conveying emotions. It's designed to enable addition of emotional features to a bland preliminary synthetic speech segment, thus enabling the artificial voice to perform – to act!

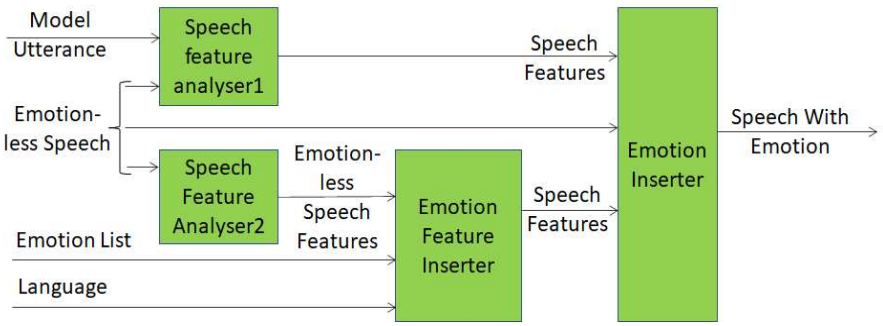

Figure 18 – Emotion enhanced speech

13.5 Speech restoration system

The goal of the *Speech Restoration System* (SRS) use case of the MPAI-CAE standard is to restore a damaged segment of an audio segment contain-

ing speech from a single speaker. An AIM is trained to create a NN-based speech model of the speaker. This model is used by a speech synthesiser which receives the text of the damaged segment and produces the synthetic version of the damaged speech. The Assembler replaces the damaged segment.

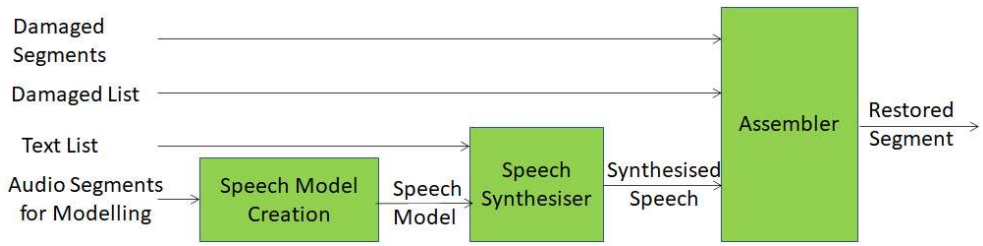

Figure 19 – Speech restoration system

13.6 Audio recording preservation

For many international audio archives, there is an urgent need to digitise all their records, especially analogue magnetic tapes, which have a short life expectancy, especially when compared to paper records. Although international institutions (e.g., International Association of Sound and Audiovisual Archives, IASA; World Digital Library, WDL; Europeana) have defined several guidelines (not always fully compatible with each other), there is still a lack of international standards.

The introduction of this MPAI use case in the field of active preservation of audio documents opens the way to effectively respond to the methodological questions of reliability with respect to the recordings as documentary sources, while clarifying the concept of "historical faithfulness". In the magnetic tape case, the carrier may hold important information: multiples splices; annotations (by the composer or by the technicians) and/or display several types of irregularities (e.g., corruptions of the carrier, tape of different colour or chemical composition).

The *Audio Recording Preservation* (ARP) use case of the MPAI-CAE standard focuses on audio read from magnetic tapes, digitised and fed into a preservation system.

Audio data is supplemented by the data from a video camera pointed to the head reading the magnetic tape. The output of the restoration process is composed by a preservation master file that contains the high-resolution audio signal and several other information types created by the preservation process. The goal is to cover the whole "philologically informed" archival process of an audio document, from the active preservation of sound documents to the access to digitised files.

Cultural heritage is one of the fields where AI can have a significant impact. This technology can drastically change the way we preserve, access, and add value to heritage, making its safeguarding sustainable. Audio archives are an important part of this heritage, but require relevant resources in term of people, time, and funding.

CAE-ARP provides a workflow for managing open-reel tape audio recordings. It is an important example of how AI can drastically reduce the resources necessary to preserve and make accessible analogue recordings.

A concise description of the operation of the ARP workflow of Figure 20 is given by:

1. The *Audio Analyser* and *Video Analyser* AIMs analyse the Preservation Audio File (a high-quality audio signal) and the Preservation Audio-Visual File [29].
2. All detected irregularity events for both Audio and Image are sent to the *Tape Irregularity Classifier* AIM, which selects the most relevant for restoration and access.
3. The *Tape Audio Restoration* AIM uses the retained irregularity events to correct potential errors occurred at the time the audio signal was converted from the analogue carrier to the digital file.
4. The Restored Audio File, the Editing List (used to produce the Restored Audio File, the Irregularity Images, and the Irregularity File containing information about irregularity events) are inserted in the Packager.
5. The *Packager* produces the Access Copy Files to be used, as the name implies, to access the audio content and the Preservation Master Files, with the original inputs and data produced during the analysis, used for preservation.

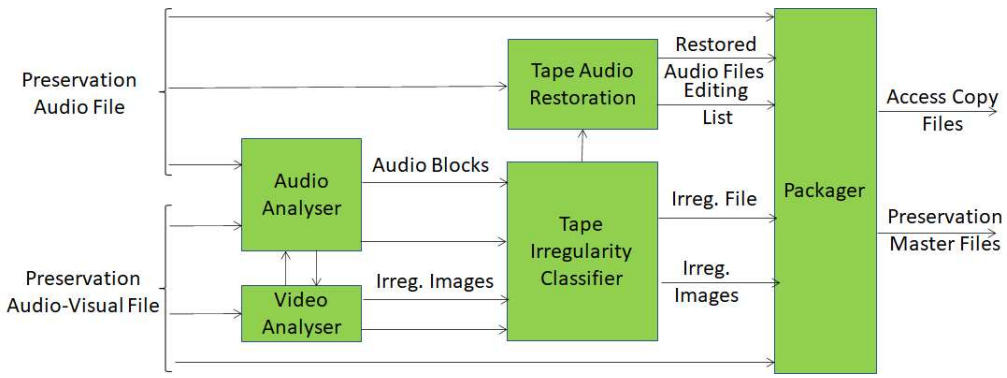

Figure 20 – Audio recording preservation

The overall ARP workflow is complex and involves different competences both in audio and video. Therefore, the MPAI "divide and conquer"

approach is well-suited to promote advancement of different algorithms and functionalities because it involves different professionals or companies.

Currently, ARP manages mono audio recordings on open-reel magnetic tape, but the objective is to extend this approach to complex recordings and additional types of analogue carrier such as audiocassettes or vinyl.

13.7 Enhanced audioconference experience

How audio for humans is addressed within MPAI can be reviewed by analysing an important use case of the *Enhanced Audioconference Experience* (EAE) of MPAI-CAE. To connect distant users requires an audio conference application whose quality is limited by the environmental and machine sound conditions. Very often, it is far from satisfactory because of multiple competing speakers, non-ideal acoustical properties of the physical spaces that the speakers occupy and/or background noise. These can lead to reduced-intelligibility speech resulting in distracted or not fully understanding participants, and may eventually lead to what is known as *audioconference fatigue*.

The main demand from the transmitter side of an audio conference application is to detect the speech and separate it from unwanted disturbances. By using AI-based adaptive noise-cancellation and sound enhancement, those kinds of noise can be virtually eliminated without using complex microphone systems that capture environment characteristics.

Similarly, the experience at the receiver side can be substantially enhanced by spatial presenting speech to offer a more life-like overall experience to the listener. This requires bridging such technology gaps as extracting a spatial of sound fields description, and speech and audio objects from microphone arrays. Transcoding HOA to OBA, generation of HOA and binaural reverberation and individualisation of binaural audio are required technologies [30].

Data-driven or ML based AI used in source separation and noise cancellation can be evaluated by analysing the limitations of the applica-tions. For example, if there is a wide range of noise in the environment, data-driven AI would fail because of limited use of dataset materials for training. In fact, ML-based AI methods can be successfully used to create more robust solutions for these cases.

The reference model of this use case (Figure 21) gets the microphone array audio as input for sound field description. Microphone array geometry also represents the locations of each microphone used for recording the scene. While the perceptual sound field can be represented in 3D hearable angles azimuth and elevation, the limitations over the number of microphones and field coverage for horizontal or vertical planes reduce the performance for the operational environment.

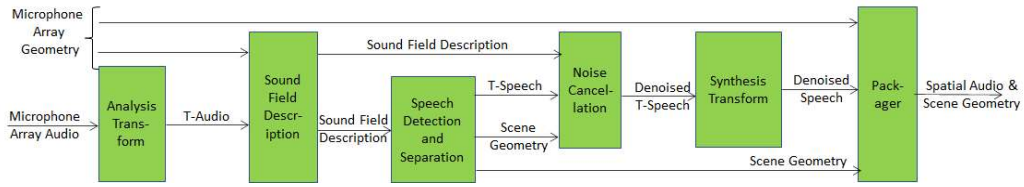

Figure 21 – Enhanced Audioconference Experience

The sequence of EAE operations of the following:

1. *Analysis Transform* AIM transforms the Microphone Array Audio into fre-quency bands via a Fast Fourier Transform (FFT) to enable the following operations to be carried out in discrete frequency bands.
2. *Sound Field Description* AIM converts the output from the Analysis Transform AIM into the spherical frequency domain. If the microphone array capturing the scene is a spherical microphone array, Spherical Fourier Transform (SFT) can be used to obtain the Sound Field description that represent the captured sound field in the spatial frequency domain.
3. *Speech Detection and Separation* AIM receives the sound field description to detect and separate directions of active sound sources. Each separated source can either be a speech or a non-speech signal.
4. *Noise Cancellation* AIM eliminates audio quality-reducing background noise and reverberation.
5. *Synthesis Transform* AIM applies the inverse analysis transform on the received Denoised Transform Speech.
6. *Packager* AIM
    a. Receives Denoised Speech and Audio Scene Geometry.
    b. Multiplexes the Multichannel Audio stream and the Audio Scene Geometry.
    c. Produces one interleaved stream containing separated Multichannel Speech Streams and Audio Scene Geometry.

This standard specifies workflow and interfaces between AIMs for enhanced audio conference experience from speech detection and separation to noise cancellation and audio object packaging. Receiver can directly reach dominant denoised speech at the other side in the audioconference call. As an immersive audio alternative, receiving the packaged separated and denoised speech which are outputs from the speaker side can be used to post-produce the same event as the receiver is spatially inside in the event.

While immersive audio applications are growing in popularity, the hardware advances in microphone arrays, i.e., increasing the number of micro-

phones, make a broad range of AI techniques available to acoustic scene analysis.

The current specification is limited to a minimum number of 4 microphones placed in a tetrahedral array shape. The next objective is to decrease the minimum necessary number of microphones without degrading the separation performance.

13.8 Company performance prediction

The fact that one of the most requested figures in the labour market is the data scientist's role shows that data analysis is increasingly assuming the role of an essential activity for companies, financial institutions and government administrations. The sub-phases of risk assessment are currently based on data management, and analysis and data collection can account for 75% of Risk Management & Business Continuity processes. Data analysis is also necessary to monitor the business situation and make more informed decisions for future strategies. At the same time, regulators require ever greater transparency from an analysis of a large amount of data. This is a highly time-consuming activity, and it is not always possible to obtain the most relevant information from large amounts of data.

With the *Company Performance Prediction* (CPP) use case of the MPAI Compression and Understanding of Industrial Data (MPAI-CUI) standard, MPAI provides initial answers to these increasingly pressing needs by introducing a versatile AI-based standard able to offer the most accurate predictions possible in both the short and medium term. Solutions characterised by a new process of tuning ML methods to improve performance because if we want good education for humans, we should do good training for AI.

CUI-CPP is the technical specification designed as a decision support system to solve these problems in the financial field targeting the assessment of a company from its financial, governance and risk data. ML and inference algorithms make it possible to predict a company's performance for a time horizon up to 60 months.

The CUI-CPP workflow depicted in Figure 22 shows that Company Per-for-mance Prediction is captured by the following three measures:

- *Default Probability*: the probability of company default (e.g., crisis, bankruptcy) dependent on financial and governance features in a specified number of future months.
- *Organisational Model Index*: the adequacy of the organisational model (e.g., board of directors, shareholders, familiarity, conflicts of interest).
- *Business Discontinuity Probability*: the probability that company operations are interrupted for a duration less than 2% of the prediction horizon.

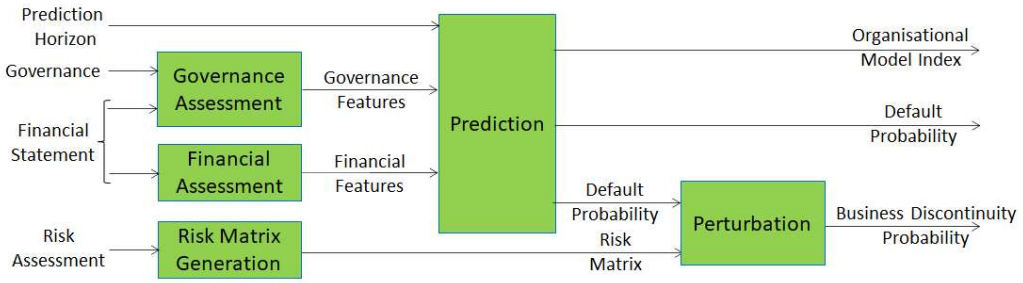

Figure 22 – Company performance prediction

MPAI-CUI can be used for several purposes:

1. *To support the company's board* in deploying efficient strategies by analysing the company financial performance and identifying possible evidence of crisis or risk of bankruptcy years in advance. The board can take actions to avoid these situations, conduct what-if analysis, and devise efficient strategies.
2. *To assess the financial health* of companies applying for funds. A financial institution receiving a request for funds, can access the company's financial and organisational data and assess, as well as predict future performance. Financial institutions can make the right decision whether funding the company or not, based on a broader vision of its situation.
3. *To assess public policies* and scenarios of public interventions in advance of their application, as well as to identify proactive actions to increase resiliency of countries. An example of this use is reported in [31] where the socio-economic effects of financial instruments on the performance and business continuity of the beneficiary companies shows how AI can support public decision-makers in creating and deploying regional policies.

In a nutshell, CUI-CPP is a powerful and extensible way to predict the performance, simplify analyses and increase efficiency of a company.

14 Structure of MPAI standards

MPAI produces AI-based data coding standards. But what is a "standard"? For sure there is a technical document specifying how things should be done, but MPAI adds to this a reference software implementation, normatively equivalent to the technical specification. Then there is a specification to test that an implementation has been implemented in a technically correct fashion. Finally, MPAI adds a specification to assess how well an implementation "performs", a multi-dimensional notion meaning, e.g., that the implementation is unbiased.

Therefore, MPAI defines a *standard* as a *collection* of four documents with associated software and data sets.

14.1 Technical Specification

The first document is the Technical Specification (TS), the document that contains normative clauses to be strictly followed by a user wishing to develop a conforming implementation. There are two types of TS: system-oriented and application oriented. The former concerns support for AI operation, such as the MPAI-AIF standard, and the latter concern application of AI to specific domains such as MPAI-CAE, MPAI-MMC and MPAI-CUI.

An MPAI application standard is typically a container of applications called use cases. For instance, the Multimodal Conversation TS contains, as of today, Conversation with Emotion (CWE), Multimodal Question Answering (MQA), Unidirectional Speech Translation (UST), Bidirectional Speech Translation (BST) and One-to-Many Speech Translation (MST). Each TS is identified by 3 characters (e.g., MMC) and each use case is also identified by 3 characters, e.g., CWE.

For each use case, the TS specifies the AI Workflow (AIW) that implements the use case with:

1. The function executed by the AIW.
2. The syntax and semantics of the AIW's input and output data.
3. The topology of the AIMs composing the AIW.
4. For each AIM
    a. The function executed by the AIM.
    b. The syntax and semantics of the AIW's input and output data.

The TS includes the syntax and semantics of all data formats used by all use cases in a single chapter. This is done because quite a few of the data formats are shared across AIMs. Some are also shared by different standards and MPAI plans on developing a standard collecting all AIMs used by more than one standard.

14.2 Reference Software

The second component of an MPAI standard is the Reference Software. Ideally, an RS is the expression of the TS in a computer language, as opposed to the natural language used to express the TS. It is a technically correct implementation of the TS in the sense that its AIW and AIMs perform as the TS specifies.

The RS is composed of:

1. Software implementing the TS released as a source code implementation of the AIF and the AIWs exposing all AIM interfaces with the full set of:
    a. High-performance source code AIMs, or
    b. Limited-performance source code AIMs, or
    c. Sufficiently high-performance compiled AIMs, not to be used for commercial implementations unless the AIM provider agrees.

2. Sample input data or a data generating environment or endpoint for trialling the RS.
3. A knowledge base conforming with the standard in case the RS requires use of a knowledge base for access by those using the RS.

The RS is distributed with the MPAI modified Berkeley Software Distribution (BSD) licence.

14.3 Conformance Testing

If a TS can be considered as the "law", i.e., the set of rules that implementers have to follow to develop correct implementations, Conformance Testing (CT) can be considered as the "tribunal" determining whether an implementation is indeed technically correct.

Conformance testing is not unknown in standardisation. Indeed, MPEG had always developed the conformance testing of its standards. However, the issue with digital media is that there is typically an "encoder" producing data that a "decoder" can decode. Therefore, conformance testing could be formulated as "provide bitstreams and check that the decoder under test can correctly decode them" and "feed the bitstreams produced by an encoder and check that the reference software decoder can correctly decode them". Although digital media are, well, digital, in general, two digital media decoders may very well not decode the same bitstream in the same way. The reason is because two decoders may have a different initial state, and different precision levels may have been used in the many computations performed by a decoder. While different, they may very well pass the conformance testing.

In the MPAI world the difference of the outputs from different implementa-tions, more than the exception is the norm, because most AIMs contain neural networks of unspecified architectures, trained with unspecified data sets. The MPAI CT specifications define the procedure, the tools, the process, and the data to be used to Test the Conformance of an implementation and specifies the tolerance of the output of an AIM given the input data used for the Test.

An example of MPAI CT is the following. In Conversation with Emotion (MMC-CWE) there is an AIM whose function is to take input speech and produce as output the text corresponding to the input speech and the emotion contained in the input speech (Figure 23).

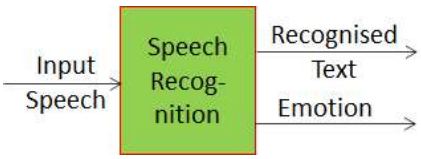

Figure 23 – An example of MPAI Conformance Testing

In this case Conformance is defined as the ability of an Implementation to produce Text expressed as Unicode and Emotion expressed as one of the MPAI standard Emotions. This is important because, to be able to interconnect and do something useful with the data, an AIM must receive them in the right format. For a user of the system, however, knowing that the data are syntactically and semantically correct is not particularly useful if Recognised Text has little resemblance with what was contained in the Input Speech and the Emotion is declared as Angry when in the Input Speech it was Happy.

Imposing that an AIM implementation is Conforming only when the AIM does a perfect job is not realistic either, because no implementation can be perfect in all cases. Giving a grade to a speech recogniser is a known problem that amounts to assigning a word error rate below which an implementation is accepted. For Emotion, however, the story is different because there is no established practice. MPAI is therefore considering three possibilities to measure the "emotion error rate": 1) use human testers, 2) train a network to measure the distance between Emotions, or 3) define an emotion space with suitable metrics.

This shows that even a seemingly "boring" topic like Conformance Testing can become an attractive field of investigation.

14.4 Performance Assessment

In the Introduction, we mentioned that AI has a subtle way of appearing reliable when it is not, whether by design or not. This most excruciating topic is engaging research in several affected fields.

MPAI addresses this issue by defining Performance of an MPAI standard implementation as the set of the following attributes: Reliability, Robustness, Fairness and Replicability. MPAI gives the following meanings to the four words:

1. *Reliability*: implementation performs as specified by the standard e.g., within the application scope, with stated limitations, and for the period of time specified by the Implementer.
2. *Robustness*: the implementation can cope with data outside of the stated application scope with an estimated degree of confidence.
3. *Replicability*: the assessment made by an entity can be replicated, within an agreed level, by another entity.
4. *Fairness*: the training set and/or network is open to testing for bias and unanticipated results so that the extent of applicability of the system can be assessed.

It should also be clear that Performance is not a yes/no attribute but can have "grades", possibly depending on the specific domain to which AI is applied.

Performance Assessment, the fourth component of an MPAI standard, is the specification that defines the data sets or their characteristics, the tools, the procedures, and the grades used to Assess the Performance of an implementation.

15 Some technologies from the MPAI repository

15.1 Emotion

Emotion descriptors and examples are widely used in current MPAI standards. For example, the Emotion-Enhanced Speech Use Case of the MPAI-CAE standard handles emotion via two different modalities: users can supply model utterances demonstrating the desired emotion for a synthetic speech segment; or they can specify the desired emotion for that synthetic segment using a label supplied by MPAI as a data type with its own digital representation, e.g., "angry".

Notably, MPAI is the first standards organisation to have standardised a numbered list of Emotions. (Note, however, that the list can be modified or replaced by implementers, as explained below.)

In MPAI, Emotions are defined by the following data set:

1. *EmotionType*, a high-level category of Emotions within the mentioned list of Emotions, e.g., "FEAR."
2. *EmotionDegree*, one of the values "high," "medium," and "low."
3. *EmotionSet*, a data structure that specifies a set of EmotionTypes and EmotionNames proposed to augment or replace the standard MPAI set.
4. *EmotionName*, the label of an emotion, whether general, e.g., "fearful/scared" or more specific, e.g., "terrified."
5. *EmotionSetName*, the name of an EmotionSet data structure.

The Basic Emotion Set is a table that currently identifies 16 EmotionTypes (high-level emotion categories), e.g., "FEAR", "HURT" and "APPROVAL, DISAPPROVAL". Each current EmotionType contains one or more general-level Emotions. For example, "FEAR" happens to contain one Emotion, "fearful/scared"; "HURT" contains "hurt" and "jealous"; and the "APPROVAL, DISAPPROVAL" category contains "admiring/approving," "disapproving," and "indifferent." EmotionTypes (catgories) can also include more specific or subcategorized emotions. For instance, the "FEAR" EmotionType includes "terrified" and "anxious/uneasy," while the "APPROVAL, DISAPPROVAL" category includes "awed" and "contemptuous." In total, the Basic Emotion Set lists 60 general or more specific Emotions.

The Emotion data type is extensible in the sense that an implementer may submit a proposal that extends or replaces the Basic Emotion Set. The proposal will be assessed by the Development Committee in charge and, if approved for consistency, posted on the MPAI web site for use.

15.2 Intention

MPAI defines Intention as the result of analysis of the goal of a question. The "intention" consists of the following elements: qtopic, qfocus, qLAT and qSAT. These are exemplified by the question: Who is the author of King Lear?" The result of question analysis concludes the domain of the question is "Literature," the topic of the question is "King Lear", and the focus of the question is "Who." More precise definitions are:

1. *qtopic* is the topic of the question, the object or event the question is about.
2. *qfocus* is the focus of the question, which is the part of the question that, if replaced by the answer, makes the question a stand-alone statement. Ex. What, where, who, what policy, which river, etc.
3. *qLAT* is the Lexical Answer Type of the question. For example, "author" is qLAT in "Who is the author of King Lear?"
4. *qSAT* is the Semantic Answer Type of the question. qSAT corresponds to Named Entity type of the language analysis results. For example, "person" is qSAT in "Who is the author of King Lear?"
5. *qdomain* is the domain of the question such as "science", "weather", "history".

The information in the Intention is used to find the answer to the user's question which matches best with topic, focus, answer types and domain by measuring reliabilities of the candidate answers extracted from sentences in the Knowledge Base.

15.3 Meaning

MPAI defines Meaning as information – semantic, but also syntactic and structural – extracted from input data, i.e., Text, Speech, and Video. The "meaning" consists of: POS_tagging, NE_tagging, Dependency_tagging and SRL_tagging defined as:

1. *POS_tagging* indicates the results of tagging Part Of Speech (POS) such as noun, verb, etc. including information on the POS tagging set and tagged results of the question.
2. *NE_tagging* indicates NE results of tagging Named Entities (NE) such as Person, Organisation, Fruit, etc., including information on the NE tagging set and tagged results of the question.
3. *dependency_tagging* indicates results of tagging dependency, i.e., the structure of the sentence such as subject, object, head of the relation, etc., including information on the dependency tagging set and tagged results of the question.
4. *SRL_tagging* indicates tagging results of Semantic Role Labelling (SRL), i.e., the semantic structure of the sentence such as agent, location, patient role, etc., including information on the SRL tagging set and tagged results of the question.

The semantic and structural information contained in the Meaning is used as features by other AIMs to decide the user's intention (Question Analysis),

the reply to the question (Question Answering) or how the dialog should continue (Dialog Processing).

15.4 Speech features

MPAI defines speech features as descriptive aspects of a speech segment. These include base speed, pitch, and volume; variations in pitch, intensity, and sub-segment duration (rhythm); vocal tension, degree of whisper or creakiness, and others. The features can be represented symbolically, e.g., indicating a certain intensity (volume) in decibels; or they can be represented via neural-network-based vectors (NNspeechFeatures). Either representation may be automatically recognised and extracted for use in, e.g., speech analysis or speech synthesis.

To describe some speech features more exactly:

1. *Pitch*: the fundamental frequency of speech expressed in Hz.
2. *Intensity*: the energy of speech expressed as dB.
3. *Speed*: the speech rate expressed as a number indicating specified linguistic units (e.g., phonemes, syllables, or words) per second.

Speech features can be used for voice analysis, e.g., to assist recognition of vocally expressed emotion; or they can be used for voice synthesis, e.g., to lend a certain emotional charge to a synthetic voice. When speech features are passed between AIMs, the receiving module will require precise specification of their format and, if the format exploits network-based vectors, pretrained models, or sufficient training data.

15.5 Microphone array geometry

A microphone array consists of several microphones placed over a platform. It is used to record the environment from different locations. These arrays are used in a variety of applications aiming noise cancellation, source separation or source localisation. Multichannel outputs of the array may be used in data-driven or ML based AI applications. EAE is a real-time use case that analyses the multichannel signals from a microphone array. Since one of the EAE inputs is multichannel audio, microphone array geometry is another input format to define the properties of the input signals. From the definition obtained from microphone array geometry, the analysis with different types of microphone arrays over AIMs is possible.

Microphone arrays are described with the following features:

1. *Microphone Array Type* defines the shape of the platform where the microphones are placed. It would be in spherical, circular, planar, linear or another format.
2. *Microphone Array* is formed with a *Number of microphones*.
3. *Microphone object* consists of the properties of specific microphone placed over the platform. It contains the microphone position in x, y, z

coordinates with respect to the central reference position. The directivity pattern of the individual microphone would also be set as omnidirectional, figure of eight, cardioid, supercardioid, hypercardioid or another. The microphone looking directions would also be set as a vector in x, y, z coordinates.

4. *Microphone array look direction* is a reference vector represented in x, y, z coordinates.
5. The type of the platform to form a microphone array may differ to place selected number of microphones. Microphones can be placed over a rigid or open surface for a compact form. *Microphone Array Scattering Type* defines the scattering surface of the platform which would have an effect over the frequency components of the sound field. During the analysis, the microphone array surface type, i.e., rigid, open or another must be regarded by AIMs.
6. Microphone manufacturing process varies depending upon the types of microphones and manufacturers. It is not possible to obtain the same frequency characteristics between the microphones even if the microphone type and the manufacturers are the same. Therefore, *Microphone Array Filter URI* is required as the equalisation filter address that defines the filter coefficients to equalise each microphone output each other.
7. Additionally, the format requires, *Sampling Rate, Sampling Type* and *Block Size as* the number of samples in an audio block.

By using the microphone array geometry information with the microphone array audio, the AIMs defined for the EAE make the speech detection and separation and noise cancellation applicable via adapting their runtime according to the input definitions.

15.6 Audio scene geometry

The EAE output contains the separated speech signals and the audio scene geometry, i.e., the information that needs to be sent to a receiver in order to correctly recreate the audio field intended by the transmitter.

EAE packages the separated multichannel speech with their spatial information given in the audio scene geometry for transmission through an audio conference application that may improve the speech source quality or enable the immersive audio.

Therefore, the EAE output should include:

1. *Speech Count* represents the number of speech objects detected during the current audio block.
2. *Speech Object* contains the spatial information of the detected speech. Each object is represented with SpeechID. The single channel identifier over multichannel audio is the ChannelID. Spatial information of the

speech object detected in the current block is represented by azimuth, elevation and distance.
3. *Block Information* represents the current index, starting and ending time of the block.

Section 3 – AI needs more than standards

16 MPAI mission and organisation

According to the Statutes, the work described so far has been the result of a collective effort of members in a not-for-profit organisation

... incorporated under the laws of Switzerland with the mission to promote the efficient use of Data by (A) developing Technical Specifications of (1) Coding and decoding for any type of Data, especially using new technologies such as Artificial Intelligence, and (2) technologies that facilitate integration of Data Coding and Decoding components in Information and Communication Technology systems, and by (B) bridging the gap between Technical Specifications and their practical use through the development of Intellectual Property Rights Guidelines ("IPR Guidelines"), such as Framework Licences and other instruments.

This text basically says that the purposes of the MPAI organisation are: (1) to explore coding techniques for any type of data, especially using AI; (2) to develop technologies to be able to integrate data coding into larger systems; and (3) to help mitigate the risk of being unable to use a standard after having invested in its development.

There are two classes of membership. Principal Members have the right to decide matters concerning MPAI. For this, they pay full membership fees. On the other hand, Associate Members pay very low membership fees – 20% of the full fees – but do not have a say in major matters such as the MPAI organisation, approval of standards and IPR matters. They have, however, the same right to participate in the development of standards.

MPAI is obviously keen on having wide participation in its work. However, it places as a condition that members 1) be legal entities or represent a university department and 2) can demonstrably contribute to the development of MPAI standards.

The MPAI organisational structure is depicted in Figure 24.

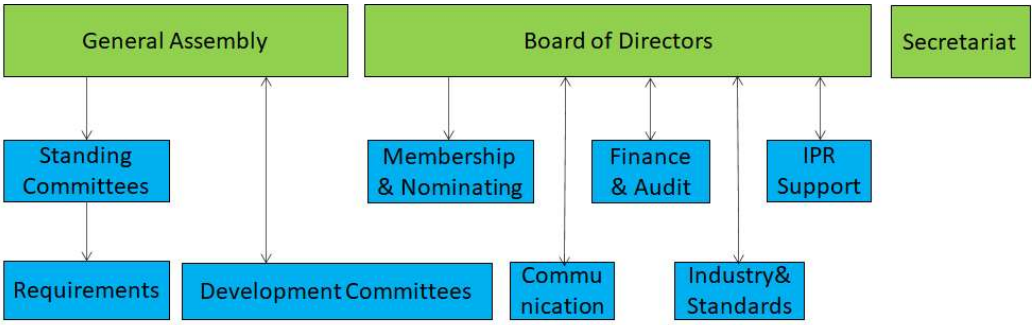

Figure 24 – The MPAI organisational structure

A summary description of the MPAI structure is as follows:

1. The *General Assembly* (GA) is the MPAI supreme body, attended by the Principal Members and Associate Members as observers. The MPAI technical work is under the supervision of the GA, currently in two branches:
    a. *Requirements* to discuss standard projects and
    b. *Development Committees* (DC) to develop standards. Currently there are 4 DCs who have all developed at least one Technical Specification:
        i. AI Framework.
        ii. Context-based Audio Enhancement.
        iii. Multimodal Conversation.
        iv. Compression and Understanding of Industrial Data.
2. The *Board of Directors* is currently composed of 5 members. The Statutes prescribe a balanced geographical representation of the main interests in data compression. Typically convened twice a month, it currently has 5 Advisory Committees:
    a. *Membership and Nominating*: Reviewing membership applications and nominating Officers.
    b. *Finance and Audit*: Advising the Board on finances.
    c. *IPR Support*: developing Framework Licences.
    d. *Communication*: Managing MPAI communication.
    e. *Industry and Standards*. Advising on matters related to relations with external entities.
3. The *Secretariat* performs basic functions such as IT management, meeting support, communication with members etc.

It was important to describe the MPAI organisation, but more important the process through which an MPAI standard progresses from an idea to a set of documents and software that industry can use. Being the vehicle that can change industry shape and consumer life, standards are serious business. Just think of MP3.

Therefore, MPAI has adopted a rigorous standards development process, whose steps are depicted in Figure 25.

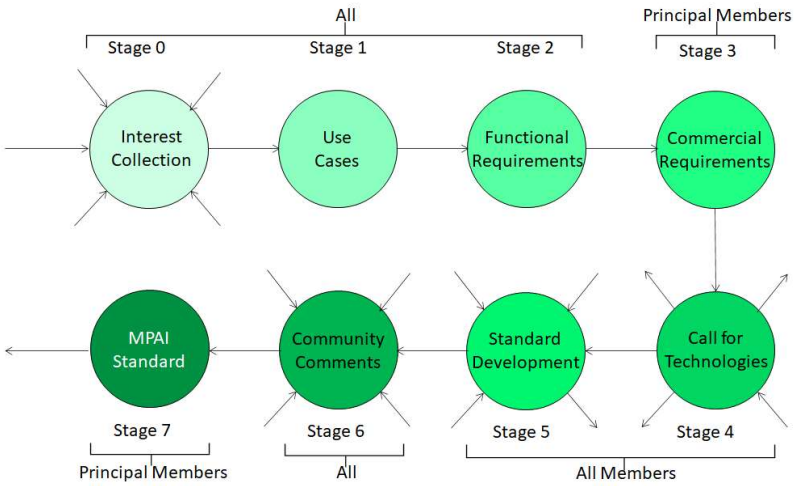

Figure 25 – The MPAI standards development process

Let's start from the top-left, from what is called 0th stage called *Interest Collection*. Members as well as non-members may submit proposals that are collected and harmonised. Some proposals get merged with other similar proposals and some get split because the harmonisation process demands this. The goal is to identify standards proposals reflecting proponents' intentions while assessing its value and breadth of use across different environments. Non-members can fully participate in this process on par with other members. The result of this process is the definition of one or more than one homogeneous area of work called "Use Case". Each Use Case is described in an Application Note. Application Notes can be found on the MPAI web site.

The 1st stage of the process, called *Use Cases*, entails the full use case characterisation and the description of the work program that will produce the Functional Requirements. The 2nd stage, called *Functional Requirements*, is the actual development of the Functional Requirements of the area of work represented by the Use Cases.

The "openness" of the MPAI process in these 3 initial stages is represented by the fact that anybody may participate in Interest Collection, Use Case and Functional Requirements. With an exception, though: when an MPAI member makes a proposal that s/he wishes to be exposed to members only.

The 3rd stage is *Commercial Requirements*. In a sense, a standard is no different than a supply contract where the characteristics of the object to be delivered (Functional Requirements) and the terms of the delivery (Commercial Requirements) are stated.

It should be noted that, from this 3rd stage on, non-members are not allowed to participate (but they can become members at any time), because their role of proposing and describing what a standard should do is over. Antitrust laws do not permit that seller (technology providers) and buyers (users of the standard) sit together and agree on values such as numbers, percentage, or dates, but it permits sellers to indicate the terms, but without values. Therefore, the embodiment of the Commercial Requirements, i.e., the Framework Licence, will refrain from adding such details.

Once both Functional and Commercial Requirements are available, MPAI is able to draft the *Call for Technologies* (4th stage). Anybody is allowed to respond and, if a proposed technology is accepted, the proponent is requested to join MPAI or else having its technology removed from consideration.

The proposals are reviewed and *Standard Development* begins (5th stage).

When the draft standard has reached sufficient maturity, MPAI may decide to make it public with a request for *Community Comments* (6th stage).

After accommodating comments, Principal Members may vote to approve the draft as *MPAI Standard* (7th stage), hence triggering its publication (an Associate Member may become a Principal Member at any time).

This explanation of the way MPAI is organised is important to make clear that MPAI achieves its goals through a proper organisation and following a rigorous process.

17 The governance of the MPAI ecosystem

As previously discussed, AI-based technologies are very powerful and also the object of a high level of concern and scrutiny by regulators over their use in sensitive or mission-critical contexts. A poorly designed AI component can conceal bias or privacy breaches, and any AI module is ultimately only as good as its training set is; in addition, hidden critical vulnerabilities might become evident only when the component is run on unexpected input data (for instance, the case of "typographic attacks").

That is why, during the MPAI design phase, special attention has been put to the concept that the AI-based components provided by MPAI should be trustworthy – anyone willing to use them, including end-users who are not necessarily technical experts in the field, should feel reassured that nothing untoward will happen during operation. The same should happen when AI components offered by MPAI are combined together – and the user should be clearly made aware of what the limits and applicability scope of each component are. In fact, such requirements are arguably the ones having the most far-reaching consequences as far as the overall design of MPAI was concerned. One could say that being able to provide trustworthy AI components has been placed at the centre of MPAI philosophy, structure, and ecosystem.

To cope with this issue, MPAI has introduced the notion of *performance* of an implementation. An implementation "performs" well if the measurement of a set of attributes is above a certain threshold. The performance attributes are described in Section 14.4.

The performance assessment specification of an MPAI standard indicates which of the four attributes an implementation should support and to what extent ("grade").

It should be noted that *performance* is not (technical) *conformance*. Conformance is measured for most technical standards as the adherence to the stated design and inter-operability parameters; however, an AI-based component might well be conformant to its specification (i.e., have the correct number and type of connection channels interfacing with other components as prescribed) and still present design flaws (such as hidden biases or poor training) that prevent it from performing well. So MPAI components shall be tested for conformance and may be assessed for performance. This is a significant difference with respect to the way technical standardising bodies have been operating so far and is a necessary MPAI response to cope with the specific problems and risks presented by AI as a technology. It complicates governance a bit, but the additional guarantees it provides the user with are well worth the effort.

In general, MPAI's commitment to pave the way for an ethical, democratic, and inclusive technology is reflected in the governance of its ecosystem and, more specifically, in the definition of the tests each novel technology being standardised by MPAI is subjected to. The full operation of the MPAI Ecosystems is depicted in Figure 26.

The foundations of the ecosystem governance rely on MPAI as a root of trust. MPAI develops the four components of a new standard as follows:

1. Technical Specification (TS)
2. Reference Software (RS)
3. Conformance Testing (CT)
4. Performance Assessment (PT)

Once the TS is defined, MPAI develops for each MPAI Standard a set of specifications containing the tools, the process and the data that shall be used in the future to estimate the level of performance of an implementation. As explained above, performance is defined as a function of a collection of attributes quantifying reliability, robustness, replicability, and fairness. Independent performance assessors are appointed by MPAI to carry out these procedures for each implementation.

The actors enabling the MPAI ecosystem to operate include the MPAI store, an independent not-for-profit entity established and controlled by MPAI. to end users, guaranteeing availability, reliability, performance, and trust.

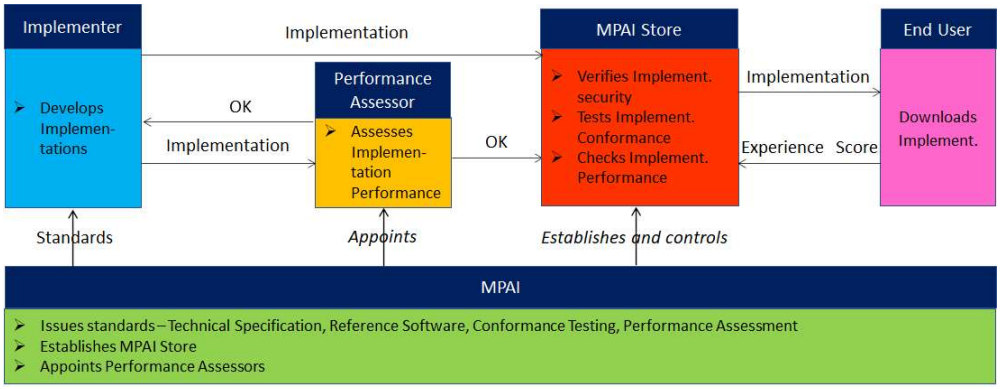

Figure 26 – The governance of the MPAI ecosystem

The MPAI Store occupies a central place in the overall architecture – it acts as a gateway through which implementers make their products available In particular, the MPAI store receives implementations of MPAI standards, tests them for security and conformance, receives the results of the performance tests and stores them. The MPAI store also assigns the AIWs and AIMs provided by each implementer to security experts for testing; it receives AIFs, AIWs and AIMs distribution licenses, and makes implementations available to users.

So, to make the implementation of an MPAI standard available to end users, a number of elements have to come into place:

1. The experts of the appropriate MPAI development committee develop the Technical Specification, provide the Reference Software, and develop the Conformance Testing and Performance Assessment specifications
2. An implementer creates an application based on the MPAI Technical Specification
3. The application is evaluated for security and correct operation by an MPAI-appointed assessor following the Conformance Testing and Performance Assessment specifications
4. The way the application performs is graded according to the assessor's evaluation.

In addition, most MPAI standards will rely on the AI Framework, MPAI-AIF, which, as previously explained, is the overarching MPAI standard specifying how different AI-based AI modules (AIMs) can be connected into AI workflows (AIWs) to generate more complex applications and executed on an MPAI controller (a.k.a. an AIF implementation). To execute an implementation of any MPAI standard, a user will also need an implementation of MPAI-AIF, which in turn, MPAI-AIF being a regular standard, will have been subjected to conformance and performance tests – all for the sake of reliable AI technology!

To summarise, the MPAI ecosystem and its AI-specific design allow end users to access implementations of AIFs, AIWs and AIMs through the MPAI Store that are trustworthy and secure. In addition, by mandating the specification of performance tests and appointing independent performance assessors, MPAI guarantees that each AI component it provides will offer a minimum level of performance, as specified in the standard defining the component. Through the governance and the structure of its ecosystem, MPAI addresses several problems inherently associated with AI applications and provides the implementers and end users with reliable solutions.

18 A renewed life for the patent system

Patents have been the engine of progress over the last few centuries since the time when the Republic of Venice (1474) – the first in Europe – adopted a patent law. The history of media in the 19th century is studded by a multitude of patents filed by the many fascinated by the prospect of conceiving – and patenting – a technology that would enable media's capture, storage, editing, transmission, and presentation. Then, in the first half of the 20th century, television captured a lot of attention, and, in the second half, consumer electronics became the rage.

The patent business model was typically based on an invention by a company, say, PAL by Telefunken or VHS by JVC, that succeeded on the market and was then licenced to competitors, the originating company enjoying a steady flow of royalties. The conversion of the company specification to a national or an international standards body was mostly a bureaucratic process.

The appearance of MPEG overturned that process. The new – digital – family of consumer electronics standards was produced by a committee that assembled technologies from multiple sources, hence embedding multiple patents. Patent pools that had gone out of fashion were revived and the patent pool administrator became the equivalent of the old "company with a successful patented product", the only difference being that the administrator allocated revenues to patent holders. For all practical purposes, however, the business model remained unchanged. Consumer electronics companies still remember the MPEG-1 and MPEG-2 time as the golden age of digital media standards.

MPEG-4 was no longer just a consumer electronics standard, as mobile telcos and IT companies started playing a role. Luckily, the patent-holder grouping model survived to the benefit of efficiency in licensing patented technologies. It was no longer a golden age, but a bronze age (in terms of revenues). Companies, however, knew that there was a standard, that there was a licence, and that there was a patent pool administrator to talk to.

Therefore, patent pooling is still a good solution to give access to a standardised technology, especially in the case where a patent pool is really a one-shop-stop, that is, by taking a license from the pool, the implementer has a licence for all the SEPs (Standard Essential Patent) that cover such technology.

The last and current decade has become the "iron age" of patents with matters drifting away from past practice. There are multiple patent pools for standards, some of which offer an incomplete and incompatible set of licences. Some standards have a single patent pool administrator, but the licence is only published years after the standard is approved. Often, at that moment, there is a new generation of technologies doing a better job than the standard for which the licence terms have finally been published.

Thus, digital media standards are returning to a state comparable with the VHS-Betamax saga. As official standards are becoming practically "inaccessible", different standards with different business models are popping up.

A strong motivation for creating MPAI has been the belief that standard are important; a healthy provisioning of good patents makes good standards; standards must be usable or they stop being enablers of progress in the market and carriers of innovation to consumers; and good patents must be remunerated. Without a fair return of their R&D efforts innovators would stop innovating.

The solution envisaged by MPAI is best illustrated by Figure 27 where the provider-customer relationship in the real world is compared with the relationship in the standards bodies requesting submitters of technical contributions to provide an FRAND (Fair, Reasonable and Non-Discriminatory) patent declaration regarding the licensing of their patents.

In the real world of product business, such as for car sellers, the supplier is committed to delivering goods that meet technical specifications on agreed commercial terms which include price and time of delivery. In the world of intangible matter, such as patent rights, with the FRAND approach, the commercial terms are simply glossed over. A company relying on a FRAND standard knows the functional requirements but not the commercial requirements, much less when the standard will become practically usable.

In the MPAI approach, those responding to a Call for Technologies know both the functional and commercial requirements with a provision that antitrust legislation does not allow MPAI to provide the real-word equivalent of the "price". However, the "customers" of the standard have a reference level ("the total cost is in line with the costs of similar technologies"). MPAI cannot give its "customers" a firm delivery date of the licence of the technology, but users know that the licence shall not come *after* the time products are on the market.

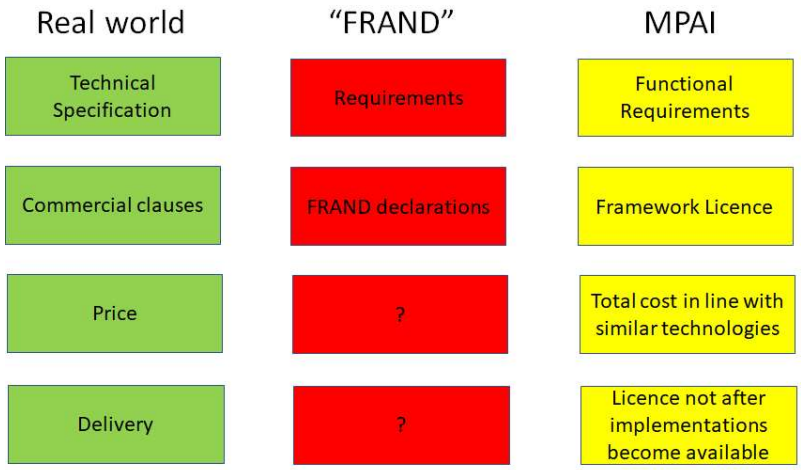

Figure 27 – The provider-customer relationship in the real world and in standardisation

The distinctive difference between the FRAND system and MPAI is the Framework Licence (FWL). This can be described as the IPR holders' business model used to monetise their IP in a standard without values: $, %, dates etc.

Here is the life cycle of MPAI standards:

1. The FWL is developed by Active Members, namely all Principal Members who intend to contribute to the standard. The text of the FWL states, at least, that IPR holders will issue licences
    a. Whose total cost will be in line with the total cost of the licenses for similar data coding technologies and will consider the value on the market of the specific standardised technology.
    b. Not after commercial implementations of the standard become available on the market
2. During the development of the standard, any Member making contributions declares it will make available its licences according to the FWL.
3. After the standard has been approved by the General Assembly, IP holders express their preference on the patent pool administrator with a 2/3 qualified majority and all Members declare they will get a licence for other members' IPRs, if used, within 1 year after publication of IPR holders' licensing terms.

More information on the differences between FRAND and FWL systems can be found in [32].

MPAI believes that patents are the engine of progress and an asset for humankind. The patent system should be saved, in the interest of industry, consumers and innovation.

19 Plans for the future

19.1 AI-enhanced video coding

Video Coding research focuses on radical changes to the classic block-based hybrid coding framework to face the challenges of offering more efficient video compression solutions. AI can play an important role in achieving this goal.

According to a survey of the recent literature on AI-based video coding, performance improvements up to 30% are expected. Therefore, MPAI is investigating whether it is possible to improve the performance of the MPEG-5 Essential Video Coding (EVC) modified by enhancing/replacing existing video coding tools with AI tools keeping complexity increase to an acceptable level.

The AI-Enhanced Video Coding (MPAI-EVC) Evidence Project is extend-ing/enhancing MPEG-5 EVC with the goal of improving its performance by up to 25%. The EVC Baseline Profile has been selected because it is made up with 20+ years old technologies and has a compression performance close to HEVC, and the performance of its Main Profile exceeds that of HEVC by about 36 %. Additionally, some patent holders have announced that they would publish their licence within 2 years after approval of the EVC standard.

Once the MPAI-EVC Evidence Project will demonstrate that AI tools can improve the MPEG-5 EVC efficiency by at least 25%, MPAI will be in a position to initiate work on the MPAI-EVC standard by issuing a Call for Technologies.

Currently, two tools are being considered: Intra Prediction and Super Resolution.

Intra prediction takes advantage of the spatial redundancy within video frames to predict blocks of pixels from their surrounding pixels and thus allowing to transmit the prediction errors instead of the pixel values themselves. Because the prediction errors are of smaller values than the pixels themselves, compression of the video stream to be achieved. Traditional video coding standards leverage intra-frame pixel value dependencies to perform prediction at the encoder end and transfer only residual errors to the decoder. Multiple “Modes” are used, which are various linear combinations of neighbours pixels of the macro blocks being considered. EVC has 5 prediction modes (Figure 28):

- DC: the luminance values of pixels of the current block are predicted by computing average of the luminance values of pixels from upper and left neighbours

- Horizontal prediction: the same mechanism but using only the left neighbours
- Vertical prediction: in this case only the upper neighbours are used
- Diagonal Left and Diagonal Right: linear combination of upper and left neighbours are used

Super Resolution creates a single image with two or more times the linear resolution e.g., the enhanced image will have twice the width and twice the height of the original image, or four times the total pixel count.

For each tool being investigated there are three phases: database building (of blocks of pixels, and subsampled and full-resolution pictures, respectively), learning phase and inference phase.

For the *Intra prediction* track, two training datasets have been built: one of 32x32 and 16x16 intra prediction blocks. A new EVC predictor, leveraging a CNN-based autoencoder is generated.

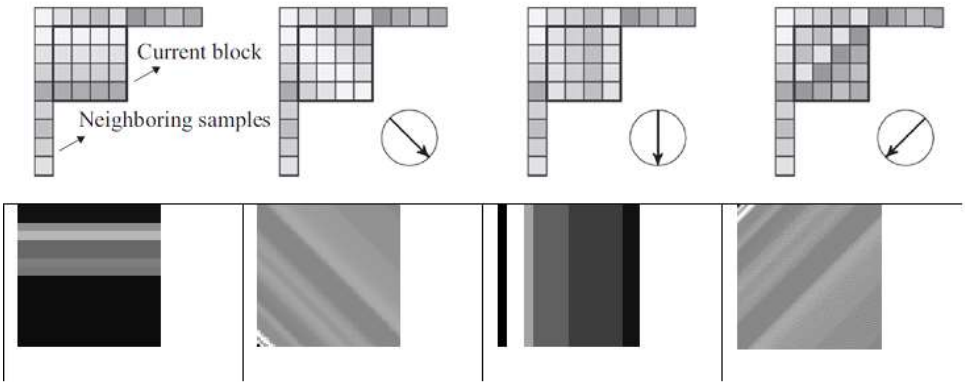

Figure 28 - Upper row: intra prediction examples (Horizontal, Diagonal right, Vertical, Diagonal Left); bottom row: some EVC intra predictors

In the training phase the autoencoder is trained on a dataset by minimizing the Means Square Error (MSE) between its output and the original image block. A communication channel based on a web socket between the EVC code (written in C language) and the autoencoder (written in python) is used to overcome the incompatibilities of different programming frameworks.

In the inference phase the autoencoder is fed with blocks neighbouring the block being predicted. In this way the problem becomes one of reconstructing the missing regions in an image. The encoder sends the 64-by-64 decoded neighbouring the block of each 32-by-32 Coding Unit (CU) and 16-by-16 CU to the autoencoder and returns the new 32-by-32 or 16-by-16 predictor, depending on the case, to the EVC encoder (Figure 29). The EVC

codeword used to signal the EVC mode zero (DC) is replaced to signal the new AI-based prediction mode to the decoder.

The generated bitstream is fully decodable under the assumption that the autoencoder network is also available at the decoder side.

The next steps in this investigation include extending the proposed approach also to 8x8 and 4x4 CUs; experimenting with other network architectures than convolutional; changing the MSE during training, enlarging the context to 128-by-128 and replace all the EVC predictors with the autoencoder-generated predic-tor.

For the *Super Resolution* track a state-of-the-art neural network (Densely Residual Laplacian Super Resolution) was selected because it introduces a new type of architecture based on cascading over residual, which can assist in training deep networks.

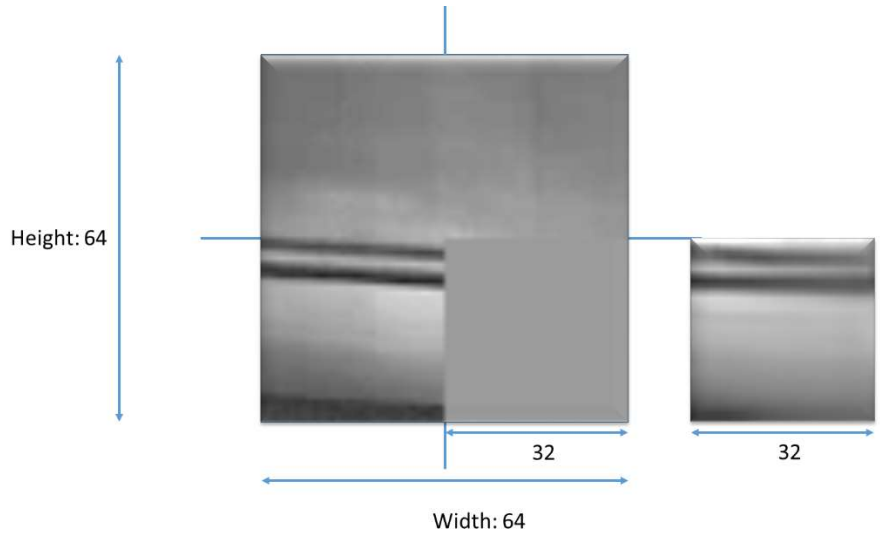

Figure 29 - Left: current block, right: autoencoder generated predictor

A dataset to train the super resolution network has been selected and 3 resolutions (4k, HD, and SD), 4 values of picture quality, two coding tool sets (deblocking enabled, deblocking disabled) for a total of 170 GB dataset.

The super resolution step was added as a post processing tool. The picture before encoding with EVC baseline profile was downscaled and then the super resolution network was applied to the decoded picture to get the native resolution.

Many experiments have been performed to find the right procedure to select a region in the picture (crop), i.e., an objective metric to choose one or more crops inside the input picture in such a way that a trade-off between GPU memory and compression performance is achieved.

The next steps include experiments with other network architectures.

Additional tools considered for further experiments are:

- in-loop filtering: reduce the blockiness effect by filtering out some high frequencies caused by coded blocks.
- motion compensation: use Deep Learning architectures to improve the motion compensation.
- inter prediction: estimate the motion using Deep Learning architectures to refine the quality of inter-predicted blocks; introduce new inter prediction mode to predict a frame avoiding the use of side information.
- quantization: use a neural network-based quantization strategy to improve the uniform scalar quantization used in classical video coding because it does match the characteristics of the human visual system.
- arithmetic encoder: use neural networks to better predict the probability distribution of coding modes.

19.2 End-to-end video coding

There is consensus in the video coding research community – and some papers make claims grounded on results – that so-called End-to-End (E2E) video coding schemes can yield significantly high performance. However, many issues need to be examined, e.g., how such schemes can be adapted to a standard-based codec. End-to-End Video Coding promises AI-based video coding standard with significantly higher performance in the longer term.

As a technical body unconstrained by IP legacy and with the mission to provide efficient and usable data coding standards, MPAI has initiated the *End-to-End Video Coding* (MPAI-EEV) *project*. This decision is an answer to the needs of the many who need not only environments where academic knowledge is promoted but also a body that develops common understanding, models and eventually standards-oriented End-to-End video coding.

Regarding the future research in this field, two major directions are envisioned to ensure that end-to-end video coding is general, robust and applicable. The compression efficiency has the highest priority and enhanced coding methods are expected to be proposed and incorporated into the reference model including high efficiency neural intra coding methods, neural predictive coding tools such as reference frame generation, dynamic inter-prediction structure, and long-term reference frames. The other aspect is the optimization techniques in end-to-end video coding, including neural rate control methods using reinforcement learning, imitation learning, parallel encoding framework, error concealment methods, and interface designation for the down-stream analysis tasks. Moreover, the high-level syntax as well as model-updating mechanism should also be considered in the future.

19.3 Server-based predictive multiplayer gaming

The general issues addressed by Server-based Predictive multiplayer Gaming (MPAI-SPG) are two of the problems affecting online video gaming. The basic idea is to record all behaviours of online games simulated by BOTs or played by humans, to create a large archive with an agent that can support the online game server in case of missing information (network problem) or alert it when a game situation does not resemble a usual game situation (cheating problem).

MPAI-SPG works by comparing the "atomic" unit of an online game, i.e., the game state and the data structure identifying the state of the whole system in a time unit. The agents will be the neural networks that learn to predict, using the data of the games that have taken place up to that moment. As depicted in the right-hand side of Figure 30, the online game server is a generic game engine composed of a unit that creates the game state (Game State Engine) and a set of simpler engines that deal with physics (Physics Engine), the management of inputs and the control of the different game entities (Behaviour Engine) and the game rules and events engine (Rules Engine). This model can be extended by adding other engines as required by the specific game.

MPAI-SPG is implemented by a neural network that interfaces to the game server, acting like a Digital Twin. The server passes the relative information of the current game state and the controller data of the clients that it has received up to that moment to MPAI-SPG. MPAI-SPG's neural networks verify and calculate a predicted game state according to the information that has been received. Usually, the game state information generated by MPAI-SPG will be identical to the game server information. In case of missing packets and information, MPAI-SPG will act to fill the missing information according to all the behaviours recorded in situations like what is happening in the game system up to that moment; it will send its "proposal" to the Game State Engine which will use it to arrive at a decision.

In case of cheating, the game server will process a state that contains anomalies compared to the state predicted by MPAI-SPG. The detection of these anomalies will allow the game server to understand where the malicious information came from, generating a warning on the guilty client. It will then be up to the game state engine to manage this warning.

After defining the architecture and studying the final form that MPAI-SPG will take (a plug-in to be used within game engines), a prototype of Pong with an authoritative server has been developed. This is being used to test and build the first example of a MPAI-SPG. A neural network is being trained to respond to the needs of these scenarios and then define the standard that will enable external producers to develop their own solutions.

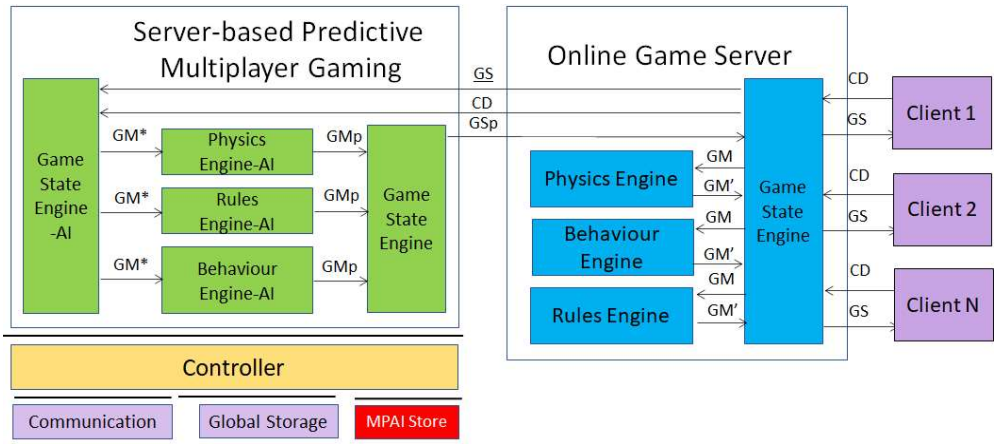

Figure 30 – Server-based Predictive Multiplayer Gaming Reference Model

19.4 Connected autonomous vehicles

Standardisation of Connected Autonomous Vehicles (CAV) components is required because of the different nature of the interacting technologies in a CAV, the sheer size of the future CAV market in the order of T$ p.a. and the need for users and regulators alike to be assured of CAV safety, reliability and explain-ability.

At this point in time, a traditional approach to standardisation might consider CAV standards as premature and some affected industries may not even be ready yet to consider them. CAVs, however, at best belong to an industry still being formed, that is expected to target the production of economic affordable units in the hundreds of millions p.a., with components to be produced by disparate sources. A competitive market of standard components can reduce costs and make CAV confirm their promise to have a major positive impact on environment and society.

Connected Autonomous Vehicles (MPAI-CAV) is an MPAI standard project that seeks to identify and define the CAV components target of standardisation. It is based on a reference model comprising the 5 Subsystems depicted in Figure 31, identifying components and their interfaces and specifying their requirements:

1. Human-CAV interaction (HCI) handles human-CAV interactions.
2. Environment Sensing Subsystem (ESS) acquires information from the Environment via a range of sensors.
3. Autonomous Motion Subsystem (AMS) issues commands to drive the CAV to the intended destination.
4. Motion Actuation Subsystem (MAS) provides environment information and receives/actuates motion commands in the environment.

The Figure depicts the 5 subsystems and their interactions.

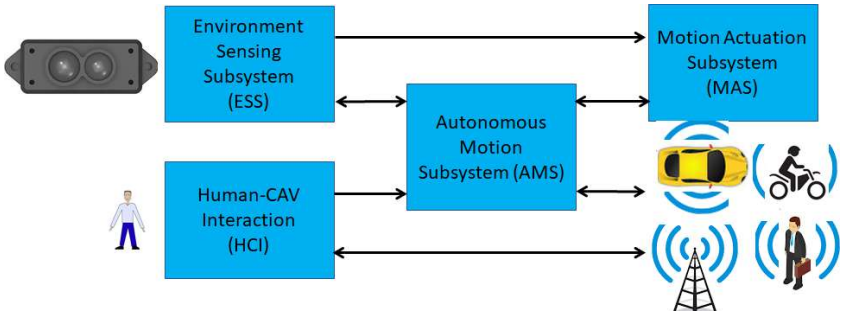

Figure 31 – The CAV subsystems

Human-CAV Interaction operates based on the principle that the CAV is impersonated by an avatar, selected, or produced by the CAV rights-holder. The visible features of the avatar are head face and torso, and the audible feature is speech embedding an emotion like it would be displayed by a human driver. This subsystem's reference model reuses several of the AIMs already developed or being developed by MPAI in addition to a few that are CAV-specific.

The purpose of the Environment Sensing Subsystem, depicted in Figure 32, is to acquire all sorts of electromagnetic, acoustic, mechanical and other data directly from its sensors and other physical data of the Environment (e.g., temperature, pressure, humidity etc.) and of the CAV (Pose, Velocity, Acceleration) from Motion Actuation Subsystem. The main goal is to create the Basic World Representation, the best guess of the environment using the available data. This is achieved by:

1. Acquiring available offline maps of the CAV current Pose:
2. Fusing Visual, Lidar, Radar and Ultrasound data.
3. Updating the Offline maps with static and moving objects, and all traffic signalisations.

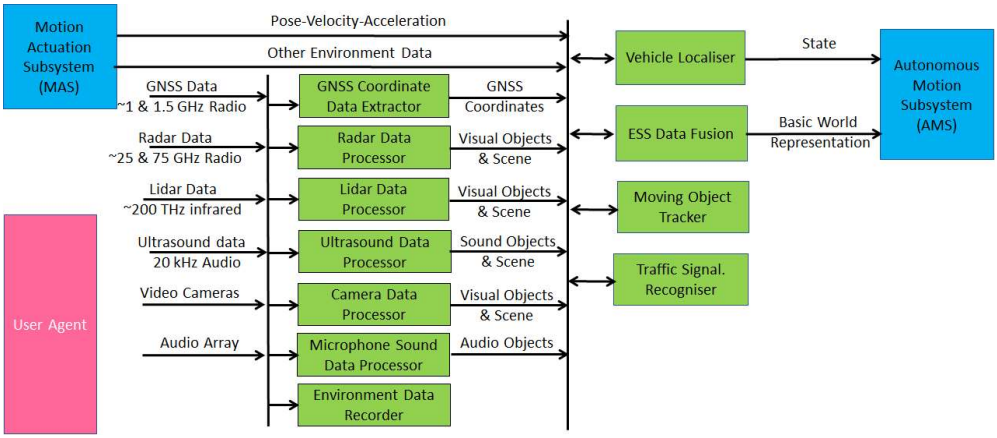

Figure 32 – Environment Sensing Subsystem Reference Model

A CAV exchanges information via radio with other entities, e.g., CAVs in range and other CAV-like communicating devices such as roadside units and Traffic Lights, thereby improving its Environment perception capabilities. Multicast mode is typically used for heavy data types (e.g., Basic World Representation). CAVs in range are important not just as sources of valuable information, but also because, by communicating with them, each CAV can minimise interference with other CAVs while pursuing its own goals.

The Autonomous Motion Subsystem is the core of the CAV operation.

1. Human-CAV Interaction requests Autonomous Motion Subsystem to plan and move the CAV to the human-selected Pose. Dialogue may follow.
2. Computes the Route satisfying the human's request.
3. Receives the current Basic World Representation (BWR) from Environment Sensing Subsystem.
4. Transmits to and receives from other CAVs BWRs and fuses all BWRs to produces the Full World Representation (FWR).
5. Plans a Path connecting Poses.
6. Selects behaviour to reach intermediate Goals acting on information about the Goals other CAVs in range intend to reach.
7. Defines a trajectory, complying with traffic rules, preserving passenger comfort and refining the trajectory to avoid obstacles.
8. Sends Commands to the Motion Actuation Subsystem to take the CAV to the next Goal.

The Figure depicts the architecture of the Motion Actuation Subsystem.

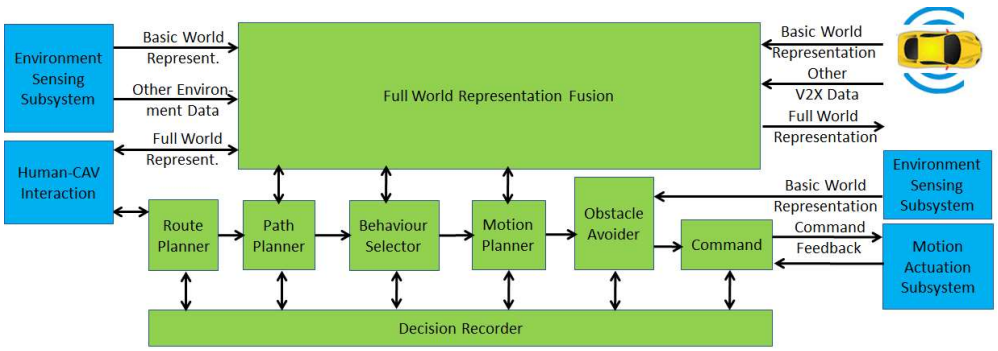

Figure 33 – Autonomous Motion Subsystem Reference Model

19.5 Conversation about a scene

In this standard project, a machine watches a human and the scene around it, hears what the human is saying, and gets the human's emotional state and meaning via audio (speech) and video (face, and gesture).

The human talks to the machine about the objects in the scene indicating one with their hand/arm. The machine understands the object the Human points at by placing the direction of the Human hand/arm from the position occupied by the Human representation of the scene and is capable to make a pertinent response uttered via synthetic speech accompanied by its avatar face.

Figure 34 depicts the solution being investigated, its AIMs, the connections and the data exchanged. It is an extension of the Multimodal Question Answering use case where the machine creates an internal scene representation by using the *Video Analysis3* AIM, spatially locates and recognises the objects in the scene, and combines gesture and scene descriptors to understand which object the Human is looking or pointing at.

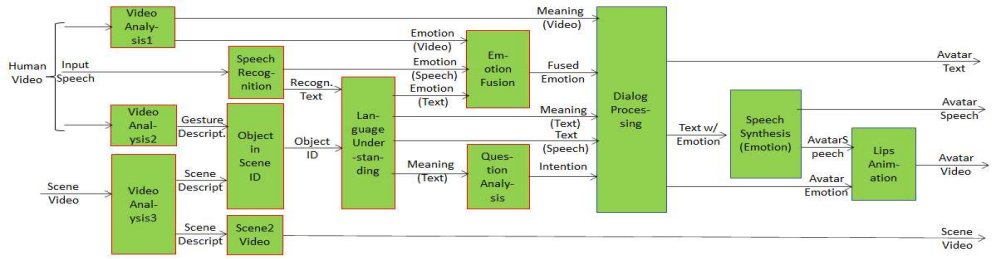

Figure 34 – Conversation about a scene

19.6 Mixed-reality collaborative spaces

Mixed-reality Collaborative Spaces (MPAI-MCS) is an MPAI standard project containing use cases, and functional requirements for AI Workflows, AI Modules, and Data Formats applicable to scenarios where geographically separated humans collaborate in real time by means of speaking avatars in virtual reality spaces called ambients to achieve goals generally defined by the use case and specifically carried out by humans and avatars.

Examples of applications target of the MPAI-MCS standard are:

- *Local Avatar Videoconference* where realistic virtual twins (avatars) of physical twins (humans) sit around a table holding a conference. In this case, as many functions as practically possible should be clients because of security issues, e.g., participant identity and no clear text information sent to the server.
- *Virtual eLearning* where there may be less concerns about identity and information transmitted to the server.

MCSs can be embodied in a variety of configurations. In two extreme configurations each MCS participant:

- Creates the MCS in its own client using information generated by the client and received from other clients without necessarily relying on a server.
- Generates media information and commands, and consumes information packaged by a server where the MCS is created, populated, and managed based on information received from clients.

In between these two extreme configurations, there is a variety of combinations where different splitting of functions between clients and servers are possible.

The Figure illustrates case #1 where most of the intelligence is concentrated in the transmission part of the client. We note *Speech Separation*, extracting speech from the participant's location, participant identification via speech and face, extraction of participant's emotion and meaning from speech, face and gesture and creation of avatar description.

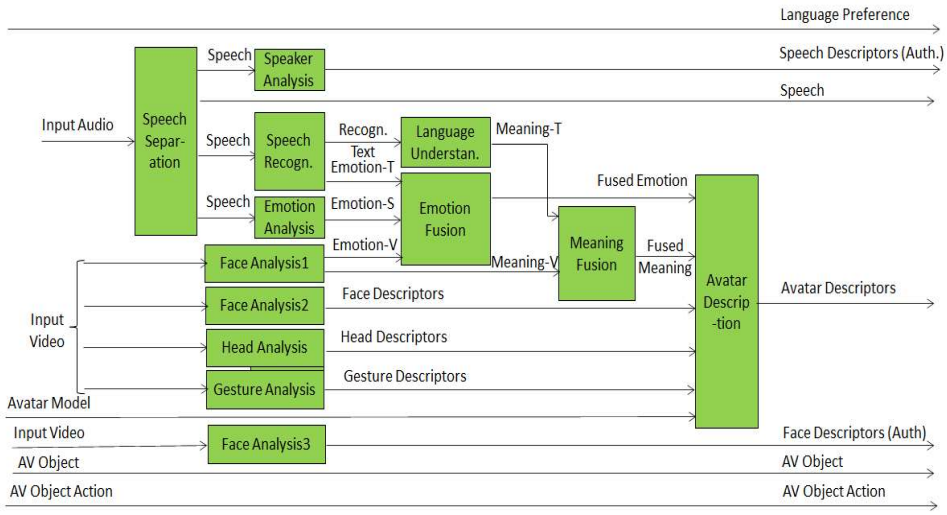

Figure 35 – Client-Based Ambient TX Client Reference Model

It is easy to see that a few AIMs can be reused from other MPAI standards. The server – not represented – contains a *Translation AIM*, again a derived from the MPAI-MMC standard and the participant identification.

19.7 Audio on the go

The *Audio-On-the-Go* (AOG) use case addresses the situation where musical content is consumed in a variety of different contexts: while biking in the traffic, at home with a dedicated stereo setup or in a car while driving.

The use case allows the reproduction of material to react dynamically to environmental conditions and perform a Dynamic Equalization based, at least initially, on a small set of environmental variables (such as time of the

day or Ambient Noise) and adjust the playback to the specific hearing profile of the listener.

As an example, while biking in the traffic, the system receives as input the audio stream of the music playing app, performs an active equalization to compress the dynamic range of the stream (to allow user to experience an optimal sound despite the surrounding noise) then equalizes the stream again adapting to the user hearing profile. Eventually, it delivers the resulting sound to the headphones via Bluetooth.

Another example encompasses a home-listening situation, where a user listens a high-quality audio stream, the system performs an active equalization, accounting for the specific user hearing-profile, tailoring the listening experience to the specific listening abilities of the user.

All the above is obtained by chaining several AIMs: (1) the sound data stream is piped into a "*Digital Audio Ingestion*" AIM, which captures and normalizes the playback input signal and can be via standard signal processing tecniques; (2) the "*Dynamic Signal Equalization*" AIM uses an ML algorithm based on RNN that operates dynamically on a given number of bands to adapt dynamically the input signal to optimize the user audio experience and performs noise reduction and environment adaptation; (3) the "*Delivery*" AIM delivers the resulting stream to available and selected endpoints in the same network, automatically selecting the most suitable protocol.

20 Conclusions

This is a small book talking about a big adventure: standards for the most sensitive objects of all – data – using the most prominent technology of all – AI – for pervasive and trustworthy use by billions of people. At the end of this book, it is thus appropriate to assess what the authors think will be the likely impact of MPAI on industry and society.

The first impact will be the availability of standards for a technology that is best used to transform the data but has not seen any so far. Standards that are driven by the same principles that guided another great adventure – MPEG – that replaced standards meant to be exclusively used by certain countries or industries with standards serving humankind.

The second impact will be a right that used to be taken for granted by implementers but has ceased to be a right some time ago. An implementer wishing to use a published standard should be allowed to do so, of course after remunerating those who invested money and talent to produce the technology enabling the standard.

The third impact is a direct consequence of the preceding two. In the mid-18th century, trade did not develop as it could because feudal traditions allowed petty lords to erect barriers to trade for the sake of a few livres or

pounds. Today we do not live in a feudal age, but we still see petty lords here and there obstructing progress for the sake of a few dollars.

The fourth impact is the mirror of the third. An industry freed from shackles, with access to global AI-based data coding standards and operating in an open competitive market will be able to churn out interoperable AI-based products, services, and applications in response to consumer needs which are known today and the many more which are not yet known.

The fifth impact is a direct consequence of the fourth. An industry using sophisticated technologies such as AI and forced to be maximally competitive will have a need to foster an accelerated progress of those technologies. We can confidently look forward to a new spring of research and advancement of science in a field to which today it is difficult to place boundaries.

The sixth impact will be caused by MPAI's practical Performance Assessor based solution to the concerns of many: AI technologies are as potentially harmful to humankind as they are powerful. The ability of AI technologies to hold vast knowledge without simple means for users to check how representative of the world they are – when they are used to handle information and possibly make decisions – opens our minds to apocalyptic scenarios.

The seventh impact is speculative, but no less important. The idea of intelligent machines able to deal with humans has always attracted the intellectual interest of writers. Machines dealing with humans are no longer speculations but facts. As objects embedding AI – physical and virtual – increase their ramification into our lives, more issues than "Performance" will come to the surface and will have to be addressed. MPAI, with its holistic view of AI as the technology enabling a universal data representation, proposes itself as the body where such issues as enabled by progress of technology can be addressed and ways forward found.

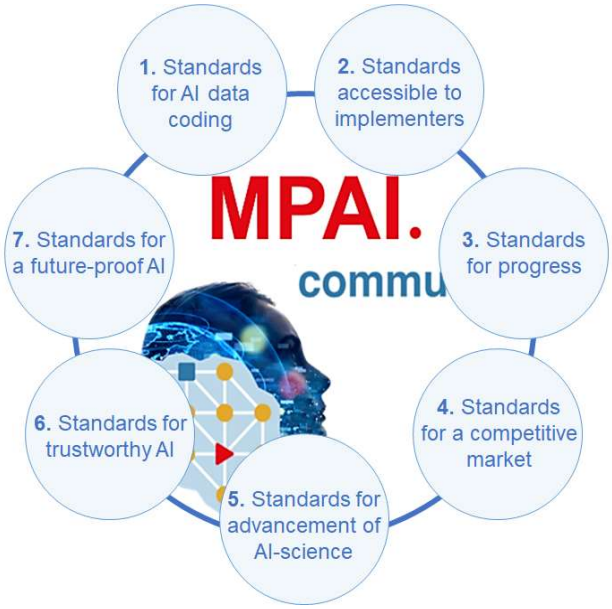

Figure 36 – The expected MPAI impacts

The results achieved by MPAI in 15 months of activity and the plans laid down for the future demonstrate that the seven impacts identified above are not just wishful thinking. MPAI invites people of good will to join and make the potential real.

21 References

[1] J. H. Wilson, P. R. Daugherty and N. Morini-Bianzino, "The Jobs That Artificial Intelligence Will Create," *MITSloan Management Review,* 2017.

[2] M. A. Hanif, F. Khalid, R. Putra, M. T. Teimoori, F. Kriebel, J. Zhang, S. Rehman, T. Theocharides, A. Artusi, S. Garg and M. Shafique, "Robust Computing for Machine Learning-Based Systems," in *Dependable Embedded Systems*, Springer International Publishing, 2021, pp. 479-503.

[3] J. J. Zhang, K. Liu, F. Khalid, M. A. Hanif, S. Rehman, T. Theocharides, A. Artusi, M. Shafique and S. Garg, "Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities," in *Proceedings of the 56th Annual Design Automation Conference*, Las Vegas, NV, USA, 2019.

[4] D. Beniaguev, I. Segev and M. London, "Single cortical neurons as deep artificial neural networks," *Neuron,* vol. 109, no. 17, 2021.

[5] P. Lewis, P. Stenetorp and S. Riedel, “Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017.

[7] I. Jang, P. Kudumakis, M. Sandler and K. Kang, “The MPEG Interactive Music Application Format Standard [Standards in a Nutshell],” *IEEE Signal Processing Magazine,* vol. 28, no. 1, pp. 150-154, 2011.

[8] ISO/IEC, "23008-3:2015, Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio," 2015.

[9] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs and M. Dietz, “ISO/IEC MPEG-2 Advanced Audio Coding,” *Journal of the Audio Engineering Society,* pp. 789-814, 1997.

[10] M. Noisternig, T. Musil, A. Sontacchi and R. Holdrich, “3D binaural sound reproduction using a virtual ambisonic approach,” in *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, VECIMS '03*, 2003.

[11] T. Chen, H. Liu, Q. Shen, T. Yue, X. Cao and Z. Ma, “DeepCoder: A deep neural network based video compression,” in *IEEE Visual Communications and Image Processing (VCIP '17)*, 2017.

[12] W. H. Beaver, “Financial Ratios As Predictors of Failure,” *Journal of Accounting Research,* vol. 4, pp. 71--111, 1966.

[13] E. I. Altman, “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy,” *The Journal of Finance,* vol. 23, no. 4, pp. 589--609, 1968.

[14] J. A. Ohlson, “Financial Ratios and the Probabilistic Prediction of Bankruptcy,” *Journal of Accounting Research,* vol. 18, no. 1, pp. 109--131, 1980.

[15] M. E. Zmijewski, “Methodological Issues Related to the Estimation of Financial Distress Prediction Models,” *Journal of Accounting Research,* vol. 22, pp. 59-82, 1984.

[16] E. I. Altman, M. Iwanicz-Drozdowska, E. Laitinen and A. Suvas, “Financial and Nonfinancial Variables As Long-Horizon Predictors of Bankruptcy,” *Journal of Credit Risk,* vol. 12, no. 4, 2016.

[17] E. I. Altman, “Predicting financial distress of companies: revisiting the Z-Score and ZETA® models,” in *Handbook of Research Methods and Applications in Empirical Finance*, ElgarOnline, 2013, pp. 428--456.

[18] S. A. Hillegeist, E. K. Keating, D. P. Cram and K. G. Lundstedt, “Assessing the Probability of Bankruptcy,” *Review of Accounting Studies,* vol. 9, pp. 5-34, 2004.

[19] A. Upneja and M. C. Dalbor , “An examination of capital structure in the restaurant industry,” *International Journal of Contemporary Hospitality Management,* vol. 13, no. 2, pp. 54-59, 2001.

[20] J. Chen, L. Chollete and R. Ray, “Financial distress and idiosyncratic volatility: An empirical investigation,” *Journal of Financial Markets,* vol. 13, no. 2, pp. 249-267, 2010.

[21] H. Son, C. Hyun, D. Phan and H. J. Hwang, “Data analytic approach for bankruptcy prediction,” *Expert Systems with Applications,* vol. 138, 2019.

[22] European Commission, "OArtificial Intelligence - A European approach to excellence and trust," 2020. [Online]. Available: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. [Accessed 17 December 2021].

[23] D. J. Trump, “Executive order on maintaining American leadership in artificial intelligence,” *Federal Register: White House,* vol. 84, no. 31, pp. 3967-3972, 2019.

[24] L. Parker, "The American AI Initiative: The U.S. strategy for leadership in artificial intelligence," 11 June 2020. [Online]. Available: https://oecd.ai/en/wonk/the-american-ai-initiative-the-u-s-strategy-for-leadership-in-rtificial-intelligence. [Accessed 17 December 2021].

[25] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang and L. Floridi, “The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation,” *AI & SOCIETY ,* vol. 36, pp. 59-77, 2021.

[26] L. Chiariglione, A. Basso, P. Ribeca, M. Bosi, N. Pretto, C. G., M. Guarise, M. Choi, F. Yassa, R. Iacoviello, A. Artusi, F. Banterle, G. Saccardi, A. Fiandrotti, G. Ballocca, M. Mazzaglia, M. Rosano and S. Moskowitz, “AI-Based media coding and beyond,” in *International Broadcasting Conference (IBC)*, 2021.

[27] M. Tsinberg, M. Bosi, A. Luthra and R. Iacoviello, "Basic Applications, Technologies and Benefits for Video Coding by means of Artificial Intelligence," in *HPA Tech Retreat*, 2021.

[28] M. Bosi, N. Pretto, M. Guarise and S. Canazza, "Sound and Music Computing using AI: Designing a Standard," in *18th Sound and Music Computing Conference (SMC '21)*, Virtual, 2021.

[29] N. Pretto, C. Fantozzi, E. Micheloni, V. Burini and S. Canazza, “Computing Methodologies Supporting the Preservation of Electroacoustic Music from Analog Magnetic Tape,” *Computer Music Journal ,* vol. 42, no. 4, pp. 59-74, 2019.

[30] M. B. Çöteli and H. Hacıhabiboğlu, “Sparse Representations With Legendre Kernels for DOA Estimation and Acoustic Source Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 2296-2309, 2021.

[31] G. Perboli, A. Tronzano, M. Rosano, L. Tarantino and F. Velardocchia, “Using machine learning to assess public policies: a real case study for supporting SMEs development in Italy,” in *IEEE Technology Engineering Management Conference - Europe (TEMSCON-EUR)*, 2021.

[32] R. Dini, "FRAND forever? Or are there other business models possible?," 2021. [Online]. Available: https://mpai.community/2021/01/05/frand-forever-or-are-there-other-business-models-possible/. [Accessed 17 December 2021].

[33] ISO/IEC , “23008-3:2015 Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio,” 2015.

Annex 1
List of acronyms

Acronym	Definition
6DoF	Six Degree of Freedom
AI	Artificial Intelligence
AIF	AI Framework
AIM	AI Module
AIW	AI Workflow
AMS	Autonomous Motion Subsystem
ANN	Artificial Neural Network
API	Application Programming Interface
AR	Augmented Reality
ARP	Audio Recording Preservation
ASR	Automatic Speech Recognition
ATSC	Advanced Television Standard Committee
AVC	Advanced Video Coding
AVS	Audio Video Coding Standard
BART	Bidirectional and Auto-Regressive Transformer
BE	Behaviour Engine
BERT	Bidirectional Encoder Representations from Transformers
BWR	Basic World Representation
CAE	Context-based Audio Enhancement
CAV	Connected Autonomous Vehicle
CDVA	Compact Descriptors for Video Analysis
CDVS	Compact Descriptors for Visual Search
CNN	Convolutional Neural Network
CPP	Company Performance Prediction
CPU	Central Processing Unit
CT	Conformance Testing
CU	Coding Unit
CUI	Compression and Understanding of Industrial Data
CWE	Conversation With Emotion
dB	decibel
DB	Data Base
DC	Development Committee
DDSP	Differentiable Digital Signal Processing
DNA	DeoxyriboNucleic Acid
DNN	Deep Neural Networks

DP	Data Processing
DSP	Digital Signal Processing
DVB	Digital Video Broadcasting
E2E	End-to-End
EAE	Enhanced Audioconference Experience
EES	Emotion-Enhanced Speech
EEV	End-to-End Video Coding
ESS	Environment Sensing Subsystem
EVC	Essential Video Coding
FFT	Fast Fourier Transform
FFNN	Feed-Forward Neural Network
FRAND	Fair, Reasonable and Non-Discriminatory
FWL	Framework Licence
FWR	Full World Representation
GA	General Assembly
GAN	Generative Adversarial Networks
GNSS	Global Navigation Satellite System
GPT	Generative Pre-trained Transformer
GSE	Game State Engine
HCI	Motion Actuation Subsystem
HD	High Definition
HEVC	High-Efficiency Video Coding
HMM	Hidden Markov Models
HOA	Higher-Order Ambisonics
HPC	High Performance Computers
HRTF	Head-Related Transfer Function
Hz	Hertz
IP	Internet Protocol
IPR	Intellectual Property Right
ISDB	Integrated Services Digital Broadcasting
IT	Information Technology
LAT	Lexical Answer Type
LSTM	Long Short-Term Memory
MAS	Motion Actuation Subsystem
MCS	Mixed-reality Collaborative Spaces
MCU	MicroController Unit
ML	Machine Learning
MMC	MultiModal Conversation
MQA	Multimodal Question Answering
MSE	Mean Square Error

NE	Named Entity
NLP	Natural Language Processing
NN	Neural Network
OBA	Object-Based Audio
OTT	Over-The-Top
PCC	Point Cloud Compression
PA	Performance Assessment
PE	Physics Engine
POS	Part Of Speech
PSR	Perceptual Sound field Reconstruction
PCA	Principal Component Analysis
RAM	Random Access Memory
RE	Rukes Engine
RNN	Recurrent Neural Network
RS	Reference Software
QA	Question Answering
SAT	Semantic Answer Type
SD	Standard Definition (SD)
SDO	Standards Developing Organisation
SEP	Standard Essential Patent
SFT	Spherical Fourier Transform
SPG	Server-based Predictive multiplayer Gaming
SRL	Semantic Role Labelling
SRS	Speech Restoration System
TS	Technical Specification
TTS	Text-To-Speech
UHD	Ultra High Definition
UHF	Ultra High Frequency
UST	Unidirectional Speech Translation
VR	Virtual Reality
VVC	Versatile Video Coding