VOLUME 11
ARTIFICIAL INTELLIGENCE, ROBOTICS & DATA SCIENCE

Topic Coordinators
Sara Degli Esposti & Carles Sierra

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030
Challenges coordinated by:
Jesús Marco de Lucas & M. Victoria Moreno-Arribas

VOLUME 11

ARTIFICIAL INTELLIGENCE, ROBOTICS & DATA SCIENCE

Catálogo de publicaciones de la Administración General del Estado:
https://cpage.mpr.gob.es

Editorial CSIC:
http://editorial.csic.es (correo: *publ@csic.es*)

ISBN Vol. 11: 978-84-00-10758-1
ISBN O.C.: 978-84-00-10736-9
e-ISBN Vol. 11: 978-84-00-10759-8
e-ISBN O.C.: 978-84-00-10734-5
NIPO: 833-21-091-1
e-NIPO: 833-21-092-7
DL O.C: M-2426-2021

Diseño y maquetación: gráfica futura

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030

VOLUME 11
ARTIFICIAL INTELLIGENCE, ROBOTICS & DATA SCIENCE

Topic Coordinators
Sara Degli Esposti & Carles Sierra

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030

What are the major scientific challenges of the first half of the 21st century? Can we establish the priorities for the future? How should the scientific community tackle them?

This book presents the reflections of the Spanish National Research Council (CSIC) on 14 strategic themes established on the basis of their scientific impact and social importance.

Fundamental questions are addressed, including the origin of life, the exploration of the universe, artificial intelligence, the development of clean, safe and efficient energy or the understanding of brain function. The document identifies complex challenges in areas such as health and social sciences and the selected strategic themes cover both basic issues and potential applications of knowledge. Nearly 1,100 researchers from more than 100 CSIC centres and other institutions (public research organisations, universities, etc.) have participated in this analysis. All agree on the need for a multidisciplinary approach and the promotion of collaborative research to enable the implementation of ambitious projects focused on specific topics.

These 14 "White Papers", designed to serve as a frame of reference for the development of the institution's scientific strategy, will provide an insight into the research currently being accomplished at the CSIC, and at the same time, build a global vision of what will be the key scientific challenges over the next decade.

VOLUMES THAT MAKE UP THE WORK

1 *New Foundations for a Sustainable Global Society*
2 *Origins, (Co)Evolution, Diversity and Synthesis of Life*
3 *Genome & Epigenetics*
4 *Challenges in Biomedicine and Health*
5 *Brain, Mind & Behaviour*
6 *Sustainable Primary Production*
7 *Global Change Impacts*
8 *Clean, Safe and Efficient Energy*
9 *Understanding the Basic Components of the Universe, its Structure and Evolution*
10 *Digital and Complex Information*
11 *Artificial Intelligence, Robotics and Data Science*
12 *Our Future? Space, Colonization and Exploration*
13 *Ocean Science Challenges for 2030*
14 *Dynamic Earth: Probing the Past, Preparing for the Future*

CSIC scientific challenges: towards 2030
Challenges coordinated by:
Jesús Marco de Lucas & M. Victoria Moreno-Arribas

Volume 11
Artificial Intelligence, Robotics and Data Science

Participating Researchers and Centres

Topic Coordinators
Carles Sierra (IIIA-CSIC) and Sara Degli Esposti (IPP-CSIC)

Challenges Coordinators
Felip Manyà (IIIA-CSIC), Adrià Colomé (IRI-CSIC); Nardine Osman (IIIA-CSIC), Daniel López (IFS-CSIC); José Javier Ramasco Sukia (IFISC-CSIC), Lara Lloret Iglesias (IFCA-CSIC); Guillem Alenyà (IRI-CSIC), Jorge Villagra (CAR-CSIC); M. Dolores del Castillo (CAR-CSIC), Marco Schorlemmer (IIIA-CSIC); Pablo Noriega (IIIA-CSIC), Txetxu Ausín (IFS-CSIC); Teresa Serrano (IMSE-CNM-CSIC), Arantza Oyanguren (IFIC-CSIC); David Arroyo Guardeño (ITEFI-CSIC), Piedad Brox Jiménez (IMSE-CNM-CSIC).

Participating Researchers
Centro de Automática y Robótica (CAR, CSIC)
Centro de Investigaciones Biológicas (CIB, CSIC)
Instituto de Análisis Económico (IAE, CSIC)
Instituto Cajal (IC, CSIC)
Instituto de Ciencias del Espacio (ICE, CSIC)
Instituto de Ciencias Matemáticas (ICMAT, CSIC-UAM-UC3M-UCM)
Geociencias Barcelona (GEO3BCN, CSIC)
Instituto de Ciencias de la Vid y del Vino (ICVV, CSIC)
Instituto de Economía, Geografía y Demografía (IEGD, CSIC)
Instituto de Física de Cantabria (IFCA, CSIC– UC)
Instituto de Física Corpuscular (IFIC, CSIC – UV)
Instituto de Física Interdisciplinar y Sistemas (IFISC, CSIC – UIB)
Instituto de Filosofía (IFS, CSIC)
Instituto de la Grasa (IG, CSIC)
Instituto de Investigación en Inteligencia Artificial (IIIA, CSIC)
Instituto de Investigaciones Marinas (IIM, CSIC)
Instituto de Lengua, Literatura y Antropología (ILLA, CSIC)
Instituto de Microelectrónica de Barcelona (IMB-CNM, CSIC)
Institución Milá y Fontanals (IMF, CSIC)
Instituto de Microelectrónica de Sevilla (IMSE-CNM, CSIC – US)
Instituto de Neurociencias (IN, CSIC – UMH)
Instituto de Gestión de la Innovación y del Conocimiento (INGENIO, CSIC-UPV)
Instituto de Políticas y Bienes Públicos (IPP, CSIC)
Instituto de Química Orgánica General (IQOG, CSIC)
Instituto de Robótica e Informática Industrial (IRII, CSIC)
Instituto de Tecnologías Físicas y de la Información Leonardo Torres Quevedo (ITEFI, CSIC)
Avantopy

VOLUME 11 SUMMARY

VOLUME 11 EXECUTIVE SUMMARY

ARTIFICIAL INTELLIGENCE, ROBOTICS AND DATA SCIENCE

Topic Coordinators
Sara Degli Esposti (IPP-CCHS, CSIC)
Carles Sierra (IIIA, CSIC)

INTRODUCTION

The world we live in is increasingly interconnected and features smooth interactions between human beings and all sorts of systems and devices which are showing increasing levels of autonomy and intelligence. Thus, Artificial Intelligence (AI), robotics and data science are already part of people's everyday life and are changing people's working structures, relationships, and learning habits. We understand AI as the ability of a computer or robot to perform tasks usually associated with intelligent beings. In the vision expressed in this book, we include classical and modern approaches to AI, the technologies that come from them and, in general, all kinds of artificially intelligent entities and systems.

Living a life which is mediated by smart technologies poses new and unexpected social and ethical challenges. For instance, the controversy around the ethical implications of the concept of autonomy in AI has led many experts to demand the banning of autonomous lethal weapons (aka killer robots). This banning has received the strong opposition of major players such as China and the US. The United Nations plays an active role in this debate, while the EU amongst other players is figuring out how to position itself within this heated controversy; in the meanwhile, the scientific community is leaning towards a human centred development of AI, a shared position amongst recently funded Research Networks on AI. Ensuring the socially acceptable and ethically aligned development of AI is the central priority of the next decade.

National and regional governments, as well as public and private entities, consider AI an area of high strategic importance and R&D priority in the next decade. AI generates a global phenomenon reshaping how organisations and individuals interact. This consideration invites us to reflect on how the collaboration between humans and machines should enfold to ensure the flourishing of humankind.

The societal and economic relevance of AI has already triggered a number of policy initiatives around the world. Overall there is general agreement on aspects related to AI that all governments need to address. Most national governments such as the UK, France, Spain, New Zealand, or Australia, realise the strategic importance of AI for industry R&D and innovation. They all propose strategies that are rather similar: raising awareness and addressing social and ethical risks; acknowledging the essential role of AI in the public sector and in the provision of public services; and responding to the growing need to promote further research, training and education in this domain.

In 2020 the European Commission put forward a European approach to AI envisioning investments for EUR 20 billion per year over the next decade. The *White paper on Artificial Intelligence — A European approach to excellence and trust,* published in February 2020, presents policy options for creating an ecosystem of excellence and trust, namely to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens. Similarly, the *Spanish Strategy for AI* identifies six priorities: creating organisational structures for AI development and assessment; identifying AI application areas with high societal and economic impact; facilitating knowledge transfer from scientific research to industry; boosting AI education and training through multi-disciplinary programmes; building big data infrastructures and a digital data ecosystem; assessing the ethics of AI.

Values, rights and ethics are of such paramount importance that in 2019, the EU set up a group of experts to discuss the ethical implications of the technology: the high level expert group on AI, that provided a large number of recommendations, some of them aligned with the already mentioned Beijing declaration on AI and Education. After the aforementioned "White Paper on Artificial Intelligence" was released, different politicians announced multi-billion investments. Things seem to move fast, but there are doubts on whether Europe is doing enough to support AI R&I.

In the United States, after an apparent lack of activity during 2017 and 2018, AI received in 2019 a new impulse from the US government. Resources were allocated to create six National AI Research Institutes to work on trustworthy AI, foundations of machine learning, and AI for accelerating molecular manufacturing, discovery in physics, AI-augmented learning and innovation in the food system. The 2019 *US National Artificial Intelligence Research and Development Strategic Plan* identifies the following key priorities: the development of effective methods for human-AI collaboration; better understanding and addressing the ethical, legal, and societal implications of AI; ensuring the safety and security of AI systems.

China, another extremely dynamic country in AI, recently announced plans to become the leading country in AI. In particular, with the Science and Technology Innovation 2030 initiative of the Chinese government which strongly focuses on AI technologies, as well as the New Generation Artificial Intelligence Development Plan, whose goals are to seize the major strategic opportunities for the development of AI, to create Chinese first mover advantage, and to accelerate the transformation of China into an innovative country and a leading power of science and technology in the world. We would like to stress the important focus of these Chinese initiatives in Education.

The role that AI can play in education is a fundamental aspect of the impact of AI on our society. It is a major element in the Chinese strategy as mentioned before, and an essential part of some countries' AI strategies, like in the case of Finland. The way in which we educate children and citizens in AI is key for the acceptance of the technology and to prepare citizens to embrace and shape the change brought by smart technologies from a critical and constructive standpoint.

It is important to stress the relevance of the May 2019 Beijing *Declaration on AI and Education*, which was signed by more than 100 States. A consensus was reached on the urgency for Governments to prepare the workforce for the changes in training and skills required by the widespread adoption of AI systems. The reason for this call for better training and education around AI is to ensure AI technologies are used to empower teachers rather than replace them, by developing the appropriate capacity-building programmes for teachers to work alongside AI systems. The declaration wishes also to promote the equitable and inclusive use of AI irrespective of people's disability, social or economic status, ethnic or cultural background, or geographical location. It emphatizes gender equality, and aims to guarantee the ethical, transparent

and auditable use of educational data, among several other recommendations. The role of research-informed education in the preparation of our society for the integration of AI in our lives is nowadays a cornerstone of public policies, regulatory development and R&D funding strategies.

The current Covid-19 pandemia has further demonstrated that the deployment of AI technologies play a fundamental role in tackling today's global challenges; it has also expanded international collaborations and accelerated the pace of R&I on AI. An example is represented by the collaboration between the Chinese company Baidu and the Oregon State University working on predicting the secondary structure of the RNA sequence of Covid-19. Several other players, such as Deepmind or US digital giants, have been offering their expertise in AI and data analytics and their computing power.

WHY A WHITE PAPER ON AI, DATA SCIENCE AND ROBOTICS?

Within this context, the Spanish National Research Council (CSIC) identified the need of assessing its internal capabilities and fostering the collaboration between different institutes and research groups working in the fields of AI, robotics and data science in order to identify open questions and challenging research priorities, and to sketch a strategic roadmap for the next decade.

The present white paper is the result of the collaborative effort between several research groups, which work on different aspects of AI, data science and robotics within CSIC. The works that led to this white paper started with a meeting convened at the Artificial Intelligence Research Institute (IIIA), part of CSIC, on 13 and 14 June 2019. During this preliminary meeting members of different research groups working on AI discussed the main challenges of the field and decided to create an informal communication channel named AI-HUB.CSIC for future co-ordination. This informal channel proved instrumental at the end of June to define the map of competences of CSIC on AI, as requested by the Ministry of Science to draw the Spanish AI Map, which complemented the Spanish Strategy on AI released in March 2019.

In July 2019, CSIC's scientific Vice-presidency announced the identification of 12 strategic thematic areas to be developed during the next decade (2021-30). One of them was Artificial Intelligence, Robotics and Data Science. In order to better identify research groups and individual researchers working in this area, all institutes of CSIC were invited to identify groups working on AI

foundations and applications across disciplinary domains. In total, 32 groups responded to the call for interest in the AI thematic area. This internal call offered an interesting overview of research activities in this domain. Group members discussed their research with their peers in a plenary meeting held at the IIIA on 28 and 29 November 2019. Researchers created working groups around sub-thematic areas which later became the chapters of this white paper. Two researchers were elected by members of each group to become the coordinator and adjunct coordinator of working group. They assumed the responsibility of coordinating communications within the group and with the coordinators of the thematic area. From December 2019 until May 2020, each group worked on developing the content of one chapter of this white paper. The list of authors of each chapter appears at the beginning of the chapter. At the end of the book, the complete list of CSIC researchers who have contributed to advance the field of AI, robotics and data science is available. Some researchers have contributed to the foundations of the area, others to application in a rich variety of domains.

STRUCTURE OF THE WHITE PAPER

The white paper offers an overview of research carried out in groups and institutes located across Spain. More importantly, the paper sketches a preliminary roadmap for addressing current R&D challenges associated to AI.

The book is structured in eight chapters. Each one represents a sub-area studied within the CSIC's research community. All chapters follow the same structure: an executive summary, a comprehensive introduction and description of state of the art and current research challenges, an overview of CSIC's strategy and advantage position to tackle the challenges identified in each chapter. Over 50 challenges are discussed along this white paper. They represent a very exciting research program for a growing community within the CSIC with more than 150 permanent researchers contributing to AI.

Chapter One addresses the importance of the integration of three classical topics in AI: knowledge (or representation), reasoning and learning. Given the importance of these three areas in the history of AI, the chapter introduces the reader to the broader domain of artificial intelligence. The chapter identifies five challenges, which mostly delve into the integration of the three areas that have so far evolved quite independently. Among the challenges, special mention deserves "large-scale problem solving" that aims at overcoming

traditional limitations in complexity and size of classical AI solutions. Attacking real size problems will require renewing old methods and adapting methods proven successful in other areas of AI. Classical AI is coming back. Our future is that of a mixed society of people and artificial intelligence. A myriad of devices around us will need not only to make intelligent decisions, but they will also need to coexist and coordinate with one another to serve humans.

Chapter Two studies the challenges associated with the development of theories, and supporting technologies, for that envisioned future society. Finding ways to relate the micro world (individual autonomous agent interactions) with the macro world (the properties we seek in those large complex societies) is a key priority. This challenge requires collective and multidisciplinary effort in order to find solutions for current pressing needs—think for example of the massive adoption of autonomous vehicles and of the coordination problems arising from the coexistence of humans and autonomous systems).

Chapter Three focuses on the thriving area of machine learning. To a large extent current interest on AI derives from the spectacular results recently obtained by machine learning techniques. On the one hand, the colossal amount of data collected by administrations, government and corporations is calling for methods to use them in forecasting and analysis. On the other hand, the growing computing power at our disposal makes the use of computing-intensive AI methods feasible. The application of ML methods to a large variety of domains is reflected in the large number of groups, and thus contributing authors, involved in this chapter. Many applied challenges have been identified. Just to mention a few, the use of ML to forecast disease propagation has a tremendous potential in the mitigation of the impact of outbreaks like the current one of COVID19. There are also theoretical challenges like the Interpretability and Explainability of ML results that require urgent attention. Otherwise, we risk that these results are not accepted by our society. The embodiment of intelligence has been one of the most attractive business cards of AI.

Chapter Four presents the fascinating area of Intelligent Robotics. A number of exciting and difficult challenges are included in this chapter. One of them is particularly relevant in the context of the mixed societies mentioned in the second chapter: How to build robots that are easy to reprogram and to adapt to changing circumstances through a process of continuous learning. Robots, among many applications, will assist us and will therefore need to learn our

preferences, adapt their pre-programmed skills to our context and understand the consequences of their action under unforeseen circumstances. These are very difficult functionalities. Also, smart robotics has ethical implications that connect this chapter with the challenges discussed later in Chapter 6.

Chapter Five explores the contribution of computational cognitive models to the design and development of AI systems. These models aim at describing and simulating human cognition and behaviour by understanding the principles of intelligent action from the study of living beings and by reproducing these principles in the development of intelligent devices that mimic, simulate, and expand the cognitive and physical capabilities of living beings. Through the understanding of artificial consciousness and the design and development of computational psychological and social models and cognitive architectures, this chapter addresses difficult scientific challenges such as the possibility of creating behavioral human clones that can contribute to enhance the capabilities of cloned individuals, detect behavioural disorders and simulate preferable reactions and approaches that the cloned subject can imitate in order to adapt to difficult situations. Finally, the clone would ensure the persistence of a person's way of reasoning and memory after death; the clone would preserve these memories and hand them over to future generations.

Chapter Six discusses the ethical, legal, economic and social implications of mass scale deployment and adoption of AI systems. From a constructive standpoint, the chapter explores how to embed ethics in AI and in engineering through innovative multi-disciplinary collaborations and educational curricula that include the study of human and machine biases, the characterisation of moral agency in artificial entities, and the development of specific "robot laws" tackling the issue of the legal personality of autonomous systems. The effects of AI on the economy and society also demand special attention, especially the development of governance and R&D mechanisms to ensure the beneficial co-evolution of humans and artificially intelligent systems.

Chapter Seven addresses the demand for implementing high-speed low-power systems that can perform intelligent tasks, while consuming an affordable amount of energy and computing resources. By drawing inspiration from the capacity of biological brains to solve cognitive problems using low-power and low-speed noisy computational neurons, this chapter sheds light on the creation of high-speed sustainable hardware for AI; a type of hardware that would overcome some of the constraints of current hardware solutions based on traditional Von Neumann computer architectures. Research challenges explored

in this chapter include the development of systems featuring high-speed reaction, high-accuracy recognition, robust learning capabilities acquired by means of neuromorphic sensors and processors, Spiking and Artificial Neural Networks.

Chapter Eight focuses on smart cyber security and addresses the challenges of developing secure, safe and privacy-respectful AI systems ranging from the hardware to the application layer by enacting security-by-default principles and a more robust, formal analysis of adversarial machine learning models. By aligning AI functionality with information security requirements, this chapter tackles research challenge highly relevant to organisations and societies. An example of these challenges is the automatic detection of misinformation campaigns, especially those involving scientific findings, and the deployment of counter-strategies to debunk false information while reducing opinion polarisation.

CHALLENGE 1

INTEGRATING KNOWLEDGE, REASONING AND LEARNING

Coordinators
Felip Manyà (IIIA, CSIC)
Adrià Colomé (IRI, CSIC – UPC)

Participant researchers and centers
G. García de Polavieja (IC, CSIC)
D. Ríos Insua (ICMAT, CSIC-UAM-UC3M-UCM)
A. Torres Barrán (ICMAT, CSIC-UAM-UC3M-UCM)
E. Armengol (IIIA, CSIC)
C. Blum (IIIA, CSIC)
T. Flaminio (IIIA, CSIC)
L. Godo (IIIA, CSIC)
J. Levy (IIIA, CSIC)
P. Meseguer (IIIA, CSIC)
J. Segovia (IRI, CSIC).

1. EXECUTIVE SUMMARY

Knowledge, reasoning and learning (KRL) play a central role in *artificial intelligence* (AI) and are instrumental in solving many AI complex problems. In such problems we may have massive amounts of data and imprecise models, and the goal is to create AI systems that scale well in such scenarios, which often requires combining KRL techniques. In this chapter, we first present KRL from an historical perspective and then identify future research directions in the KRL domain in which CSIC is or could be very competitive. In particular, we present challenges integrating KRL from several perspectives, such as:

- General planning: finding AI agents that combine learning and planning techniques to be able to learn solutions to a problem and generalize them to others.
- Problem solving: create algorithms with learning capabilities for solving complex optimization problems with huge amounts of data.
- Learning adaptable AI agents from a reduced amount of data, keeping the most information possible from data, while being computationally and sample efficient for adapting to changing situations.
- Enhancing logics for conditional, causal reasoning and their integration with *machine learning* (ML).

- Integrating uncertainty, similarities, knowledge and learning by using stochastic methods that better represent the variability in real-world scenarios.

These challenges are considered key topics in the current AI research and, with CSIC already having a good knowledge on them, pushing research in those directions could place CSIC as a reference institution in the world on these fields. We detail a plan and resources needed to boost research on the integration of KRL techniques.

2. INTRODUCTION AND GENERAL DESCRIPTION

Little can be truly understood on any human endeavor if its history is unknown or comprehended superficially. The AI discipline is not an exception. Trying to compose a coherent view of this topic, we start with a brief summary of its history (Buchanan, 2006). Later, we provide a more detailed description of some particularly relevant aspects.

2.1. Historical Background

Returning to the first AI times is as coming back to the initial computer days. The decryption of the Enigma code during the Second World War was crucial for its outcome. This task was accomplished by a group of cryptographers—including Alan Turing—using very primitive computers. After the war, USA and UK started the development of general-purpose computers[1]. Turing was one of the few able to conceive entirely one of those, and he suggested that computers could be used to simulate intelligent behavior. He wrote the first computer program to play chess in 1948. He published these ideas in his famous paper *Computing Machinery and Intelligence* in 1950. Turing died (probably committed suicide, after convicted from homosexuality) in 1954 (Hodges, 2012).

Two years later, the Dartmouth College, a small university in the US East Coast, hosted a workshop entitled *Summer Research Project on Artificial Intelligence*. It grouped a handful of researchers interested in using computers for intelligent action. This meeting was the starting point for public research on AI (Feigenbaum and Feldman, 1963) (Newell and Simon, 1972). The following years were a period of enthusiasm, where several researchers made *exaggerated claims* about results that could be expected in the near future. None

1 ENIAC (1946, Pennsylvania Univ.) was one of the first general-purpose computers.

of them became true, which caused an increasing scepticism for these techniques (Lighthill, 1973). In the 70's, there was a shortage of public funding for AI, a period called "the first AI winter". Researchers thought about the difficulties they were facing. The expression *combinatorial explosion* became familiar, referring to the multiplicative effect that happened when considering all options when unfolding a problem, a huge quantity soon unmanageable. They realized the role of domain knowledge, able to reduce the enormous search spaces of the initial AI formalizations to spaces of reasonable size. Then, knowledge became the golden element of any AI system, and ways to store, represent and use it were developed, causing the area of *knowledge representation* to become central in AI (McCorduck, 2004).

In the 80's, the interest for AI renewed, specially due to the popularity of *expert systems* and *neural networks*. Expert systems were AI programs trying to imitate the reasoning of a human expert in a particular domain (Mitchie, 1979). Their elements were symbols and their knowledge was codified using mostly production rules (Buchanan and Shortliffe, 1984). Neural networks were AI programs inspired in the brain structure, with many elementary artificial neurons that implemented a non-linear function and with high connectivity (Rumelhart and McCleland, 1986). Typically, these neurons were disposed forming layers, each neuron connected with a weight to each other neuron of the preceding layer. Trained with a set of examples, the algorithm of backpropagation was able to adjust weights to minimize the error. After discovering the role of knowledge, many daydreamed with the existence of a universal form of knowledge representation, to be used in any AI application, but it turned out not to be possible.

A "second AI winter" ocurred at the end of the 80's and the starting of 90's, after becoming apparent the limitations of expert systems and neural networks (McCorduck, 2004). From the 90's on, the attention has moved, slowly but continuously, towards particular knowledge representations with very efficient inference that allowed to face complex problems with competitive solving times.

2.2. Current State Overview

AI has become a large field of science, and it encloses a wide variety of theories and applications. Here, we present some of them:

Problem solving. In the early days, this term denoted a number of puzzles, cryptoarithmetic and logical games: well-defined problems of small size that could

be solved by searching its *state space*: the space of all possible configurations and transitions between them. The idea of *heuristic* was introduced (Pearl, 1984) as an estimation of the distance between two states. The notion of optimization appeared naturally, when some solutions were preferred to others, and integrated in procedures like *A** (Hart, Nilsson, and Raphael, 1968). While all of the above appeared in the context of complete and *exhaustive search*, the growing size of the spaces in which to search and optimize were a serious drawback for this approach. Trading completeness for performance, *local search* followed a different approach. The term *metaheuristic* (Blum and Roli, 2003) was introduced in order to describe higher level heuristic techniques that combine basic heuristic components such as greedy algorithms and local search with the aim of exploring search spaces. In the case of problems completely defined by *constraints*, a number of models and procedures have been developed. All the mentioned problems are single-agent: only one agent acts in the resolution process. Nowadays, *adversary games* (with two or more agents) are also considered inside problem solving (Russell and Norvig, 2010).

Knowledge representation. As mentioned before, one of the recurrent difficulties of AI was *the issue of scaling*. New solving methods were demonstrated on instances of small size, and worked perfectly well. But when instances grow, methods did not scale up efficiently to be considered actual solving options. This was the basis for the criticism on *toy problems* raised in the 70's. The introduction of specific *domain knowledge* (Buchanan, Sutherland, and Feigenbaum, 1969) for the considered problem was essential to reduce the huge search spaces of many AI applications when formalized under *good old-fashioned AI*. Given that it was required to operate with existing knowledge to produce new one, *classical logic* was considered as a direct candidate to represent knowledge (Kowalski, 1979). Also, other methods were proposed as production rules, semantic nets, frames, blackboards, etc. Soon, the necessity of *control knowledge*, to focus the part of domain knowledge that should be used, become apparent. Coming back to classical logic, it appeared not to be well-suited to capture all the peculiarities of human knowledge, full of exceptions, with a few absolute true facts, invaded by uncertainties and inaccuracies, and time-dependent. This caused a movement towards *non-classical logics*: multivalued, fuzzy, temporal, modal, non-monotonic, description and others. Also, non-logic based formalisms were developed (Branchman and Levesque, 2004). *Hierarchies* try to solve issues like default reasoning. In a broader sense, *ontologies* aim at capturing the elusive aspects of common sense reasoning.

Symbolic vs subsymbolic AI. Over the years, the inclusion of knowledge in AI systems has followed two main approaches: symbolic and subsymbolic. On the one hand, in the symbolic representations of knowledge the smallest elements composing it are symbols. For instance, when we write "all men are mortals" ($\forall x, men(x) \rightarrow mortal(x)$), the properties *men* and *mortal* are symbols of functions that apply on variable x that instantiates over other symbols (identifiers of agents). *Symbol manipulation* has been very intense in AI: remember that (i) the programming languages LISP, one of the more used in early AI applications, offered an elaborated ways to handle symbols; (ii) Newell and Simon, in the reception of the Turing award in 1975, formulated their famous hypothesis of *physical symbol system* (Newell and Simon, 1976). On the other hand, in the subsymbolic representation (also called connectionist or neuronal) the smallest elements are of lower entity than symbols. These representations are inspired in brain models, where the elementary computing unit is an *artificial neuron* (its name and function are inspired by biological neurons). These artificial neurons are highly connected. A popular model in the 80's and 90's was the *feed forward neural network*, where neurons were disposed in layers and each neuron was connected with all neurons of the preceding layer (Rumelhart and McCleland, 1986). Neural networks evolved until today, where they capitalize the success of *deep learning* (DL) (LeCun, Bengio and Hinton, 2015).

Learning. Many consider that learning is a nuclear part of intelligent behavior. In AI, however, it has been considered as another element of intelligence, not included in every AI system. For example, the computer chess program Deep Blue (Campbell, Hoane and Hsu, 2002)—famous because it won the world championship of chess in 1997—did not have the learning capacity: its strategy was parallel alpha-beta search and its behaviour did not change because of past experiences. ML (Mitchell, 1997) encapsulates these issues. ML systems perform learning from a set of experiences—denominated the training set—and its performance is measured on a set of different cases—called the test set—. In *supervised learning*, each case includes the correct answer (in classification tasks, frequent in ML, the class to which it belongs). This information is missing in *unsupervised learning*, where cases are grouped by similarity or closeness of some features. *Semi-supervised learning* seems to be between these two extremes, where some limited forms of external solution are provided. An example is *reinforcement learning*, where an agent tries different elementary actions to achieve a goal that requires a sequence of actions (for instance, to exit from a maze), and the only information provided is when it reaches the goal.

The agent has to distribute the *merit* of finding the goal among the actions of the successful sequence. With a random generation of actions, it is expected that *good actions* will receive enough reinforcement to make them to stand out among the others. The dichotomy symbolic vs subsymbolic appears in ML with special strength, because the principal subsymbolic systems have appeared in this area.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

3.1. Basic Science

During the last twenty years, AI has been growing rapidly and has been widely used in applications that vary from robot learning (for planning, control, navigation, etc.) to data mining (for suggesting products to buy to users), with other applications like for cleaning robots (roomba) and even washing machines with fuzzy logic.

In the early 2010s, the theoretical developments –from the previous 20 years– on neural networks were revisited, originating DL which caused significant improvement on computational capabilities. However, DL techniques required huge amounts of data to adjust the high quantity of weights to learn complex neural networks.

In the year 2014, the need of huge labeled datasets for DL was partially solved with *generative adversarial networks* (GANs) (Goodfellow et al., 2014). A GAN consists of two neural networks that compete with each other, where the first learns to improve predictions (generative) while the second learns to make the former to fail (discriminative). This enables models to tackle unsupervised learning in large datasets such as image datasets.

However, *new trends in AI* also account for issues that DL has brought up, such as explainability, sample efficiency or uncertainty. Currently, the level of performance of some AI applications are close to human level, and these systems are being deployed in human society. This poses new challenges of *ethical* and *legal* nature, specially for AI systems able to take decisions and to operate autonomously. A new subarea of AI to address these new issues is currently under development.

Successful cases

- *Expert systems (1984)*. In the 70's and 80's several expert systems appeared (Dendral, Mycin, Xcon...), emulating the reasoning of human experts. Their knowledge was represented in declarative form (production rules, other forms), to deduce new knowledge by inference.
- *Handwritten text recognition (1995)*. Several companies allowed to recognize handwritten text, using from hidden markov models to neural networks. In the last years, DL offers better performance.
- *Deep Blue (1997).* The world champion, Gary Kasparov, was defeated in a chess competition by the computer program Deep Blue, with international tournament rules. Developed by IBM, used special hardware and did not incorporate learning technology.
- *Autonomous cars (2005, 2007, 2009).* DARPA launched several contests to promote research on self-driving cars. Google took these advances; later, Waymo inherited this know-how. Such cars integrate a wealth of KRL techniques.
- *Image recognition (2010).* Now dominated by DL, it allows internet searchers to find images by their content.
- *Laundry Robot (2010).* In 2010, UC Berkeley integrated learning algorithms with vision, planning and navigation methods in a robot that could fold laundry. While the execution was very slow, being able to perform such complicated tasks was a major success.
- *Watson (2011).* In 2011 the Watson program competed in a question answering TV show against two qualified human participants. Watson won. It was developed by an IBM team; now this technology is commercialized to handle large quantities of information.
- *Deep Face (2014).* A face detection algorithm with a success ratio that matched that of humans (27% improvement over previous algorithms) was developed using DL.
- *Alpha Go (2015).* A Google team was able to win the world champion of the Go game, on which the strategies of Deep Blue did not work. They combine *Monte Carlo tree search* and *deep reinforcement learning*.
- *Automatic translation (2017).* Today neural machine translation, based on DL, dominates the world of translators offered in internet.
- *Robots doing parkour (2018).* The DARPA robotics challenge (2012-15) aimed at developing better robotic platforms. The company Boston Dynamics combined the latest control techniques with AI methods in a humanoid robot that can jump, do backflips and run.

3.2. Potential Applications

With the growth of the internet, cell phones, social networks, online shopping, there is more data every day, resulting in a new subfield of AI: *data science* (DS). Additionally, with the availability of better software and hardware, robotization of the industry is also becoming a reality. ***While robotics is a multidisciplinary field,*** including electronics, dynamics, control and programming, ***AI is what makes robots learn to act, recognize their environment, optimize their behaviour, and so on.***

We foresee a number of potential applications based on pattern recognition, such as voice, face and gesture recognition, biometry, etc. In a broader sense, vision and video analysis can also improve.

Nevertheless, important issues remain to be solved in the future. For instance, the problem of adding *common sense knowledge*: it is the knowledge about how the world works, learned by children at early age, that everybody knows but nobody explicitly includes. It acts as a kind of "background knowledge" and it is really important to understand our word and forsee correctly the events that happen in it. Integration is another big issue. Although some early attempts have been done (Veloso et al., 1995; Kaelbling et al., 1998; Geffner and Bonet, 2013; Ghallab et al., 2016), the whole problem remains largely unsolved: we are able to develop very sophisticated applications but we still do not know how to connect them in a strong sense, allowing information to easily flow and affects other applications (using a construction metaphor, now we are able to produce very sophisticated bricks but still we do not know how to make beautiful houses with them).

4. KEY CHALLENGING POINTS

In this section, we present the key AI challenges that, given the current state of CSIC, should be tackled by its research institutes.

4.1. Searching General Plans

Automated planning (Ghallab, Nau, and Traverso, 2004) is one of the branches of AI (also named AI planning) where the representation of the world is known and agents should follow a strategy, plan or sequence of actions to solve a planning problem. This decision-making technique is considered general in that solutions to problems can be found independently of the world representation. However, once a plan is computed it is no longer valid to solve

other problems, and for every new problem a new plan must be computed from scratch.

The challenge of finding ***solutions that work not for one but for many problems*** or even solving all problems on a given planning domain (representation of the world), got researchers attention in the planning community. This extension of automated planning is called *generalized planning* (Srivastava et al., 2008; Bonet et al., 2009; Hu and De Giacomo, 2011) which aims to compute an algorithm-like structure that solves a set of planning instances in a given domain. We refer to this as *intradomain* generalization, where all planning problems must share some properties or features which conditions the algorithmic solutions to make the next decision. More recently, the problem of *intra-domain* generalization has been tackled as a classical planning problem, where solutions have the form of programs (Segovia-Aguas, Jiménez, and Jonsson, 2019) or automata (Segovia-Aguas, Jiménez, and Jonsson, 2018) and take advantage of succinct representations allowing hierarchical structures and recursivity. However, all these approaches suffer from scalability when computed, the number of controller states and the size of the features set are the bottlenecks when solving the problem. Thus, one of the challenges in this field is on ***improving the scalability when computing solutions that generalize***.

Decision-making problems have also been addressed with success by other techniques such as transfer (Taylor and Stone, 2009) and reinforcement learning (Kaelbling, Littman, and Moore, 1996). These techniques are more flexible than automated planning in that they can be used without a representation of the world (in other words, these approaches can be model-free). While generalized planning is focused on *intra-domain* generalization, learning techniques are focused on *inter-domain* generalization, where a solution to a domain or policy can be used to solve problems from a new domain. Although, the challenge when transferring knowledge from a domain to another one depends on ***learning the correct set of features,*** and converging to a general policy may be hard whether rewards are sparse. Hence, this requires long sequences of actions which yields to useless explorations, so finding ***new methods that minimize exploration*** when rewards are sparse is an open challenge in the reinforcement learning field.

Thus, the main motivation and challenge of generalization in planning is on ***bringing together planning and learning communities to find approaches and solutions that generalize*** in both ways like we humans do, *intra-domain*

and *inter-domain*.

4.2. Creating Robust and Hybrid Optimization Solvers with Learning Capabilities for Large-scale Combinatorial Problem Solving

The class of *combinatorial optimization* (CO) problems is very important in practise. Each CO problem has a set S of solution candidates, which is exponential in the input parameters of the problem. Moreover, an objective function f assigns a value $f(s)$ to each solution candidate $s \in S$. The goal is to find a solution $s^* \in S$ with $f(s^*) \leq f(s)$ for all $s \in S$. Prominent examples include routing, transportation and bioinformatics.

Many relevant CO problems today need more precise models and involve a huge amount of data. On the other hand, only few optimization solvers incorporate sophisticated learning mechanisms. As a consequence, optimization techniques used so far do often not scale well on those problems and must be revisited. New solving techniques must also be devised to adapt to this challenging scenario. Therefore, as a challenge in the field of CO we propose the creation of ***robust and hybrid optimization solvers with learning capabilities for large-scale problem solving.*** In the rest of the section we point out three research directions to address this challenge.

Hybrid problem solvers

Problem solvers in CO can be divided into (1) the *class of exact techniques*, and (2) the *class of heuristics*. Exact methods compute a provenly optimal solution in finite time. The class of heuristics comprises approximate optimization algorithms. It includes simple heuristics and also metaheuristics. Ant colony optimization, genetic and evolutionary algorithms, iterated local search, simulated annealing and tabu search are typical representatives of metaheuristic algorithms. Each of these metaheuristics has its own historical background. Some metaheuristics are inspired by natural processes such as evolution and swarm intelligence, others are extensions of less sophisticated algorithms such as greedy heuristics and local search.

Different research communities working on metaheuristic techniques co-existed without much interaction during two decades. Their focus was exclusively on the understanding of how a certain metaheuristic worked, and how this metaheuristic could be applied to different CO problems. Only when it became clear that pure metaheuristics had reached their limits, research on

metaheuristics for CO experienced a noteworthy shift towards the combination with other techniques for optimization. Over the last years, algorithms were reported that do not purely follow the paradigm of a single metaheuristic. On the contrary, these hybrid approaches combine various algorithmic components, often originated from algorithms of other research areas on optimization. These approaches are commonly referred to as hybrid metaheuristics. The main motivation behind the hybridization of different algorithms is ***to exploit the complementary character of different optimization strategies,*** that is, hybrids are believed to benefit from synergy. Unfortunately, developing an effective hybrid approach is in general a difficult task which requires expertise from different areas of optimization. Moreover, the literature shows that it is nontrivial to generalize (a specific hybrid algorithm might work well for a certain optimization problem, but it might perform poorly for others). There are very few hybrid algorithms that are general enough in order to be applicable to a wide range of problems. Probably, the most well-known example is *Large Neighborhood Search* (LNS). Other approaches such as *Construct, Merge, Solve & Adapt* (CMSA) (Pinacho, López-Ibáñez, and Lozano, 2016) are based on the idea of problem instance reduction. One of the challenges of the near future concerns the development of generally applicable hybrid techniques for those cases in which algorithms such as LNS and CMSA do not perform well.

Large-scale problem solving

ML, compressed sensing, natural language processing, truss topology design, computational genetics and transportation are some examples of application domains in which it is nowadays usual to design optimization problems with tens of thousands or even millions of variables. ***The growing need for solving large-scale optimization problems has a number of reasons, including the following ones:***

- In early years, researchers and practitioners modelled optimization problems in a simplified way, resulting in problem sizes that the available optimization techniques were able to deal with. However, with the steady rise of computational power and advances in research on optimization techniques, ***problems are modelled in a more and more detailed way,*** leading to problems of increasing size and complexity.
- The general tendency to decompose complex problems into a series of sub-problems of reduced size has been changing during the last two

decades. Instead of decomposing a problem and solving it in parts—potentially loosing in this way high-quality solutions to the original problem—researchers and practitioners rather try ***to solve such complex problems in one go.***

- Due to globalization, ***the scale of industrial players has been constantly increasing*** during the last decades. For example, optimizing the time table of Ryanair DAC, with 152.4 million of total scheduled and chartered passengers in 2019, generates an optimization problem of a size that is not comparable to those from 20 years ago.

Unfortunately, many modern optimization algorithms, while exhibiting great efficiency for problems of rather modest dimensions, are not designed to scale to large-scale problems and are hence often not applicable since large datasets make standard operations expensive: evaluation of the gradients, computing hessians, Newton steps, line searches, etc. In general, an exact minimizer of the stated objective is not necessary in ML and often an approximation is good enough.

At least the following two trends can currently be observed in the research community when faced with large-scale problems: (1) rather simplistic methods, some of them having been proposed decades ago, are experiencing a comeback, even though in modern forms (examples are simple randomized methods such as simulated annealing); (2) adapting modern algorithms from the AI and OR communities in order to be applicable to large-scale problems. Examples for this latter trend can be found in (Byrd et al., 2016; Deb and Myburgh, 2016) (concerning real-valued optimization) and in (Yang, Tang and Yao, 2019; Hooker, 2019) (concerning combinatorial optimization). Another example is adapting existing optimization algorithms to take advantages of recent advances in computing such as automatic differentiation frameworks (Tensorflow, PyTorch), parallelization in multiple CPUs/GPUs and distributed versions of current algorithms. This is not an easy task since it must be performed on a case by case basis.

Integrating machine learning in combinatorial optimization

ML can serve at least two purposes in CO (Bengio, Lodi, and Prouvost, 2018). The first one concerns CO problems in which expert knowledge on the problem is available. Some computations, however, might be very heavy and time-consuming. In this context, the term expert knowledge refers to, for example, well-working greedy functions or to efficient neighborhood operators

for local search. Heavy and time-consuming computations might focus on the calculation of decisions or the objective function value. ***ML can be useful for such problems in order to obtain a cheap approximation for the heavy computation tasks in a generic way, without the need to develop new and explicit algorithms.*** The second one of the possible ways to make use of ML techniques within optimization algorithms concerns those cases (problems) for which either no expert knowledge is available, or for which the available expert knowledge is often misguiding the optimization algorithm. ***The goal is therefore to make use of ML for the discovery of new and generic policies for guiding the decisions of the optimization algorithm.*** In the former case (ML to approximate decisions), a policy is often learned by so-called imitation learning. The expert knowledge serves as an oracle that provides demonstrations for learning, and the goal is to minimize the distance between the decision of the expert/oracle and the decision provided by the approximation obtained after learning. In the latter case (discovering new policies), the policy is learned by reinforcement learning through experience, that is through a maximization of the obtained reward.

One of the main challenges in this area of research concerns the fact that algorithms for optimizations that make use of ML techniques for solving subtasks may suffer from the same problems as ML techniques in more traditional application scenarios. That is, current ML algorithms are generally good in generalizing to examples that are close to the distribution of examples used for learning; they are, however, not so good in generalizing to examples far away from the distribution. In general, the research on both lines outlined above has just started, and not many examples exist. Therefore, there is a lot of room and potential for ground-breaking research in this field.

4.3. Learning Adaptable Policies from Small Datasets

DL, a recent advance in ML, is showing impressive results. However, they rely on the availability of having a large dataset. While this can be true for vision applications, with billions of images available on the web, the majority of real-world datasets are still in the gigabytes range, with robotics manipulation being one example application.

For these applications simpler (first-order) algorithms are usually preferred, specially for non-convex problems like deep neural networks, but more complex algorithms and/or accelerated versions can help. One example are active

sampling strategies, which help to select the next batch of samples to optimize in the ubiquitous stochastic optimization algorithms. Another example are old acceleration approaches that are recently being rediscovered, like Nesterov's acceleration of gradient methods, that dates from 1983.

Imagine a case where a robot wants to learn a certain motion that involves manipulating an object that cannot be precisely modelled. This results in an impossibility to simulate a reinforcement learning scenario, and requires real-world executions, which are usually of the order of a few hundred samples, compared to the millions (or even billions) used for *big data* (BD) applications. While robots can be automated to sequentially execute samples from a policy, those samples are also costly due to time constraints and robot usage. In such situations, the sample complexity - the amount of samples we will execute in a real world scenario for learning is crucial: ***we want to learn an optimal (or sub-optimal) policy with as fewer samples as possible.***

This field of AI is gaining importance over the last years (Deng et al., 2018), and it focuses several challenging aspects:

- Minimal policy representation: Search spaces for real-valued policies are often high dimensional, depending on the functional used for representing a policy. Moreover, searches in high-dimensional spaces are very inefficient, partly due to the *curse of dimensionality*, referring to the increase (w.r.t. dimension) in sparsity of samples of a dataset. Therefore, policies with thousands (or millions) of parameters to optimize such as neural networks are not valid when samples are at the tens or hundreds. In such cases, more compact policy representations need to be used, or apply dimensionality reduction techniques on the policy functionals in order to reduce the search space.
- Active sampling: Having a compact policy representation might not be enough. Therefore, latest works on sample-based policy optimization often generate new samples not only based on the stochasticity of the model, but also in an expected information gain. As an example, stochastic surrogate models for a reward function in reinforcement learning are built, and samples are generated by estimating the expected reward and variance (Chatzilygeroudis et al., 2020).
- Information theory - minimum information loss: Not only it is necessary to have a compact representation and efficient sampling, but also to use as much information as possible. In those terms, information theoretical approaches, based on terms such as entropy, are often used to both have

a model that does not lose any information from data and for optimizing a policy.

- Adaptability: As stated before, generalization is an open problem in AI. In most cases where there is a contextual variability provided by an external feedback, an AI agent must be capable of adapting to such contextual feature. In the case of small scale problems, approximate and stochastic models are often used, but their success is still limited with a reduced amount of data. Another expression of adaptability is to make policies that are learned in a particular platform adaptable to a new platform, transferring their knowledge.

4.4. Pushing Logic-based Knowledge Representation Methods for Advanced AI

Logic-based methods have always played an important role in AI, and prominent members of the AI community (Darwiche, 2018) agree that logic has still an important role to play as a tool to formalize reasoning and representation models, and to serve as a common ground for the integration between symbolic and sub-symbolic AI. The relations between AI and logic depart from issues having strictly to do with reasoning, like representing conditional knowledge and drawing correct causal inferences, or reasoning about uncertainty are central topics of knowledge representation that need logic to be correctly formalized. Besides that, a new crucial challenge for logic in AI is concerned with providing tools and methods to be integrated with ML and, by doing so, to contribute to the development of a more transparent and better explainable AI.

Enhancing conditional and causal reasoning

Notwithstanding the prominent role that classical logic has played, and still plays, in the area of knowledge representation and reasoning, the logic community of AI have developed a plethora of alternative systems capable to handle more general reasoning patterns and deductive models. Among them an important role is surely played by the wide family of non-monotonic logics, i.e., formal frameworks devised to handle *defeasible inference*. Under the wide umbrella of non-monotonic reasoning two general topics aim at capturing different aspects of intelligent agents and have occupied a special place in the logic for AI: reasoning with conditionals and causal reasoning.

Conditionals are expressions that represent a general statement of the form "*b given a*", usually denoted by $(b \mid a)$. Although a and b may be formulas of

classical logic, there is no way to define $(b \mid a)$ in the language of classical logic without expose ourselves to known triviality results and paradoxical behaviors (Flaminio, Godo, and Hosni, 2017). This limitation of the expressive power of classical logic called for alternative approaches for representing and reasoning about conditionals, and it paved the way to several lines of research which, if from one side directly involve conditional uncertainty measures in general, and conditional probability in particular, from the other side try to understand what conditionals are as pure logical objects and hence outside the bounds of uncertain reasoning. The latter is known as the area of measure-free conditionals. Although this line of research initiated in the 60's of the last century, it is still a challenging problem to determine alternative methods that could shed light on reasoning with conditionals while, at the same time, being sufficiently general to reconcile the measure-based and the measure-free approaches.

Causal modeling and causal reasoing are nowadays recognized as one of the central topics in logic-based AI

The best known causal models are probabilistic graphical models, in particular Bayesian networks, based on conditional (in)dependence relations. However, other models which transcend probability theory have been introduced to interpret causal query about a specific domain. These try to capture several characteristic features of causality and they can be classified in several distinguished classes each of which is meant to represent, for instance, *associations* (typical questions that these models try to answer are "how are the variables related?" or "how would seeing X change my belief in Y?"), *intervention* ("what would Y be, if I do X?", "how can I make X happen?") and *counterfactuals* ("Was it X that caused Y?", "what if X had not occurred?"). Research on causality have mostly favored the model-based approach, while determining sound and complete axiomatizations for specific classes of causal models is a logical task that, besides a first impetus almost twenty years ago, has not been systematically investigated so far.

Combining methods in differentiable and mathematical fuzzy logics

A key point in the success of ML, and in particular DL, concerns with the availability of high-performance computing architectures allowing to process a large amount of data. However, the collection of training data is often a slow and expensive process, requiring an extensive human intervention. A way to overcome this limitation consists in equipping a learning algorithm with a prior knowledge describing some of the desired behaviors of the functions to be learned.

For representation and reasoning purposes, this prior knowledge is formalized by theories of classical first order logic. However, non-classical logics have also been designed to handle them already at the propositional level. A particularly relevant class of logic of this kind is that of *differentiable fuzzy logic* (DFL) (van Krieken, Acar, and van Harmelen, 2020). In this setting, truth values of ground atoms are not discrete as in the classical case, but continuous, and logical connectives are interpreted using real-valued functions, in particular fuzzy t-norms, t-conorms and S-implications. The objective of DFL is to maximize the satisfaction degree of the full grounding of the (fuzzy) knowledge base. Although DFLs have been already employed and integrated with learning algorithms, they lack of a precise formal definition and, more importantly, of methodological formal tools that may be used to deepen the investigation in that field.

For a completely different purpose, formal logical systems based on t-norms, called t-norm based fuzzy logics, have been developed and well-studied in the last two decades. These logics are studied by a branch of mathematical logic called *mathematical fuzzy logic* (MFL) which encompasses a wide family of logical systems all of which share the common property of evaluating formulas in the real unit interval [0, 1], rather than the classical two truth values 0 (false) and 1 (true). The development of this field took place essentially within the bounds of pure mathematical logic and nowadays several methodological tools such as proof theory, game semantics, abstract algebraic logic, categorical equivalences, complexity classification and many others are available. Moreover, several points of contact with other areas of mathematics such as functional analysis, probability theory, universal algebra, real convex geometry, have been explored and are a fruitful ongoing line of research (Cintula et al., V1, 2011; Cintula et al., V2, 2011; Cintula et al., V3, 2015) .

A fundamental step towards integrating subsymbolic and symbolic methods in AI is hence ***to understand up to which extent the already mature methodology developed for MFL can be transferred to DFL and hence integrated in learning algorithms.*** Furthermore, a knowledge transfer between DFL and MFL would be instrumental to endow the former with reasoning models, proof-theoretical calculi, algebraic semantics and other important tools that, as we already recalled above, have been already developed for MFL.

4.5. Integrating Uncertainty, Similarity and Learning

In many areas of AI, for example in robotics, some autonomous systems should be able to perform in an environment that only resembles partially the one in

which they have been trained. Hence, ***the underlying reasoning models should be flexible and capable of recognizing the similarity of the actual context with the artificial one to facilitate decision making.*** So as to act coherently and make rational decisions, agents should be endowed with uncertainty/utility models which perform suitably and robustly under operational conditions. Bayesian methods can facilitate the integration of uncertain knowledge (under the form of a probabilistic graphical model), reasoning and learning in a coherent framework.

Integrating similarity and uncertainty

Betting methods, of which de Finetti's Dutch Book is by far the most well-known, are uncertainty modelling devices which accomplish a twofold aim. Whilst providing an (operational) interpretation of the relevant measure of uncertainty, they also provide the formal setting to tell apart admissible from inadmissible quantifications of uncertainty. Although their theoretical value, these features are particularly important also for applications as they allow to characterize each uncertainty theory by a bunch of human-intuitive criteria which are usually described by a game between several opponents. By doing so, this methodological approach to uncertainty permits to distinguish, at a human-explainable level, which uncertainty theory better performs in a given context and in a given scenario.

Features like similarity, uncertainty and imprecision have been tackled from the perspective of formal methods by exploring several logical systems which, separately, are able to provide a faithful representation of these attributes and also allow to perform reasoning tasks in their presence. However, besides some exceptions, little effort has been made to systematically address the issue of combining two or more of the above features. ***This task of merging similarity-based, uncertain and approximate reasoning would provide applications with a solid background and is also particularly challenging from the theoretical perspective.*** Indeed, formalizing the interaction between different types of reasoning calls for the development of tools and methods which go far beyond the logical systems which, in the last years, have been proposed to reason about each of those features separately.

Integrating knowledge, probabilistic reasoning and learning

De Finetti's arguments also lead to Bayesian methods which facilitate the integration of knowledge, reasoning and learning in a coherent framework, supporting most of the tasks within earlier challenges: knowledge is structured

through probabilistic graphical models; expert knowledge is coded through prior distributions, being specially relevant in small data contexts; evidence learning is based on Bayes rule; reasoning essentially deploys probability calculus; for decision making, planning and optimization, the maximum expected utility principle is usually employed coupled with sensitivity analysis methods to cope for imprecision (for such issues, see e.g. (French and Rios Insua, 2000). The power of *Markov chain Monte Carlo* (MCMC) methods enabled the growth of Bayesian methods within AI. However, recent scenarios brought forward by BD and DL have ballasted these methods with respect to the maximum likelihood paradigm: MCMC methods do not scale at the BD regime. Variational Bayes techniques are much more efficient computationally yet they tend to underestimate uncertainty. Moreover, as cogently shown in many scenarios, including the COVID crisis, we shall have to deal with problems with little data, requiring decisions based mainly just on expert opinion; although there are relevant methodologies for *structured expert judgement* (SEJ) they are pervaded by classical methods. Finally, a major problem related to the ever increasing use of AI systems is that they are subject to malicious attacks, with the field of adversarial ML (Vorobeichyk and Kantarcioglu, 2019) devoted to applying game-theoretic concepts to enhance the security of such systems, despite inherent common knowledge conditions which are hardly tenable. Therefore, three major challenges in this area for the next decade include: (i) ***developing efficient methods for large scale Bayesian computations;*** (ii) ***developing Bayesian methods for SEJ,*** including cases when uncertainties are created by adversarial agents; and (iii) ***developing methods for securing AI algorithms*** possibly based on adversarial risk analysis (Banks, Rios, and Rios Insua, 2015), as an alternative to game theory.

CHALLENGE 2

MULTIAGENT SYSTEMS

Coordinators
N. Osman (IIIA, CSIC)
D. López (IFS, CSIC)

Participant researchers and centers
V. Gallego (ICMAT, CSIC-UAM-UC3M-UCM)
D. Ríos (ICMAT, CSIC-UAM-UC3M-UCM)
M. San Miguel (IFISC, CSIC)
M. Toboso (IFS, CSIC)
P. Noriega (IIIA, CSIC)
J. A. Rodriguez (IIIA, CSIC)
J. Sabater-Mir (IIIA, CSIC)
C. Sierra (IIIA, CSIC)
D. Zurro (IMF, CSIC)
S. Degli Esposti (IPP, CSIC)
F. Herreros (IPP, CSIC)
G. Alenyà (IRI, CSIC)

1. EXECUTIVE SUMMARY

Our future is that of a mixed society of people and AI artifacts. A multitude of devices in our homes will need not only to make intelligent decisions, but they will also need to coordinate with each other to serve us well. Cars will have to coordinate to allow safe road crossings, avoiding accidents. Also, the industry is already beginning to integrate teams of humans and robots collaborating to solve complex problems.

Once there are a multitude of autonomous agents that interact in a collective system, a *multiagent system* (MAS), the notion of autonomous software agent takes a completely new perspective. On one hand, there are the agents, as a primitive entity and, on the other, there is the shared social space where they interact as a different primitive element. Research in this field has commenced with the engineering of autonomous agents and their reasoning and decision making capabilities. This was followed by interest in the interaction and coordination mechanisms of such complex systems, with solutions like agent communication languages, organisational and institutional approaches, competition and collaboration, and team formation popping up. Lately, agreement technologies have been gaining ground, which focus on normative systems, argumentation and negotiation mechanisms, trust and reputation, computational social choice, and semantics.

Over the years, MAS have evolved from systems composed of interacting software agents into socio-technical systems where humans and agents interact. Today, the challenge is governing these socio-technical systems through self-regulation. We envision systems that adapt to the ever evolving needs and values of its actors (be it humans or software agents) (Osman, 2018). This challenge requires a multicentric approach, bringing in fields like agreement technologies, learning, collaboration and coordination, natural language processing, and semantic technologies. These are all mature and well-established fields, but their cross-fertilisation is key for achieving self-regulation in MAS. Furthermore, an interdisciplinary approach that brings in fields like ethics, law, and the social sciences is also critical for helping with issues like wrongful agreements (agreements that are unethical or illegal), human control, and value alignment. This close-knit collaboration will naturally leave a profound impact on the different relevant fields.

Since the birth of MAS, the CSIC has positioned itself as a key player at an international level, contributing to the development of the field along many of its main areas, such as social coordination frameworks, automated negotiation, and trust and reputation, and resulting in a wide and impressive range of application areas. Today, CSIC has proposed a roadmap for self-regulating MAS (Osman, 2018), and its advantage position is its interdisciplinary collaborations between the AI researchers and the social sciences and humanities, both at a national and international level. Although relatively young and growing, CSIC's interdisciplinary work has been rich and promising.

Nevertheless, to address the challenge of self-regulating MAS, there is a pressing need for ***building a multidisciplinary multicentric MAS group*** that allows for even stronger and more effective interdisciplinary research, a key for addressing the big challenges of MAS. Such a group will have the CSIC play an active role at an international level in shaping the future of MAS in particular and AI in general. Overcoming the limits of a CSIC institute, this group will bring in the top researchers of the existing and relevant CSIC institutes to lead this challenging work, opens the door for necessary collaborations with other national centers (for example, the CSIC does not have a law institute, although the contribution of legal studies is critical for MAS research), and attracts top international researchers and laboratories, especially those with a multidisciplinary background.

2. INTRODUCTION AND GENERAL DESCRIPTION

Although we usually talk about artificial intelligence, we should talk about *artificial intelligences*, in the plural. A multitude of devices in our homes will need not only to make intelligent decisions, but they will also need to coordinate with each other to serve us well. Cars will have to coordinate to allow safe road crossings, avoiding accidents. Also, the industry is already beginning to integrate teams of humans and robots collaborating to solve complex problems. Our future is that of a mixed society of people and AI artifacts. Once there are a multitude of autonomous agents that interact in a collective system, the notion of autonomous software agent takes a completely new perspective. On one hand, there are the agents, as a primitive entity and, on the other, there is the shared social space where they interact as a different primitive element.

Generally speaking, what makes *agent-based social coordination* interesting is that some or all of the agents that interact in the shared social space are software agents. Thus the designers of the system and the agents have the possibility of encapsulating behaviour that is amenable for digital content, may be situated in a digital environment and is programmed into the agents endowed with some degree of autonomy.

The way those coordination conventions are established serve to characterise two main types of (agent-based) social coordination systems: *organisations and institutions*. Agent based organisations are inspired on conventional organisations and thus capture their usual constitutive features in such a way the members of the organisation may be software agents (as well as humans in some agent-based organisations). Consequently, coordination conventions in the case of organisation consist of the elements that define the organisational infrastructure with its control of roles, resources and tasks. Agent-based institutions, in contrast, are inspired on conventional institutions and thus capture the notion of devices whose purpose is to establish and enforce conventions that articulate interactions of the autonomous agents, be they software or human, that enter the institution to achieve a particular collective objective.

For this reason, many issues in AI, machine learning (ML) and robotics refer to dealing with the strategic and cooperative interactions in systems with multiple agents. For example, *game theory* emerges as a major mathematical formalism for studying such interactions. The recent large scale computational

AI domains demand new efficient algorithmic ideas for computing equilibria and related solution concepts.

Another key requirement is to have agents (and humans) discuss, argue, negotiate in order to reach agreements (or, in a more general term, to make collective decisions). *Agreement technologies* is an active field with a number of sub-areas, from argumentation, negotiation and social choice theory, to trust and reputation and semantic alignment. *Automated negotiations*, for instance, deal with the case of two or more purely self-interested agents that try to come to a mutually acceptable agreement. Although these agents are not necessarily targeting any form of social welfare optimisation, each agent would be willing to cooperate if (and only if) that allows it to achieve a solution that is better for itself than what that agent could achieve on its own. The field has not only studied the development of negotiating agents, but also the protocols that are used by those agents to negotiate.

It must be noted that many problems in MAS today are understood as *combinatorial optimisation* (CO) problems, like sensor networks, disaster management, and transportation. *Composition and formation of effective teams*, a fundamental problem in MAS where teams of agents may be required to work together may also be understood as a particular instance of combinatorial optimisation. It is also a hot topic, for both companies to assure their competitiveness and for a wide range of emerging applications exploiting multiagent collaboration. In this domain, so far the MAS literature has focused on developing algorithms that help automate team formation and composition by casting it as an optimisation issue. Hence, the problem of continuously forming teams, which is common in *dynamic*, actual-world scenarios where tasks arrive along time to be served (e.g. in crowdsourcing), has been overlooked.

A rising field in MAS whose impact is gaining ground is that of *agent-based modelling and simulation*. In an *agent-based simulation* (ABS), each entity possesses an internal state and an autonomous behavior. *Agent-based modeling* (ABM) is a type of modeling in which the local actions and interaction of autonomous entities (agents), both with each others and the simulated environment, are modeled using an ABS approach. ABM is based on the assumption that higher-level system properties emerge from the interactions of lower-level (and usually simpler) subsystems and that this process can be observed and studied using computer (agent-based) simulation. ABM allows evaluating social norms and *human agency*. We analyse the role played by agents and by variables involved in social and historical phenomena which are

unrepeatable also fosters transdisciplinary research, giving, for instance, the possibility to analyse socio-ecological systems (human-environment relationships).

Last, but not least, the issue of ethics in AI has attracted a lot of attention lately, and the concerns are just as applicable in MAS. AI scandals on topics like algorithmic bias[1] and the appalling use of our private data[2] have been highlighting the urgency of the issue even further. Developing MAS that are beneficial, accepted and trusted by society is one of the main challenges of this field today. A technology should not be evaluated only on how it contributes to efficiency and productivity, but also for the way in which it can create certain forms of power and authority (Winner, 2001). The social adoption of technological systems implies a series of determinants related to human relations, habits and customs that often favor ethical, moral and political values such as inclusivity or exclusivity, centralization or decentralization, equality or inequality, empowerment or disempowerment (López et al., 2020).

A late roadmap (Osman, 2018) proposes building the foundations for *responsible open systems*, where users' needs and values are at the core of the design, development and evolution processes of these open systems, and where the users are given back control over the technologies that mediate their interactions. Such responsible MAS are in fact inspired by the European Union's Declaration on Responsible Research and Innovation (RRI), which states that RRI is the on-going process of aligning research and innovation to the values, needs and expectations of society. The proposal bases MAS on people's needs and values, and puts people in control. This is aligned with the many initiatives on ethics and AI, all of which place *value alignment* and *human control* as some of the basic principles of ethical AI (Floridi and Cowls, 2019).

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

3.1. Basic Science Panorama

One main contribution of MAS to other disciplines is a means to describe and implement collective rationality. Thus one may use social coordination technologies to *formulate theories, define models and implement applications involving several rational entities under a given form of governance*. This

1 https://www.technologyreview.com/f/614626/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/
2 https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation

characteristic is quite significant for the social sciences and the general impact for ethics, law, economics and applied social sciences in general will be discussed in more detail in Chapter 6. But it is also significant for other areas like computer architecture, and programming because agent-based social coordination involves specific forms of coordination devices like multi-objective planning, task allocation, resource sharing and collective problem solving.

Traditional game theoretic concepts are at the core of many basic science developments in areas such as medicine (e.g cancer research); biology (evolution models); economics (market competition); political science (dynamics of international relations); computer science (communication protocols); to name but a few. Mathematics and computer science also provide foundational concepts and algorithms for game theory. Therefore, we would expect that developments in this field to impact the aforementioned sciences.

Combinatorial optimisation problems appear in a wide range of scientific areas. For instance, in bioinformatics (e.g. determining the 3D-structure of proteins); economics (clearing of combinatorial markets); operations research (e.g. finding shortest or cheapest round trips in graphs).

The computational reputation and trust models will also naturally play an important role in the areas of social psychology, trustworthy AI and ethics, as well as social robotics.

In the last few years, the incorporation of agent-based simulation to the study of social phenomena has resulted in a revolution in the social sciences. In Cristiano Castelfranchi's invited talk at AAMAS 2013, he said "*simulation is so important and crucial because it finally provides to the social sciences a truly experimental method, for the validation and adjustment of the models and in particular of working architectures not simply formal*". The incorporation of formalisation processes to social and humanistic disciplines is resulting in narrowing the gap between "research tradition" and will create a new research culture aligned with new methods fostered by the European Research Area, such as quali-quantitative research approaches. Agent-based modeling is the keystone for the creation of what have been called the computational social sciences (computational economics, computational sociology, computational history, etc.). It is a revolution similar to the birth of the cognitive sciences, due to the impact of information and computer sciences on the human sciences: psychology, linguistics, logics, philosophy of mind. Within the humanities, agent-based modelling has been, in particular, applied to archaeological

TABLE 1–The interplay between the MAS field and other fields.

Distributed computing	computation by agreement; agent-based system architectures; crowd-based computing; market inspired process management; agent-based resource optimisation management
Distributed decision-making	action organisation; commitments and collective agreement; group deliberation; action entitlements and policies; task and resource allocation; coordination languages; conceptual and implementation platforms for socio-cognitive systems; crowd based action
Organisation/institutional theories	organisational structures, roles, goals, incen tives, resources, tools; organisation-based programming; transaction costs; institutional meta-models and platforms; norm-aware agents; institutional constraints and incentives; anchoring on-line institutions
Natural language processing	responsible MAS; formalisation of norms; argumentation and argument extraction/identification
Social Psychology	theory of mind; reputation and trust; attribution; attention; collective decision making; values; augmented reality; team formation
Sociology	collective epistemology; collective action; group behaviour; social influencing; discrimination and social specialisation; agent-based modelling of second order social phenomena; socio-cognitive technical systems
Archaeology	agent-based modelling; agent-based simulation
Political science	game theory, e-government; e-democracy; collective political action; crowd behaviour; political rhetoric; agent-based policy-support systems; policy means and ends; collective stake-holder modelling; institutional agency; policy assessments; stakeholders' motivations and values; agent-based simulation
Governance of hybrid (artificial and natural agents) populations	norms; norm-based programming; institutional constraints and anchoring; norm enforcement in hybrid population; self-enforced regulations
Law	normative systems; agreement technologies (illegal agreements); artificial agency; artificial personality; accountability; blame allocation; reparation and punishment; and in general, robot laws
Ethics	responsible MAS and the value alignment problem; normative systems (norms for enforcing ethical requirements); agreement technologies (wrongful agreements); trust and reputation
Economics	game theory; mechanism design; market-inspired programming; organisational theory; experimental economics

studies. It has been already recognised that the use of modelling in Archaeology is providing a new paradigm of research that will soon have the same impact in the discipline as the radiocarbon revolution did in the '50s (Kristiansen, 2014; Cegielski and Rogers, 2016).

Last, but not least, the latest work on responsible MAS has been having profound impact on fields like ethics, law, and the social sciences, along with other technologies like natural language processing. In natural language processing, there is a pressing demand for online extraction of norms and arguments. In ethics, a new interest is emerging to assess the boundaries and limitations

of using norms to enforce ethical requirements (an issue being studied by the ongoing European funded WeNet project). In both the ethics and law fields, there is an interest in assessing "wrongful" agreements (agreements that are unethical or illegal) and how to deal with them. Finally, the need to understand the interplay between ethical requirements, legal requirements, and social requirements and their evolution is paramount for fulfilling the vision of the proposed roadmap for responsible MAS that evolves with people's evolving needs and values.

In summary, and from a wider perspective, *the interplay between MAS technologies and specific social sciences has had a bidirectional influence,* which, in some specific topics may start from the borrowing of terminology and intuitions but in many cases it grows into a rich cross-pollination of disciplines. Table 1 presents only a sample overview of this interplay.

3.2. Application Areas

A comprehensive study by (Müller and Fischer, 2014) in 2014 illustrated that the impact of MAS is evident in at least 22 sectors, with 11 of those covering 86% of all applications, whereas the top six cover 59% of all applications. Those top six sectors were *logistics and manufacturing*, *aerospace*, *energy*, *defense*, *security and surveillance*, and *telecommunications*, followed by the following five sectors: *business process and IT management*, *e-commerce*, *healthcare*, *traffic and mobility*, and *robotics*. This is despite the fact that MAS is a relatively young field—only two to three decades old—when it is argued that the impact of technologies and their prototypes might take at least two decades to fully manifest (Osterweil et al., 2008).

We divide the application areas of MAS according to the class of problems that the MAS technologies address: coordination and collaboration, agreement and decision making, and agent-based simulations.

- *Impact of Coordination and Collaboration Technologies.* The work on coordinating agent interactions, followed by the work on norms and normative MAS systems has left an impact on *legal systems*, *e-governance*, and other virtual communities where coordinating interactions is key, like *gaming*, *security and secure knowledge management systems*, and *e-health systems* (e.g. each stakeholder would have to follow different regulations and each clinic would have to follow their internal regulations in addition to governmental regulations, and so on).

The proposed developments in collaboration and competition can be used in *social robotics* to facilitate the development of communities of robots interacting with communities of persons, *national security* and helping fight crime and terrorism and cybersecurity (to facility security resource allocation), *business competition* in problem areas such as bidding in auctions or marketing, and *autonomous vehicle* (AV) management to facilitate the interaction between groups of AVs and groups of standard vehicles.

The combinatorial optimisation algorithms have a wide array of applications that are ubiquitous in modern society and arise in *networking*, *manufacturing*, *transportation*, *distribution*, *reservation systems*, and *emergency response systems* (Van Hentenryck and Bent, 2009). Team formation is a particular instance of a combinatorial optimisation problem, and they have much potential for application in a wide range of emerging application domains in *education* (collaborative learning), *crowdsourcing*, *citizen science*, and *social innovation* (to compose teams that solve challenges), and for collaborative work in industry (allocating and scheduling teams to projects).

- *Impact of Agreement Technologies.* Technologies like argumentation have had a direct impact on domains such as *law* and *medicine* where other approaches failed to capture the richness of reasoning. Argumentation has been applied to *conflict resolution* and *decision making* in general, joint *deliberation* and *e-governance*, as well as giving *structure to online information* (e.g. the Argument Web).

  Many real-world problems can also be modelled as negotiation problems. In the *sharing economy*, automated negotiations can create many great new opportunities for companies like Uber and Airbnb to extend and improve their services, allowing clients to interact with many drivers or apartment hosts at the same time to discover the best possible deals. *Self-driving cars* can use negotiation mechanisms to come to mutual agreements about who will cross an intersection first, or who will enter a road first. *Travel agencies* can offer tailor-made holiday packages, by allowing the client to negotiate every detail of the package, such as train bus and plane-tickets, hotel room reservations, and tickets to activities and excursions. Other areas of potential impact include: *cloud service providers, crowdfunding and crowdsourcing, logistics, insurance companies, retailing and auction sites, telecommunication providers, and time tabling scheduling*.

Trust and reputation models along with recommender systems were key to the success of many *e-commerce* companies, such as Amazon, eBay, and Netflix, and are critical for the success of distributed systems like the *internet of things* and *social robotics*. Whereas computational social choice has had a profound impact on *voting*, *resource allocation*, *self-governance*, *web page ranking* and the *fair division and allocation problems* like fair division of food donations, car sharing, doctor allocation, etc.

- *Impact of Simulation Technologies.* Today, agent-based simulations are being used to analyse *human mobility*, *climate change*, the *spread of epidemics*, while in business it is being used in studies of *organizational behaviour*, *team working* and *supply chain optimization*, just to name a few examples. In the area of computer-based training, agent-based modelling and simulation can help training skills in complex scenarios with multiple actors like for example those associated with emergency situations.

3. KEY CHALLENGING POINTS

The big challenge in the MAS field today is achieving self-regulating socio-technical systems where agents and humans collaborate, and take an active role in the regulation of their systems. This requires tremendous interdisciplinary work, transcending micro challenges into macro challenges, and careful ethical considerations that ensure neither the human loses control nor do they abuse their power. These challenges are discussed in what follows in further detail.

3.1. Achieving Self-Regulating Socio-Technical Systems by Bridging Micro and Macro Modelling

The MAS field has evolved over the years from systems of interacting software agents to socio-technical systems where humans and agents interact and collaborate. Today, the challenge is to develop self-regulating mechanisms that adapt to the ever evolving needs and values of their actors (be it humans or software agents) (Osman, 2018). We envision an active role for humans in the evolution of their system's behaviour. We imagine software agents helping with detecting potential problems with the interactions and suggesting changes in the system's behaviour, and humans discussing and agreeing on the best path for change. However, all of this requires: (1) much more resilient interaction and

coordination mechanisms (which is addressed in the third challenge in Section 2.3.3), and (2) a very strong multicentric and multidisciplinary collaboration.

Up until lately, AI has been building upon many different fields (and vice versa) without a true close-knit collaboration between the fields. ***We strongly believe such a collaboration could benefit all, and is key for addressing MAS's future challenges.*** For example, the technologies required to address self-regulation (and apart from the ethical challenges, which are key for the success of such systems and are addressed shortly in Section 2.3.2) are agreement technologies, learning, collaboration and coordination mechanisms, natural language processing, and semantic technologies. Each of these are well-established and mature fields. Nevertheless, their cross-fertilisation is key for addresses challenges like self-regulating MAS. For example, how can natural language processing help extract arguments from text in natural language, or describe the systems' norms to human users in a comprehensible way? How can norms agreed upon by users get automatically formalised and ready to be enforced by the system? How can the system verify users' needs are addressed and their values adhered to? ***All of this requires a novel multicentric approach***. Furthermore, the fields of ethics, law, and social sciences become critical for helping with issues like wrongful agreements (human decisions that may be un-ethical or illegal), or assessing whether norms can be used to ensure ethical requirement are fulfilled.

In summary, ***interdisciplinary research is key for moving from micro to macro challenges***. And this is evident not only for addressing self-regulation in socio-technical systems, but within the sub-fields of MAS as well. For example, we can see that although the work on negotiation is already advanced, the initial dream of building a realistic fully multimodal autonomous negotiator, dealing with other artificial agents (negotiators) and humans, is (also) yet to be achieved. The potential impact of such a negotiator could be huge as tedious and lengthy negotiation processes are costly activities in areas like commerce and industry. Addressing such a challenge requires the integration of language technologies (text and voice), semantics, preference modelling and learning (to build and learn models of negotiators), vision (for gesture analysis), and ethics (to determine the legality and value-alignment of offers).

Exploiting cross-fertilisation opportunities between MAS and organisational psychology can also help team formation algorithms by helping us understand: how to model humans in a non-simplistic manner; the human factors

that drive and help predict team formation (e.g. motivation, empathy, synergies); subjective measures of team performance; etc. Understanding social complexity and its integration within the natural domain is one of the significant challenges of current research, as there is a specific need for approaching social and historical processes based on historical and social knowledge. Understanding how individual behaviour and agency allows emerging social phenomena is paramount for pursuing a real understanding of social complexity.

3.2. Building Responsible Human-Centered Multiagent Systems

Ethical considerations are currently one of the biggest issues in AI, with the public, media, big tech organisations and AI scientists preoccupied with ethics and AI. As such, this topic has a dedicated chapter in this white book (see Chapter 6). Nevertheless, we believe that while many efforts focus on policies, education, or research culture (and rightly so), some work on a technical level is necessary. ***For this, "responsible MAS" is considered one of MAS' future challenges, with the specific goal of having human-centered MAS.*** This is especially critical today as the role of humans in MAS becomes more prevalent. For example, it is important to answer questions such as how to make sure humans retain control of the system, understand the system's decisions, or how to ensure human values are respected by the system.

Explainability is one key requirement for many MAS mechanisms, such as reasoning and decision making. For instance, when using a team formation algorithm, a teacher or project manager need to understand the output that it generates, and the general challenge would be to design methods capable of explaining the output of a combinatorial optimisation algorithm.

Human control is another key requirement in the ethical considerations put forward by the many initiatives laying out the requirements for ethical/beneficial AI. It stresses the importance of the human having control. Again, there are technical challenges here to provide this control. For example, as noted by the roadmap in (Osman, 2018), humans should be able to comprehend the norms governing their system's behaviour, as well as discuss and agree on those norms when needed, with the system automatically adapting to implement the agreed upon norms. All of this raises challenges in natural language processing, agreement technologies, and logic and reasoning. However, the criticality of the interdisciplinarity of this research cannot be stressed enough. For example, although human control is promoted, but what if humans end up agreeing on unethical or illegal norms? Or what if humans abuse their

power (say by resulting in the creation of online minorities that get marginalised)? As such, it must be stressed that technical solutions today should be *designed* in collaboration with the social sciences and humanities.

Value alignment is another of the key requirements put forward for ethical/beneficial AI that requires an AI system to be aligned with our human values. Some of the technical challenges here are: How to formally define values? How to relate values to norms? How to develop a value-driven decision making process? How can we verify the adherence to values? Can norms be used to enforce ethical requirements in a system? What if people's values contradict the law or other absolute values?

Although we talk about technical solutions here, again, we note that only a true close knit collaboration with other fields, like ethics and law, can truly address these issues.

3.3. Strengthening the Resilience of Coordination Mechanisms: Breaking the Boundaries

Coordination is key to the success of MAS, it is the glue that holds a society together. It is common for research lines addressing MAS coordination (just like other fields) to build their assumptions and work within those set boundaries. However, ***one of the main challenges of MAS coordination today is developing resilient algorithms that go beyond those boundaries to address real life requirements***. For example, there is a traditional divide in game theory between **competitive** and **cooperative** domains, which is not relevant in many application domains. Thus, unifying theories and broader operational concepts beyond those in evolutionary game theory are today necessary (Esteban et al., 2020). Standard reinforcement learning approaches assume the environment is *stationary* and only advanced ones consider unforeseen situations or even exogeneous effects (Martínez, 2017). When the agent is faced with an adversary or collaborator, who may interfere with the reward/utility the first one perceives, these approaches may become suboptimal. Thus, it is of customary importance to develop reinforcement learning algorithms that can make the supported agent learn whether another entity can be a friend, an adversary or just irrational (Gallego, Naveiro and Rios Insua, 2019).

Many algorithms are designed to work offline, to make use of available resources at their own pace. ***Making online decisions under uncertainty and time constraints represents one of the most challenging problems for robust intelligent agents.*** For example, tackling online team formation, an open

challenge that has been barely addressed by the MAS literature with the notable exception and pioneering work of Lappas et al. (Lappas, Liu, and Terzi, 2009), calls for the development of novel online stochastic combinatorial algorithms that leverage on ML techniques to predict the future (Van Hentenryck and Bent, 2009).

Another issue is designing algorithms for large-scale problems. For instance, MAS research has largely studied the coalition structure generation problem (of which team formation is an instance) and developed complete (optimal) algorithms to solve it. Nonetheless, such algorithms have been reported to have limited scalability, hence making impossible their applicability to solve actual world problems. These call for the design of novel algorithms, be them either approximate (e.g. along the lines of (Andrejczuk et al., 2019)) or hybrid (e.g. along the lines of (Blum et al., 2016; Bistaffa et al., 2019)).

CHALLENGE 3

MACHINE LEARNING AND DATA SCIENCE

Coordinators

J. J. Ramasco Sukia (IFISC)

L. Lloret Iglesias (IFCA, CSIC)

Participant researchers and centers

D. de Hertog (Avantopy)

A. Pizarro (Avantopy)

M. D. del Castillo (CAR, CSIC-UPM)

J. Villagra (CAR, CSIC-UPM)

N. Campillo (CIB, CSIC)

H. Mueller (IAE, CSIC)

L. Menéndez de la Prida (IC, CSIC)

A. Navas-Olive (IC, CSIC)

H. Domínguez (ICE, CSIC)

J. Aurentz (ICMAT, CSIC-UAM-UC3M-UCM)

F. Borondo (ICMAT, CSIC-UAM-UC3M-UCM)

M. Fontelos (ICMAT, CSIC-UAM-UC3M-UCM)

V. Gallego (ICMAT, CSIC-UAM-UC3M-UCM)

A. Kosgodagan (ICMAT, CSIC-UAM-UC3M-UCM)

R. Naveiro (ICMAT, CSIC-UAM-UC3M-UCM)

D. Ríos Insua (ICMAT, CSIC-UAM-UC3M-UCM)

I. Villanueva (ICMAT, CSIC-UAM-UC3M-UCM)

A. Villaseñor (ICTJA, CSIC)

J. Fernández Recio (ICVV, CSIC)

M. Romero Durana (ICVV, CSIC)

D. Ramiro (IEGD, CSIC)

F. Aguilar Gómez (IFCA, CSIC – UC)

J. Baño (IFCA, CSIC – UC)

J. M. Diego (IFCA, CSIC – UC)

J. M. Gutiérrez (IFCA, CSIC – UC)

I. Heredia (IFCA, CSIC – UC)

D. Herranz (IFCA, CSIC – UC)

P. Martínez (IFCA, CSIC – UC)

D. Rodríguez (IFCA, CSIC – UC)

F. Albiol (IFIC, CSIC – UV)

L. Fiorini (IFIC, CSIC – UV)

J. E. García (IFIC, CSIC – UV)

A. Oyanguren (IFIC, CSIC – UV)

R. Ruiz (IFIC, CSIC – UV)

A. Argyris (IFISC, CSIC – UIB)

P. Colet Rafecas (IFISC, CSIC – UIB)

M. Cornelles Soriano (IFISC, CSIC – UIB)

I. Fischer (IFISC, CSIC – UIB)

D. Gomila (IFISC, CSIC – UIB)

V. Martínez Eguíluz (IFISC, CSIC – UIB)

S. Meloni (IFISC, CSIC – UIB)

C. Mirasso (IFISC, CSIC – UIB)

D. García-González (IG, CSIC)

J. L. Arcos (IIIA, CSIC)

J. Cerquides (IIIA, CSIC)

J. Hernández (IIIA, CSIC)

J. Antonio (IIIA, CSIC)

M. R. García (IIM, CSIC)

L. Campillos Llanos (ILLA, CSIC)

1. EXECUTIVE SUMMARY

There is no doubt that the progressive digitalization of the world has a ground-breaking impact on every sphere of people's lives. Since the beginning of the XXI century, digital technology has permeated every aspect of modern society, becoming an integral part of our everyday lives. This brings both thrilling opportunities and new challenges for the research communities in this ever-changing and somewhat revolutionary context, as it implies shifts in established paradigms and application of completely new study approaches. However, it is not only scientists who are facing the challenge on

how to adapt to these new realities in the most efficient way: many governmental bodies across the developed countries are determinedly moving towards new analysis tools such as *machine learning* (ML) and making greater use of data, routinely collected by state departments, agencies or hospitals. On the other hand, both the public sector and the industry (e.g. telecommunication, energy or transport sector) have started to realize the urge to leverage the masses of data they possess and to master the adequate tools to exploit the knowledge behind it, causing the new field of *data science* (DS). All of this has led to a fast development of ML and DS departments, both addressing fundamental problems and developing applications in the most diverse research topics. Data collection, management and processing have undergone a revolution in the last few years. The increase in the availability of data and the growing processing capacity given by the advances in ML, have a tremendous impact in areas such as physics, social sciences medicine and economy just to mention a few. Many of the tools used and the challenges encountered share a lot of similarities even if applied to different areas of research. This chapter is born from the need to coordinate the current efforts on ML and DS, in an attempt to provide a big picture of the advantaged position of CSIC on these cutting-edge topics and to help foster the collaboration among groups that are currently working isolated. This cooperation is key, since most of the challenges that are presented in this document, can only be successfully addressed with a multidisciplinary approach.In this chapter, the focus is precisely set on these challenges and on the state of CSIC to face them. It includes a first part summarizing the impact in basic science of the current ML and DS techniques. The main key challenging points spotted during the elaboration of this document on the different topics, have been gathered in the second part of the chapter, followed by an overview of the institutes working on the areas described, motivating thus the advantage and experienced position of CSIC to tackle the aforementioned challenges.

2. INTRODUCTION AND GENERAL DESCRIPTION

The current avalanche of data comes with a cost, as it imposes very serious challenges in the way to extract valid information from them. In this chapter, the impact and challenges are organised in subsections. The first subsection includes the fundamentals of ML in an era of massive data. The next ones consider the different applications divided following the three CSIC global areas: Life, Society and Matter.

2.1. Fundamentals

Mathematical foundations. Mathematical foundations of ML and AI aim to achieve two main goals. First, mathematics may provide tools to improve our understanding of the mechanisms that make AI techniques so efficient in the tasks of *supervised, unsupervised and reinforcement learning* as well as systematize our knowledge around them. Secondly, based on this better knowledge we can improve the existing methods and create new ones (for an introduction to the area see (Marsland, 2015)).

Probabilistic graphical systems. Probabilistic models are a mathematically well founded approach for *dealing with uncertainty* (Koller and Friedman, 2009). Furthermore, probabilistic models are interpretable and can be integrated with available domain knowledge coming from experts as well as with available data.

Causal inference. Human understanding, and in particular scientific knowledge is structured by means of the definition and identification of *cause-effect relations* (Pearl, Glymour, and Jewell, 2016). Standard ML methods, mostly founded on probability theory, are not capable of ascertaining the difference between correlation and causation.

Data-driven simulations. In the context of socioeconomic and socio-technical systems, modeling efforts have been limited to abstract configurations. It was possible to explore the relative importance of different mechanisms but the results were far from predictive. Such limitations have prevented the systematic implementation of these models in *decision and policy making,* when they have been indeed used was under the risk of producing radically flawed recommendations (Vespignani, 2009).

Deep learning. *Deep learning* (DL) is a ML sub-field that has become a vital and integral part of it. The DL approach implies that more powerful algorithms must be used, together with a whole set of new mathematical tools, to handle and extract the knowledge behind the incredible amount of data being generated nowadays coming from very different sources (LeCun, Bengio, and Hinton, 2015). One of the main advantages with respect to more traditional ML techniques, is the capacity of the DL algorithms to perform feature engineering by themselves reducing the need of human intervention in the learning process.

Recurrent neural networks. State of the art DL techniques have proven to be very powerful to tackle tasks that were deemed impossible until recently.

Nevertheless, they also exhibit certain *weaknesses* mainly in terms of required training speed and effort, and computational efficiency. Recurrent neural networks, where connections between nodes form a directed graph along temporal sequences, represent a potential solution that intends to solve some of these shortcomings (Lukoševicius and Jaeger, 2009).

2.2. Applications

Life

Neuroscience. In neuroscience, the different data classes have been traditionally explored with dedicated *quantitative approaches,* e.g. Fourier spectral analysis of electroencephalography time series or multivariate statistics for numerical and categorical data. However, the intrinsic interacting nature of the brain as a complex system suggests that, to better understand how it works, one might adopt rather *holistic approaches* (Vogt, 2018). More over, as big mapping initiatives go, there is an increasing number of open-access repositories sharing observations at different scales from genes and proteins to networks and behavior, thus merging data altogether.

3P (predictive, personalized and precision) medicine. The current trend to move from a reactive medicine to a *preventive medicine* and to consider the patients as *active actors* in the management of their health is changing the health paradigm. Moreover, the advances in medicine imply, as a side effect, that the complexity of health challenges to tackle are also rising. This increased complexity together with a higher demanding population for health care, and the population ageing in developed countries are the key items that challenge sustainability and delivery of universal health care (Pathinarupothi, Durga and Rangan, 2018). Additionally, the number of IoT biomedical devices has increased, providing an unprecedented opportunity to enhance medical studios by incorporating automated data acquisition, and provide a more sustainable evidence-based medicine and preventive medicine.

Epidemic spreading. In the last decades, the growth of human mobility at the global scale has transformed the way in which epidemic diseases unfold and increased the risk for *new pandemics,* making infectious and emergent diseases one of the most serious threats to our society. While the time-scale for new outbreaks has passed from years to weeks, the time needed for response and control has increased due to the global nature of the problem. This is particularly true for *emerging diseases* where information about new pathogens is hard to obtain. On the other side, the explosion in the availability of digital

traces of our lives is changing epidemiological research, leading to the establishment of a new field where *big data* (BD) and ML are used to inform and guide epidemiological models: *digital epidemiology.*

Drug development. The drug development process starts from existing results obtained from various sources such as *high-throughput compound* and *fragment screening, computational modelling* or *information available* in the literature. This inductive–deductive iterative process eventually leads to optimized hit and lead compounds. ML will play a major role in this field (Dobchev, Pillai and Karelson, 2014).

Structural bioinformatics. Structural bioinformatics provides methods and tools to *analyse, predict, archive* and *validate* the three-dimensional (3D) structure data of biomacromolecules such as proteins, RNA or DNA (Gu and Bourne, 2009). Even though the application of ML algorithms has been a common practice in the field, in recent years there has been a surge in the use of AI algorithms.

Food science. Food science and technology requires advances in data processing to solve the current challenges on food quality and safety. *Identifying problems (current, emergent and potential)* in the food chain and *to detect and explain certain fraudulent practices* require interpreting data from the most updated analytical tools for data-driven decisions. Many of these techniques tend to be non-targeted or they are based on spectroscopic techniques that generate a great amount of data (see, for example, (Vithu and Moses, 2016)).

Precision agriculture. Clearly falling within the *UN Objectives for Sustainable Development,* the Food and Agriculture Organization has declared 2020 the International Year of Plant Health, as a way to make people aware that plant health can help to finish hunger, reduce poverty, protect our environment, and boost economic development. The new digital technologies can enormously contribute to this field (Liakos, 2018).

Society

Natural language processing. The field of *natural language processing* (NLP) has grown exponentially in the late 2010s, as shown by the increase of contributions to high rank conferences (ACL, COLING, EMNLP, IJCAI) and the rising birth of start-ups. Although the area is not new (the Spanish Society for the Natural Language Processing was created in 1983) the ecosystem is ripe, especially from an industrial viewpoint. Technology advancements are a key factor, as the access to powerful processing tools has been democratized. Thus,

if NLP in the 70s and 80s relied on logic and symbolic approaches (e.g. rich computational grammars, thesauri, ontologies and lexicons), the 90s made way to more statistically based methods using annotated data (e.g. treebanks) and parallel corpora (see, for instance, (Collobert et al, 2011; Goldberg, 2017)).

Demographic studies. There is a large number of ongoing projects in the field of demographic studies that represent the enormous interest for both academia and industry on the topic and which involve successful collaboration in between these two. Among such projects appear the use of *mobile phone data* for mobility, urban and tourism statistics; application of BD for *highway and waterway* transport statistics; *web scraping for labour market indicators,* and many more. Additionally, within statistical agencies and academia, there is a growing interest on the linkage of current administrative data from very different sources and the creation of long longitudinal registers based digitised historical sources (for a short review see (Billari and Zagheni, 2017)).

City science and mobility. Over 55% of the global population is living today in cities, a fraction that is expected to further increase in the coming decades. Such overpopulation put under strong stress many of the services that a city must provide to facilitate citizens life: from *resources like food, water or energy* to *efficient transportation, health care, environmentally friendly waste disposal* at the same time that it grants *clean environment (air, noise, etc.)* and *a secure and friendly social context.* In all these aspects, urban authorities must implement policies that in many cases are based on previous practices but whose efficiency lacks otherwise empirical evidence. The advent of the BD era is modifying this picture (Batty, 2018).

Autonomous vehicles. The biggest challenge to massively deploy autonomous vehicles is to find solutions that address *variable uncertainty in dynamic complex environments* (Litman, 2017). In fact, the acceptance of future mobility concepts related to autonomous driving will depend on the confidence in their behavior that the new generation of intelligent vehicles can transmit to passengers.

Economy. ML both supervised and unsupervised, provides a huge potential for social science research. In economics and political science, *the use of unstructured data (social media networks, text and images)* has led to a boom in research directly related to the EU challenge areas. In economics, the areas of *conflict studies, development economics, labour economics, industrial economics, migration and political economy* are covering aspects of these

challenges directly (related to banks and risks see, for instance, (Vitali, Glattfelder and Battiston, 2011)).

Collective behavior & fake news. The flow of information has attained unprecedented levels today. Instead of a few newspapers, the circumstantial radio or TV channel, citizens receive inputs through mobile devices at high frequency. This phenomenon has, of course, an impact on issues such as *opinion formation,* the dynamics towards *consensus* in controversial topics or, just the contrary, *the divergence of opinions* leading to polarization and the creation of *ideological ghettos* or *echo chambers.*

Electric energy, renewable sources & power (smart) grids. The power grid is one of the most important infrastructures of modern society. It is composed of power plants, both conventional and renewable, transmission and distribution lines and consumers and it is managed in a centralized way by the system operator, which adapts in real time the energy generation to the demand. The progressive penetration of *renewable energy sources* (RES), intrinsically dependent on weather variability, together with the fluctuating demand, is increasing the difficulty in *balancing production and demand.* Moreover, renewable sources can be distributed over the grid reducing energy transport needs, and therefore losses, but hindering centralized control by the system operator.

Matter

Climate and weather. Earth observations and climate models are key resources for studying *the evolution of the climate system at different temporal scales,* from meteorological forecasting to multi-decadal climate change projections. These activities have important societal applications, such as alerting of severe weather (in the short range, a few days in advance) or analyzing the potential risks of climate change in different socioeconomic sectors (in the long range). ML and DS techniques have already proven a great impact in this field (Lakshmanan et al., 2015).

Earth science. ML use is growing rapidly in geosciences because of recent advances in DL and the development of new powerful ML tools and the increased availability of open datasets (Bergen et al., 2019).

Particle physics. AI and in particular ML were protagonists of an explosion of applications in particle physics during the last decade, and the state of the art in several areas of particle physics is already represented by ML techniques (Radovic et al., 2018). Identification of *high energy physics events and particles,* as

well as *energy scale calibration* and *noise suppression* are examples of applications where AI has replaced successfully traditional techniques in particle physics nowadays.

Astrophysics and cosmology. Astronomical and cosmological observations show that atomic matter (e.g., gas, stars, galaxies) accounts for only 5% of the mass and energy in the Universe, while 95% remain unknown. It is one of the central goals of fundamental physics and astronomy to uncover the nature of these unknown forms of matter (called *dark matter,* about 25%) and energy (*dark energy,* about 70%). The nature of this mysterious form of matter is expected to become clear in the near future thanks to existing and upcoming experiments. ML will play a major role in this field (Ball and Brunner, 2010).

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

As it was already mentioned in the introduction, the way in which this section is organised, allows us to address interdisciplinary research, initiatives of excellence and knowledge transfer, and tackle frontier issues in virtually all areas of knowledge. The great amount of applications that expect major advances by the use of both DS and ML techniques, is an illustration of the impact of these tools in the development of all areas of research.

3.1. Fundamentals

Mathematical foundations. Sparked by the success of applications, the mathematical theory (still in its infancy) is undergoing a very fast development and growth around several areas that we shall describe below. Stated in mathematical terms, the problem of *supervised learning* is to approximate a given target function using a finite sample of function values. Then, using those sample values one minimizes some "empirical risk" over the "hypothesis space" (set of functions used to approximate the target). The problem, formulated in this way, requires to identify the right function spaces, e.g. Sobolev/Besov spaces, for which the direct and inverse approximation theorems are valid; i.e., a function is in a certain function space if and only if it can be approximated by the given approximation scheme with a specified order of accuracy provided by the corresponding error estimates. Traditional approximations methods suffer from the so-called "curse of dimensionality",which is the fact that the number of data needed to keep errors small grows exponentially fast with the dimension of the data set. DL strategies, on the other hand, have proven

impervious to the curse of dimensionality. The reason why this happens represents a major mathematical open problem. The *manifold hypothesis* suggests that high dimensional data actually live on a lower dimensional manifold that DL algorithms manage to naturally uncover. Other paradigms of ML, such as *unsupervised and reinforcement learning* are also, perhaps even more, in need of refinements of the underlying mathematical foundations for instance by establishing formal guarantees to the quality of the learned representations in the deep setting. *Optimization* methods for ML and AI would benefit from additional developments. The recovery of the classic *Robbins-Monro stochastic gradient descent* have to, a large extent, enabled the development of DL models based on the mini-batch setting. The recent celebrated Adam optimizer has enhanced performance recovering ideas from nonlinear optimization of the 70's and there is still much space to rediscover and incorporate new ideas. The above presumes a maximizing likelihood plus a regularizer approach to ML. Probabilistic ML is emerging as a convenient unifying framework bridging the gap between ML models and *Bayesian statistics,* with the benefits that this methodology entails through properly quantifying uncertainties, inherent regularization, taking profit of prior knowledge, some security against attacks and coherent embedding in decision making structures. Yet there is main crux of Bayesian approaches due to the intractability of integrals over very large dimensional spaces: the MCMC methods that made Bayes so popular about twenty years ago do not scale well in the BD regime of the last ten years. *Variational Bayes* (VB) techniques are much more efficient computationally yet they are biased and underestimate uncertainty. Drawing on recent developments like stochastic gradient MCMC and recent hybrids between MCMC and VB, there is an urgent need to develop efficient large scale computational Bayes methods.

Probabilistic graphical systems. Despite all these benefits probabilistic approaches, both Bayesian and frequentist, bring about some difficulties, being the more challenging the fact that probabilistic inference is *NP complete* for model classes which are minimally complex. Probabilistic graphical models arise as a way to solve this problem by represent ing the statistical independence relations that hold in our model by means of a graph.

Causal inference. There is a growing need for methods, which can make joint use both of the available knowledge and of the available data to automatically enact causal relationships. During the last years, the interest of causal inference has grown and the scientific work has crystallized in two different

theoretical frameworks: that of *potential outcomes* (based on the work of Donald Rubin) and that of *causal diagrams* (based on the work of the recent Turing award winner Judea Pearl). Developing a deep understanding of the causal inference process will have a large impact on the understandability and actionability of AI systems.

Data-driven simulations. The advent of the present data era, characterized by the massive collection of information on how people interact between them, with the economic system and with the environment, is bringing a revolution to modeling human systems. For the first time, it has been possible in the last five years to obtain trustable model predictions for the spreading of emerging diseases at a global scale, for the spreading of delays in the air transportation network of Europe and the US, for the evaluation of risks and how they transfer in the bank relation network, just to name a few, where model predictions adjust well to the empirically observed patterns. The change is of such magnitude that prominent researchers in the field have proposed a name for the new discipline: *computational social science.* The ingredients are simple, namely, agent-based modeling as described in the challenge 11b, and detailed information (data) on the system structure and functioning. This is basic for obtaining model results that can really be validated against empirical observations. For example, in the case of epidemic spreading the ingredients are population levels by age, gender, etc., mobility patterns, social interactions and the basic mechanisms of the disease progress in a single individual. This is enough to build an agent-based model with potential to assess the capacity of contention measures. Despite the large advances of the last few years, important challenges remain both at application and at general levels.

Deep learning. DL algorithms can extract information from unstructured data such as images or text in a way that was unthinkable a decade ago. The development of these techniques is still ongoing, giving impressive results in very different research areas, but for the moment, all the greatest advancements have been achieved thanks to either *supervised* or *reinforcement learning.* One of the challenges for the next decades is to give a quantitative leap forward by developing more efficient algorithms capable of learning in an *unsupervised manner.* Improving the Bayesian approaches in deep models is also one of the main bones to be tackled.

Recurrent neural networks. One of the main limitations of current intelligent systems is that they are usually built such that software and hardware are developed independently. This approach certainly leads to inefficiencies in

power budgeting, computational power, or processing speed. We envision that there is plenty of room for novel concepts in ML that better exploit the properties of *hybrid systems,* combining both hardware and software approaches.

3.2. Applications

Life

Neuroscience. Nowadays, there is no question that ML and DS approaches are the tools of excellence to process and analyze such an amount of data. More importantly, while neuroscience has originally inspired the design and operation of the artificial neural networks used in AI, we are now taking advantage in return by using AI to develop and test ideas about how the brain operates. Thus, the tandem AI-neuroscience is more circular than ever. Applying ML and DS to neuroscience will have an enormous impact at basic science level and potential applications:

- To better interrogate *multimodal neuroscientific big datasets,* from individual labs to multi-center and open-access repositories, with transformative computational tools.
- To develop novel analytical tools to aggregate *large-scale multimodal non-stationary and noisy datasets* characteristics of brain activity.
- To help in testing and developing *novel models of behavior and cognition* that subsequently inspire next-generation ML algorithms.
- Another important topic where the use of advanced data analysis techniques will be of paramount importance is building *brain's functional and structural models* of cerebral palsy, acquired brain damage, Parkinson's disease or essential tremor diseases, just to name a few, for characterizing, diagnosing and predicting them or building brain-computer interfaces for translating human intention in communication with the environment, device control or neurorehabilitation.

3P (predictive, personalized and precision) medicine. Nowadays modern perspectives of medicine are *predictive* (in terms of providing better understanding of the outputs of medical devices, and analysis), *precision* medicine by incorporating the lessons learned from drugs medical evidences, and *personalized* medicine, by incorporating specific environmental data from the patient. This is what is called 3P medicine. An increased twist is the so called participative medicine, involving challenges such as the continuous tracking, better information to the patient and better tracking of periodical diseases such as seasonal flu. ML is envisioned as the best tool to provide solutions to these challenges.

Epidemic spreading. Detailed human mobility data at both global and local scale are allowing the creation of integrated models to predict the unfolding of new outbreaks (Pastor-Satorras et al., 2015; Salathe et al. 2012). *Social media posts* and *search engines queries* are used to track current outbreaks and seasonal diseases like influenza, while *participative platforms* where users can report symptoms are becoming a central tool for *disease surveillance.* However, this revolution also poses new challenges. Current modelling approaches are limited by the high variability of human behavior and data availability, strongly limiting the validity of epidemic forecasts to a horizon of 2 to 4 weeks; too few for efficient policy making. Moreover, as heterogeneous sources of non-clinical data (i.e. human mobility, mobile phone records, social media posts, etc.) are being used for epidemiological purposes, a major challenge is to assure representativeness and fairness in covering the entire population. Finally, the incredible amount of digital records also demands new information mining and clustering techniques to extract relevant information to be integrated into conventional data sources.

Drug development. The AI in the drug development process is being applied in various stages *to predict target's role* in different diseases, *identify novel target* as prediction of SAR/ADMET properties, or even *during pre- and clinical phase,* for instance for the selection of patient population in clinical trials or during the pharmacovigilance phases. Currently, there are no drugs developed using AI methods available in the market, but the use of AI has already drastically accelerated the typical path, reducing from years to just some months, the time needed to enter clinical trials.

Structural bioinformatics. Currently, *deep neural networks* (DNN) and NLP solutions that have probed successful in the computer vision and speech recognition fields are being adapted to solve a wider range of challenges in the field of structural bioinformatics such as *protein structure prediction, protein-protein and protein-ligand modeling, binding affinity prediction, etc.* A major break-through in the field of protein structure prediction is the new AlphaFold program, using DL-based prediction of residue distances for the generation of an accurate potential for proteins, who has beaten all other methods in the recent blind assessment experiment CASP13. Advances in other areas would have a profound impact both in basic and applied science and could lead to the development of new drugs and therapeutic alternatives.

Food science. The perfection of advanced data processing is necessary to (1) solve complex problems such as sophisticated adulteration that are difficult to

detect; (2) to extract the information in such a way that it can be applied in food quality and safety assurance integrated in early warning systems; (3) to gain knowledge in the interpretation of the effect of chemical compounds on human sensory perception and to understand the interface between chemical composition of foods and neural responses responsible for consumers' attitudes (4) to promote self-sustained system of data sharing between different agents to implement efficient data treatment in a daily use practice so promoting the interoperability of food actors; (5) to identify pattern of changes in food composition that can be associated to climate change and geographical traceability; (6) to establish intelligent platforms of functional databases with fusion of data from different nature (chemical, physicochemical and market data) for anticipation and rapid reaction to food quality/safety alerts; (7) to develop adapted validation schemes for the qualitative/quantitative methods including complex data processing to facilitate their recognition and their use and to permit meta-analysis and the establishment of reference values.

Food science applications should redefine (1) the theory of identification, optimization and control of complex dynamic systems, subject to uncertain and varying systems (2) the foundations of the predictive microbiology and (3) the incorporation of the principles behind the 4.0 industry (or the internet of things) to the food industry. In addition, food science would require and impact the development of the theory to manage bacterial resistance to disinfectants, antibiotics and food preservation techniques: a major health threat expected to increase considerably in the following 50 years.

Precision agriculture. AI and particularly DL helped by *remote sensing* can help a lot in this task, allowing to build a BD platform suitable for the characterization and management of crops and to detect crop diseases from air and satellite images.

Society

Natural language processing. Current trends in NLP move towards data-intensive methods such as ML (which rely on *feature engineering*), DL and *embedding-based representations.* There is a move to knowledge inference; the explicit representation or formalization of linguistic information is underrepresented. Because language pervades most forms of human interactions, NLP is necessarily heterogeneous and promotes multidisciplinarity. A brief panorama includes *text-based applications* (morphosyntactic analysis, parsing, semantic role-labelling, spelling correction, natural language generation)

and *support to speech processing* (automatic speech recognition, audio indexing, or text-to-speech). Multicomponent technologies are used in information retrieval, machine translation, text simplification, text and visual question-answering, or dialogue systems (chatbots). NLP systems spread from educational applications (e.g. computer-assisted pronunciation teaching or intelligent tutoring systems), automatic sentiment and opinion analysis (e.g. product reviews in blogs or Web fora), text mining for domain-specific texts (finance, legal or administrative documentation, or biomedical natural language processing) and detection of plagiarism, cybercrime, hate speech and theft of digital identities. Results from the NLP field have enriched close areas in linguistics, cognitive modeling, lexicography and digital humanities.

Demographic studies. The avalanche of new data on this topic can be seen as a genuine golden mine for both academic fields, industry and government, which will provide *deeper, richer and timely insights* into the essential demographic motions and societal changes when tackled with the appropriate tools and methods. Relying on the accumulated experiences and lessons learnt over the past 20 years since the start of the digital era, it seems to be feasible and evidently necessary to build a thorough framework to study the processes of human fertility, mortality and migration through continuously generated BD, involving the latest technological and methodological achievements like AI and ML.

City science and mobility. The abundance of data that underlies the concept of smart cities with ubiquitous sensors allows for *a more scientific approach to urban management decision taking.* This must necessarily pass through the use of the rich information obtained from the field as input for data-driven models able to assess the effect of policies before their implementation. Examples are exploring the impacts of new transportation infrastructure as the implementation of a new metro line, new configuration for the bus lines, changing directions for car traffic in one-way streets, etc. In some cities, there is already incipient modeling efforts to understand the policy impacts on transportation.

Still, the approach must be further developed and the scope needs to be extended to other issues beyond mobility and transport as, for instance, health, pollution, economic inequality, security, etc.

Autonomous vehicles. Artificial embodied decision-making aspires to respond to the current challenges on this topic and is articulated around three fundamental mechanisms: *the interpretation of the scene, the motion*

prediction of the agents involved in the driving scene and *the application of behavioral planning.* All of them have a critical transversal element to manage a great variability of situations: embedded explainable ML. Furthermore, from a system perspective, the advent of autonomous vehicles in transportation is further encouraging the study of ridesharing solutions, such as (electric) *shared autonomous vehicles* (e-SAVs). This calls for the design of novel ridesharing algorithms for e-SAVs that consider both physical constraints (e.g., where and how to charge vehicles) and passengers' preferences. Ridesharing solutions are typically conceived as on-line optimization. Algorithms supporting such optimization might potentially benefit from the automatic learning and prediction capabilities of ML techniques to take more informed decisions. The major challenge, therefore, is to put ML methods to the service of optimization (ridesharing) algorithms.

Economy. The adoption of new methods is happening fast. Many of the changes have been driven by the increasing level of programming skills in economics and political science graduates and the development of DS graduate programs which combine knowledge from both the social and DS. The possibilities for further growth in this area with an expert centre on issues ranging from *political conflict monitoring, refugee flow forecasts* or *political risk evaluations* are enormous.

Collective behavior & fake news. The current phenomenon of fake news has been around in the topic for many years under the name of rumor propagation. The general problem with geopolitical fake news is that there are few real time mechanisms to deal with such challenges, basically because even though the creators of such strategies are people, they are frequently performed through automatic systems against which traditional response mechanisms are essentially useless. As today, the detection is based on singling out malicious spreading strategies but not so much on the content itself. Another challenge to be faced in the fake news detection area, is the recent interest in text-generating AI systems, raised partially due to the creation of the GPT-2 system by OpenAI. This fact manifests the urgent need to introduce *real time* and *automatic fake news detection methods.*

Electric energy, renewable sources & power (smart) grids. A control distributed over the grid would maximize the use of renewable sources at the same time that would increase the efficiency of the system. There are several proposals to achieve such decentralized control, going from virtual power plants that integrate a large number of small renewable sources distributed over a

portion of the grid, to the use of smart devices that autonomously adapt their operation to the needs of the system, what is known as *dynamic demand control.* On the other side, the generalization of electric vehicles will introduce a stress on the demand. The main issue is to avoid the peaks in demand due to the synchronized recharge of these devices, besides the batteries can be used as stabilizers for the system recharging or injecting power as needed. In fact, the massive use of smart devices taking autonomous decisions would convert the power grid in a truly complex system composed of many interacting units. Smart devices and smart meters provide huge quantities of information in real time than can be used to understand and control the system. In this context, the use of BD and ML techniques to simulate the system, train the behavior of its agents, and predict its dynamics will then be essential to achieve a stable and efficient operation of a future CO_2 emissions-free power grid.

Matter

Climate and weather. Comprehensive long-range climate simulations are obtained by running global and regional climate models (driven by different future emission scenarios) in the framework of international coordinated efforts such as the CMIP (global) and CORDEX (regional) initiatives, producing huge amounts of data characterizing the future state of the climate system. Earth observations (of essential climate variables) and climate simulations result in a huge volume of available data (around 350 petabytes estimated for the above initiatives). Thus, there is an increasing interest *to exploit these vast sources with AI techniques* and extract knowledge about the climate system. In the last years, there have been notoriously advances in ML in this field, mainly boosted by DL. Some first applications have already tested the suitability of ML in diverse climate applications. Some examples are the tracking and intensity estimation of hurricanes, the detection of extreme weather patterns, the emulation of model components (such as subgrid parameterizations), or downscaling model outputs to regional/local resolution. However the possible benefits of ML in climate are not limited to the latter and research in this field is an area of growing interest.

Earth science. In geosciences, ML use can be classified into three interconnected categories: *automation* (e.g. labeling data when the task is difficult or time-consuming for humans), *inverse/optimization problems,* and *discovery* (extract new patterns, structure, and relationships from data) as for example with the satellite imagery reconstruction at super-resolution and their posterior treatment. In spite of the availability of large datasets from Earth and

ocean observing systems, often extending over long observation times, many of them remain largely unexplored. A wider adoption by the community of open-science principles such as open source code, open data, and open access would allow to take advantage of the rapid developments in ML and AI taking place.

Particle physics. The next decades will be characterized by new challenges both in quantitative and qualitative terms. New experimental projects, such as the *High Luminosity Large Hadron Collider,* are characterized by unprecedented size and amount of events, data rates and complexity. The success of this new generation of experiments will depend on our capability to develop effective algorithms and optimal usage of computational resources. Only a sagacious usage of AI techniques will allow to exploit the full potential of future particle physics experiments.

Astrophysics and cosmology. Several strategies can be adopted to look for *dark matter,* all of them requiring realistic physical simulations which are very time consuming and the development of techniques to handle enormous amounts of data in the search for new physics as data anomalies. This connects the scientific problem to advanced DS and ML. One of the strongest evidences for dark matter comes from gravitational lensing. Galaxy clusters and galaxies contain large amounts of dark matter.

Regarding *the classification of galaxies,* they have traditionally been classified by visual inspection. However, this is a very time consuming approach, prohibitively expensive given the large amount of data which compel to use automated algorithms. DL models have demonstrated to be very successful for morphological classification of galaxies, reaching or even improving, human performance. Other useful applications are e.g., the identification of galaxy lenses, mergers and tidal streams or anomaly detection.

4. KEY CHALLENGING POINTS

Each of the different global research areas mentioned in the previous section, presents a series of long term ML and DS related challenges to effectively gain new knowledge and, in turn, to make a positive impact on applications for society. These key challenging points generally require the participation of multidisciplinary research groups, so that the challenge can be addressed with a global perspective and with the necessary instrumental and computational techniques.

In this section, the general challenges related with the use of ML and DS techniques and crosscutting to all the aforementioned applications are summarized first. Then, more high level challenges are described by the global research area:

4.1. Fundamentals and Theoretical Challenges

Machine learning

- **Interpretability and explainability.** Within this label coexist many ideas and methods, some of them such as DL, or its relatives under the name of recursive networks and reservoir computing, are the most successful in terms of applications but have the major drawback of not providing insights on their behavior and decision making processes. This challenge considers ***the development of a theoretical and practical framework for coping with uncertainty in ML.*** Very often ***interpretability and explainability*** are intermediate steps for achieving further goals such as ***fairness or unbiasedness,*** but they are also of ***paramount importance standalone*** since they can potentially unveil new insights on how the systems work and which mechanisms must be taken into account in mechanistic models. Issues mentioned such as causal inference can bring advances in this direction but this is the most relevant challenge in the area for the next decade.
- **Coping with uncertainty, stability and robustness.** ML models need data for training or, as in the case of data-driven modeling, as basic inputs for the simulations.In all cases, ***data may come with biases, errors and in partial sets.*** This is most common out of the laboratory, in observational fields, but it can happen as well due to unexplored control parameter regions in the lab. ***The outcome of these models can be affected by such uncertainty sources in undetermined ways,*** which can later propagate across the models. Changes in the analyzed system situation may require a prompt reaction for the tool to be robust. It becomes thus crucial to develop theories and suppression measures to avoid flawed solutions.
- **Fairness.** This topic deals with ***understanding and mitigating ML*** algorithms taking discriminatory decisions based on sensitive characteristics, such as race, religion, gender, physical ability, sexual orientation, etc.

Data science

- **Representativity: correcting biases.** In many applications, there exists a fraction of data that gets missed. There are even tools to recover them assuming that the statistics of the missing data is similar to the one observed. However, this assumption not always holds and the data, which are going to be the basis for modeling, may present strong biases. In physical systems, it may be relatively uncommon even though there can be systematic errors. In biological and social systems, on the other hand, data representativity is one of the major sources for model uncertainty. Developing methods ***to measure the representativity, to correct biases and to minimize their impact on the final model outcomes*** is a substantial challenge to be solved in the future.
- **Accessibility to the data.** The creation of new tools, to improve the availability of reliable and accessible data sources, is fundamental for the development of the field. Data in many fields can be a property, either belonging to the scientists that performed the experiments, to the companies providing a service, to agencies that make the observations possible (satellites, detectors, etc), there are different circumstances. ***A balance between the natural right to exploit the data for obtaining results of the owners and the access of the general public or scientific community must be searched.*** An example is mobile records from the population, this data is of paramount importance for applications such as computational epidemiology or urban and infrastructure planning. Still most of the data are owned by technological companies that provide a service to their users and must protect their privacy. Finding an equilibrium between these antagonist interests is an important challenge, which may have a technical aspect as well as a legal one. The solution could pass through the development of new technical solutions able to satisfy the diverse interests.

4.2. Main Applied Challenges

- **Communication with the machines using natural language.** From a theoretical point of view, the main issue in NLP deals with processing and formalizing linguistic phenomena beyond the limits of the sentence (ambiguous co-reference analysis, and discontinuous features such as out-of-context dialogic turns) and also within the sentence, at the word or phrase level in the realm of creative language (figurative language, euphemism, metaphor, certainty and speculation, humour or irony).

From the perspective of resources, promoting NLP for under resourced languages and documenting less studied varieties (e.g. American varieties of Spanish) is critical. Recent DL algorithms (e.g. BERT, XLNet, ELMO) need exhaustive evaluations by both ML researchers and computational linguists to explore the semantic modelling of words by means of semi-supervised or unsupervised methods. There is an increase ***in processing multimodal corpora —with visual, text and speech information for human-computer interfaces— and a standing growth in tasks in specialized domains*** such as medicine or e-commerce or law.

- **Describing how the brain handles uncertainty.** In neuroscience, the challenge consists in representing and modeling how biological brains are able to cope with uncertainty, beliefs and generalization. Research in this direction is especially fruitful ***provided neuroscientists, data and AI experts work together to learn from each other.*** As has already happened before, in the process of characterizing the brain new algorithms for AI are likely to emerge.
- **Reducing cycle times for drug development.** One of the most important challenges is the reduction of the time and the cost to take a new drug to a clinical trial. AI systems should play a major role ***allowing for a selection of the most promising substances*** and deducing the effects out of testing subjects monitoring information.
- **Personalized, preventive and high precision diagnosis of individual health risks: 3P medicine.** The challenge is to be able ***to connect genetic anomalies or peculiarities with the risk of development of illnesses*** such as cancer and with the potential response to the available treatments. Naturally, this is a data rich environment where correlations are in some cases difficult to get established. DS and AI can provide tools essential to attain this goal.
- **Forecasting the propagation patterns of emerging diseases.** As with the weather with certain time in advance, the objective is to be able ***to track in real time and predict the evolution of epidemics and pandemics, as well as to evaluate measures and scenarios and assess the risks for public health.*** The prediction and control of new outbreaks and emerging diseases requires the combination of data and modeling. ***As demonstrated by recent outbreaks*** — i.e. the SARS in 2002, H1N1 flu of 2009, the MERS in 2012 and the 2019 Wuhan Coronavirus— ***emerging diseases pose an increasing threat to our society.*** Their containment is particularly challenging due to lack of data and infrastructures for

identification. Computational epidemiology has a major role to play in shedding light on the situation and helping public health authorities to take informed decisions.

- **Understanding conformational dynamics of proteins and other biomolecules.** Proteins are dynamic entities, and large conformational changes are often key for their function. Molecular dynamics and elastic network models are currently used to model conformational movements of biomolecules, but ***new computational approaches are required to cope with large systems or biological time scales*** (over ms). An additional challenge is ***to structural modeling of all protein assemblies in the cell.*** While experimental determination of the structure of biomolecules is rapidly expanding, structural data are available for only a small fraction of all possible complexes in human and other organisms. Structural bioinformatics approaches such as template-based modeling, ab initio docking and integrative modeling aim to complement experimental efforts, but more efficient computational approaches are needed for flexible cases, weak interactions and multi-molecular assemblies.
- **Predicting food quality and improving safety** is a critical scientific challenge for ensuring a healthy diet under, among others, fast changes in consumers' demands and patterns in food outbreaks. Inside the broad variety of systems and behaviours under this challenge, ***the emergence and spread of resistant food pathogens and spoilers require an special mention for its expected impact in the following years,*** in the context of both the human and animal health. Model-based predictions would allow to optimize and control conflicting objectives such as the maximization of consumers' demands (for example organoleptic properties or minimally processed foods) satisfying hard restrictions in food safety. Additionally, the development of advanced data processing, including ML algorithms, ***to interpret the sensory impact (aroma, taste, texture, etc.) of food composition and to extract information to be used in the control of complex authenticity issues*** (sophisticated fraudulent practices, geographical traceability, and emerging safety problems). In the case of human-food interaction concerning sensory perception, one challenge is the establishment of causal relationship of sensory active compounds with quality characteristics. The principles that govern the connections between these compounds and the final perception needs to be elucidated through several scientific areas (e.g. chemical characterization, psychophysics and cognitive neuroscience) that

require specific data studies to combine linear and non-linear information with a relevant dynamic component. Likewise, food authenticity requires establishing causal relationship of complex data (e.g. non-targeted analysis) with specific authenticity issues through data studies that are adaptable to standardization and validation.

- **Monitoring crops and real-time response in case of anomalies.** Precision agriculture deals with the presence of toxins, plagues or diseases. It requires the construction of reliable datasets suitable for following the state of agricultural crops mainly based on data collected by remote sensing (hyperspectral cameras airborne in drones or from satellites, such as the ESA Sentinels), the validation of the tools to detect abnormal situations with measurements from the fields and in robotized greenhouses at the lab (e.g., the *Personal Food Computer* (PFC) from OpenAgriculture Initiative at Media Lab, MIT). In this line, it will be important to build ***low cost multi or hyperspherical cameras with sensors based in interferometric customized filters electrochemically obtained from porous silicon,*** to be used in our PFC and/or drones. This can be used for general crop remote sensing purposes and also to be tuned to specifically detect the existence and extent of pollutants and poisons in plants.
- **Characterizing the state of the population.** The emergence of information and communication technologies has brought new data that after processing can allow to probe the population state with unprecedented immediacy and precision in classical demographics. The main challenges here are ***detecting the presence of possible biases*** (representativity of the data samples) and ***how to correct to obtain a faithful image of the society.*** AI will play a central role in all these issues. The accessibility to high quality proprietary data represent a major issue as well.
- **Improving the live quality in cities.** Regarding urban systems and mobility, the main challenge is to improve the quality of life in cities. Facing this challenge is a multifaceted and multidisciplinary task. It requires the development of reliable sensors to monitor the state of cities. In several aspects such as mobility, the arrival of the smart city concept is facilitating the advances in this direction even though some issues as economic activity and inequality are still out of the radar of these sensors. This probably will improve in the next years or decades. Once the data are available the work has just started: ML techniques must be developed to extract ***as much information as possible on the***

state of the system and from there ***data driven model frameworks must be introduced to be able to assess policies before their implementation.*** The use of these tools should then be popularized among the authorities, stakeholders and decision makers. The path is, thus, clear but the advance in these fields are necessarily slow and will take decades before attaining the full technological deployment.

- **Reliable autonomous vehicles.** Autonomous mobility as a service: learning (1) for embedded intention estimation and decision-making, and (2) to model manual automated cooperative behaviors in dense environments. Enacting shared autonomous vehicles: (1) learning individual passenger preferences via automated, non-intrusive preference elicitation; (2) learning the compatibility of a potential group of passengers in the same SAV; (3) predicting passengers' demand.
- **Towards data-driven economic models.** Two areas are particularly important: ***the integration of unstructured data into existing models and the adoption of ML to help in the identification of causal links.*** Economics, in particular, has developed a formal model language which lends itself to the integration of new tools for causal identification and unstructured data.
- **Mitigating the damage of fake news**. There are several current examples of collective decisions taken under the pressure of misinformation. ***Collecting data from the different spreading media*** (online social networks, newspapers and traditional media) ***is fundamental to characterize and identify the spreading patterns.*** Modeling and AI can help to design new strategies to reduce the impact and to explore ways to expose the population to un-skewed information.
- **Stable and efficient distribution of electricity.** Given the present configuration of the network and, especially, the expected changes such as the massive adoption of electric mobility solutions, ***this passes through the efficient analysis of massive amounts of data to be used as input*** for data driven modeling and forecasting demand variations, the establishing decentralized control protocols based on weather forecast and BD analysis to cope with fluctuations and the introduction of adaptive behavior of the power grid under system failures or RES unavailability.
- **Improving the temporal and geographical scale of the weather and climate prediction.** Multivariate calibration/downscaling of model outputs using higher resolution observational records (satellite data or in-situ observations) is a challenging applied problem in practical

sectoral applications (climate, agriculture, hydrology, health, energy, etc.). There are a number of statistical ad hoc approaches, but ML could provide ***an efficient general approach, making the best of all the available data for a particular problem*** (postprocessing model outputs). Additionally, climate models numerically solve the physical equations governing the climate system, but they also include parameterizations (empirical subcomponents) to balance and model subgrid processes. ML has been already applied to emulate those sub-components, allowing to speed up computations. Therefore, there is evidence of the potential of integrated models benefiting from both approaches.

- **Providing new information for Earth science.** In geoscience, data are acquired using a variety of methods (satellite sensors, field measurements, computer simulations) with varying spatial and temporal resolution. The most common approach to handle multi-resolution data is to upsample the datasets (e.g. using interpolation methods). New approaches are needed ***to identify patterns at different resolutions while using the uninterpolated datasets.***
- **Triggering and real time analysis of particle physics data.** The future particle physics experiments will require AI to cope with an ***unprecedented level of data processing in real time to select the events of interest among a massive data rate*** (several times larger than the bandwidth of internet giants like Facebook and Google) and ***reduce the event size by performing real time event reconstruction and store only a reduced set of key elements of the event.*** Additionally, the reconstruction of physics objects will need AI techniques different from the current ones based on pattern recognition, analytical filtering and iterative clustering, to achieve the level of performance required by the future experiments.
- **Estimation of distances in astrophysics and cosmology.** The accurate distance estimation based on photometric measurements is one of the biggest challenges. ML techniques have ***the potential to tackle this complex problem and play a central role in these surveys.*** Additionally, they can contribute as well in accelerating expensive numerical simulations like hydrodynamical simulations involved in galaxy formation and evolution and in the simulation of the dark matter structure of the Universe.

CHALLENGE 4

INTELLIGENT ROBOTICS

Coordinators
G. Alenyà (IRI, CSIC – UPC)
J. Villagra (CAR, CSIC)

Participant researchers and centers
R. Fernández (CAR, CSIC)
P. González de Santos (CAR, CSIC)
R. Haber (CAR, CSIC)
A. R. Jimenez (CAR, CSIC)
A. Ribeiro (CAR, CSIC)
E. Rocon (CAR, CSIC)
J. Borràs (IRI, CSIC – UPC)
F. Moreno (IRI, CSIC – UPC)
C. Torras (IRI, CSIC – UPC)

1. EXECUTIVE SUMMARY

Intelligent robotics are called to be the next revolution by providing AI with the capability of interacting with the physical world. Robots are overpassing their cages in the industry to become intelligent machines that can live among us, helping in the service sector, as tools in rehabilitation and assistive tasks, and also as companions.

Robotics poses especial problems and AI research must be reshaped and redefined to meet robotics special needs in areas like perception and scene understanding, decision making and learning, and actuation. Besides these classical robotics areas, modern robots need to take into account the central role of human-robot interaction: unstructured environments, unforeseen situations, user preferences, and safety.

The challenges to frame this revolution are multiple. We highlight the seven where we identify CSIC has a strategic advantage and thus can cause a better impact. Modern robotics implies robots in human environments, what we called here *robots for everyone: easy reprogramming and continuous learning*. Deployment can include big-scale mobile robots and cars for *autonomous navigation for cities*, or small-scale robots for *intelligent manipulation for new applications*, possibly making use of *effective and adaptive coordination of robot fleets*. Robots in human environments require *safe and ethical human-robot interaction*, that can take advantage of *seamless cooperative and everywhere*

localization solutions and *dexterity and efficiency through bio-inspired and parallel mechanisms*.

Advances on intelligent robotics will have a great impact on science, industry, and society in general. Robots have the potential to change people's lifestyle and thus, require special attention from rule bodies and policymakers. However, robotics is highly experimental and requires special efforts in physically building the prototypes. To make this possible, we believe a new joint lab or infrastructure must be established to facilitate research and testing, foster collaboration and involve industry and policy-makers.

2. INTRODUCTION AND GENERAL DESCRIPTION

Robotics has become one of the key drivers of the economy thanks to the maturity of the hardware and the use of powerful AI algorithms. Robotic systems must have the capability to adapt to the environment and task requirements. This entails the perception and understanding of the state of the surrounding environment, the understanding of the different actions that can be executed to achieve the high-level goal, and the actual execution of the actions (Siciliano and Khatib, 2016).

Robotics involves challenges in these three areas that are particular because of the inherent embodiment in robotics, making most of the AI developments not appropriate. General *perception* algorithms are usually not useful and must be tailored to the particularities of robotics. *Decision making* methods must take into account that the robot can, and will, interact with the environment changing the state, causing sometimes catastrophic effects. *Actuation* must be safe, but also must be adaptable and generic.

Finally, robots have left the safety of industrial cells to go to the unconstrained world, usually close to humans. Safety, human-robot interaction, explainability, and adaptability are some of the concepts that become crucial (European Union, 2014). AI provides ways of making this possible, for example, by facilitating the understanding of the surrounding environment, the task and the people, and also by providing generalization and transfer of the knowledge.

Perception

The development of perception systems is a key ingredient for building the next generation of robots. We expect contributions in several domains. (1) *Passive perception methods* for creating category and instance-level object

models. These developments can leverage recent *deep learning* (DL) approaches. (2) *Linking perception and action.* Designing new perception algorithms tailored to the action to be executed by the robot. This includes the prediction of object affordabilities and object properties beyond the geometry, such as material type, roughness, elasticity or graspability. (3) Developing *interactive perception* algorithms for capturing properties of objects that cannot be estimated passively and involve the interaction of the robot with its surrounding environment. (4) *Sensor fusion* has been a topic researched for many years in robotics. However, current characteristics of the sensors concerning their volume, weight and power consumption make it possible the integration and fusion of several sensors in most robotic platforms. This will require the development of novel data fusion algorithms. Again, DL seems an appropriate technology for such purpose. (5) *Location and context estimation.* Positioning systems for intelligent robots have experienced a significant evolution during the last years, however more research is needed in order to accurately and reliably operate both in outdoor and indoor environments. This improved performance evolution is giving rise to new location-based services and applications in a variety of fields, such as military, healthcare, tourism, city planning, agriculture, and logistics, among others.

Decision making

We expect robots to *adapt*: to the task, to the changes in the foreseen plan, to unexpected situations (Martínez, Alenyà and Torras, 2017), to the human preferences (Canal, Alenyà and Torras, 2019). They exhibit intelligent behaviour when choose at every moment the most appropriate action. This entails decision-making algorithms able to understand the task and the context, and able to reason about the consequences of the robot actions (Andriella, 2018). The solution naturally encompasses not only one but the combination of several complex AI techniques, as this is one of the most challenging aspects in robotics.

In contrast to other areas, decisions in robotics have to be made in *real-time,* taking into account the complexity of the different execution levels (*from symbolic to execution*), and considering the effect of the action in the *physical changes* produced in the environment (Savarimuthu et al., 2018). Handling complexity and uncertainty is paramount because producing real-time decisions entails sometimes focusing on partial information, on limited horizon planning, on partial learning of the involved models, on relying on a belief of the current state instead of a completely observable state.

Intelligent robots must share space and task with humans (Jevtic et al., 2019; Olivares-Alarcos et al., 2019) and possibly with other robots. This adds complexity to the robotics task as aspects as safety, explainability, adaptability, and coordination become important. Considering the human-in-the-loop of decision is a very promising and fruitful line of research, as allows to take advantage of human interactions to learn, for example, models of the task.

Thus, AI algorithms developed to learn and take decisions in intelligent robotics tasks are sophisticated and need to be specially tailored to the specifics of the problem.

Actuation

The role of control theory has been central in robotics solution over the last fifty years (Mattila et al., 2017; Tzafestas, 2018). The initial control systems were conceived under simplistic assumptions and considerations (e.g. linear systems, single-input-single-output, nonexistent or very reduced use of environmental information...). However, in most recent (systems of) robots, kinematics/dynamics are often challenging (Thuruthel et al., 2018; Cortés and Egerstedt, 2017), and requirements stringent, thus demanding *adaptable, efficient* and *robust control* solutions. To that end, novel theoretical results in nonlinear, robust, and adaptive control have been applied to specific and complex problems, which have cross-cutting connections with larger application domains, such as networked decision systems or cyber-physical systems.

One of the main challenges actuation for robots must face is its ability to *adapt to the operation context*, while remaining human-aware (safe and comfortable (Bansal and Tomlin, 2019)). In this connection, as many of the systems to be controlled will be *safety-critical*, the rising and inescapable *learning-enabled strategies* (Bing et al., 2018; Polydoros and Nalpantidis, 2017) will need be intelligently combined with *fail-operational mechanisms* (Guiochet, 2017). As a result, model-based techniques and model-free (or data-driven) control approaches will have to co-exist bringing out each advantage, namely explainability/robustness and context-dependent adaptation/learning, respectively.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Robotics has reached such maturity in the design of mechanisms and control systems that several practical applications could be now feasible. What is sometimes still missing is the intelligence that can make these robotic applications practical, adaptive, generic, safe and reliable. Aware of that, research groups have concentrated on investigating new AI techniques that can effectively bring these required characteristics to robotics. It is worth to mention again that most of the standard AI algorithms do not apply to robotics. Robotics has its own requirements in terms of the time of computation, available resources, and most importantly, safety. We have identified five of the most relevant applications of robotics with high social impact where CSIC has a strategic advantage: assistive robotics, wearables, autonomous driving, smart manufacturing, and agricultural field robotics.

3.1. Assistive Robotics

Robots at home, hospitals and daycare facilities helping people: this is one of the highest social impact applications of robotics. This is the next robotics revolution and will boost the development and mass production of robots in the next decades. We have to distinguish two main areas: *social robots* and *assistive robots*. While social robots do not require physical intervention in the world, assistive robots interact physically with the environment and are in charge of helping people performing very basic and important activities. These activities involve self-care tasks defined in the known list of *activities of daily living* (ADL) like personal hygiene, dressing or feeding, and activities to maintain individual life independence, detailed in the list of *instrumental* ADL like cleaning, moving, safety and emergency.

The adoption of robotics in these human environments is very challenging because are highly non-structured, and because people are expecting the robot to seamlessly adapt, something that is very easy for human caregivers but extremely difficult for assistive robots. We also expect robots to *use* human-made tools, and to *manipulate* deformable objects. There are numerous other challenges, like assuring safety, ethics behavior, and an efficient bi-directional communication so the robot *can understand human intentions and preferences,* but also the person can understand *robot internal states* and reasons behind a given decision. We need new AI techniques to build effective robots, capable of understanding the underlying semantics of actions, not only their objective but also the way they are performed.

As we will see, all the developed techniques to improve human-robot interaction and understanding will also influence the use of robots in industry and have fostered the development of Industry 4.0 *collaborative robotics*.

3.2. Wearable Robots

The new technological advances opened widely the robot application field. Robots are moving from the classical application scenario with structured industrial environments and tedious repetitive tasks to new application environments that require more interaction with humans. It is in this context that the concept of *wearable robots* has emerged. These robotic mechanisms have been applied in telemanipulation, man-amplifier, neuromotor control research, rehabilitation and to assist impaired human motor control. In particular, wearable robot technology is a rapidly expanding field in research and clinical applications. The use of robotic *exoskeletons* for neurorehabilitation applications has increased in the last decades, both in childhood and later life, and in several motor diseases such as stroke, spinal cord injury, cerebral palsy or Parkinson disease. This approach has interesting advantages compared to traditional therapy because robotic therapy *integrates functional tasks* with accurate and assembled movements instead of repetitive movements without a goal. In order to achieve this goal, robotics needs not only new technology but also more science. Future research in this area will take inspiration from biological models (in particular from the human model) to the design of innovative, dependable, inherently friendly and highly acceptable robotic systems.

This field is currently undergoing dramatic changes as a result of significant advances in *robotics* (sensors and actuators), *biosignal recording* techniques, and *signal processing*.

These developments will represent a game-changer in the rehab robotics field by the development of solutions that integrate *emerging technologies* (smart sensors and actuators, novel signal processing techniques, and advanced biomimetic control strategies) into *robotic platforms* to promote motor learning. These solutions should revolutionize rehabilitation of humans since devising protocols which optimize motor learning is a state-of-the-art research question that promises to deliver scientific, clinical and societal impact.

Injuries and diseases of the nervous system such as stroke, spinal cord injury, Parkinson's disease or cerebral palsy are extremely debilitating and pose a formidable socioeconomic challenge. Developing *effective treatments* that

alleviate their symptoms has a tremendous potential to improve the quality of life of many.

3.3. Autonomous Driving

Although important challenges are still ahead of us, intelligent vehicles technology is advancing at a vertiginous pace. Cars are going to be increasingly connected, automated, and shared. As a result, they will have the potential to not only radically change personal, freight, and public transportation, but also to influence cities life and growth.

Autonomous vehicles (AVs) will help to reduce the number of driving-related accidents, reducing the fatal impact of human errors, which is the cause that lies behind 95% of crashes, thus shrinking the costs of both injuries and deaths in terms of insurance, rehabilitation, or medical leaves.

AV technologies will be key, together with electromobility and sharing-oriented infrastructure, for the achievement of the *triple zero vision* (emissions, congestion and accidents). Indeed, shared AVs will drastically reduce traffic problems, and therefore the footprint per journey, as the vehicle *occupancy* will be dynamically optimized. Besides, AV can permit *emissions reductions* by enabling more-efficient driving patterns, reducing drag, and actually increasing the use of public transit by solving the first/last mile problems. These benefits are also reinforced through the envisioned V2X communication, for example in cruising towards an intersection, where the vehicle can *intelligently coordinate* with other vehicles and infrastructure to properly schedule its passage, avoiding significant waste of energy and time.

In addition to the above, the combination of AV and the sharing economy will also have positive equity impacts by increasing access to employment, education or health services, and by enabling higher degrees of social inclusion, offering broader access to discretionary travel for disabled and ever-growing elderly, thus satisfying their mobility requirements.

3.4. Smart Manufacturing: Towards Industry 5.0

Today, the manufacturing industry is aiming to improve competitiveness through the convergence with cutting-edge ICT, in order to ensure sustainable development. Smart manufacturing, which is the fourth revolution in the manufacturing industry, is an emerging form of production *integrating* manufacturing assets of today and tomorrow with sensors, computing platforms, communication technology, control, simulation, data-intensive modelling

and predictive engineering. internet-of-services and AI-based methods will be essential for qualitative change in existing complex systems. Smart manufacturing shall be able to create not only the effects in the economic indicators, such as cost-saving and productivity increase, but it must also be able to create new values that can constantly contribute to societies towards Industry 5.0 paradigm. Smart manufacturing shall not only be able to simply construct components and devices with embedded artificial cognitive functions –self-adaptive, self-resilient, self-organize, self-repair, self-maintenance, self-reconfigure, cutting-edge information technologies, but it will also be able to develop as a continuous growth engine for manufacturing with human and society oriented philosophy through 'sustainable development'. The fifth pillars of future applications require *intensive use of AI* to address the new manufacturing processes (e.g., 3D printing), *smart materials* (e.g., shape memory alloys), *predictive and proactive engineering* (e.g., digital twins), *sustainability* (e.g., eco-friendly factories) and *resources sharing and networking* (e.g., collaborative modelling and shared manufacturing). Industry 5.0 will be focused on combining human beings' creativity and smart production machines with the speed, productivity and consistency of robots. Industry 5.0 means to better appreciate the cooperation between robots, smart production machines and human beings by combining their diverging strengths, in order to create a more *inclusive* and *human-centred* future.

3.5. Robots for Agriculture

Much has been said about the desirability of introducing robots into agricultural fields; nevertheless, despite the effort made to penetrate agricultural markets and the potential benefits that robotics presents, the use of robots for agricultural tasks is far from being generalized.

Introducing robotic systems to perform agricultural tasks provides several advantages over classical methodologies: (1) An increment in the precision, repetitively and overall quality of the process. (2) An extension of the operation time. (3) It allows applying thorough and homogeneous decision base. (4) It improves interoperability and coordination and reduces unit costs.

There are two main differences between industrial and agricultural domains: (1) The operational environment of an agricultural robot is only partially structured and subject to biological variability. Due to this, agricultural robots require a *good perception* system (sensors) to be able to operate within this environment. In addition, it must allow making appropriate decisions

without relying on complete and precise information regarding its environment. (2) During their operation, agricultural robots *are not placed at a well-known, fixed location* with precise and secure (for example a tree branch can intercept in seconds the path of a manipulator) access to supplies, unlike manipulators in a production chain. Moreover, the product to be handled (plants/crops) *is fixed to the ground,* so that the robot must move to it in a safe manner even though the environment is only partially known and may change quickly, for example, a tree branch can change its position in seconds as a result of gravity.

Therefore, dealing with the management of mobile and cognitive robots is required to combine robotics and agriculture. In other words, smart mobile robots able to learn and make decisions with complex objectives in a partially-known and changing environment must be developed. Furthermore, it must be addressed how to increase the level of reconfigurability, adaptability, decisional autonomy, dependability, interaction, perception and cognitive ability and moving and manipulation capabilities for the agro-robotics systems. Thus, only by reaching *higher ability levels* than currently available, the required goals will be reached.

4. KEY CHALLENGING POINTS

4.1. Intelligent Manipulation for New Applications

Achievement of human-like dexterous manipulation skills will be the last frontier for robotics, in close relationship to full intelligence. In the next years, the manipulation community faces challenges at many levels to improve the limited skills that robot have nowadays in terms of dexterity, from end-effectors hardware development to control and planning methods for real-time physical interaction with the environment and humans.

Hardware level. On the hardware level, we will integrate sensors and new materials to enable grippers for safer and meaningful interactions. The challenge will be to utilize ***soft materials in the most optimal way*** to take the interaction and safety advantages they offer, but combine it with rigid materials to increase controllability and sensing capabilities.

Bi-manual manipulation. On the manipulation level, dexterous manipulation will move towards the next level to fully exploit the potential of ***human-like bi-manual manipulations to accomplish tasks,*** and not only with rigid objects but with ***flexible and textile-like objects,*** and ultimately also with humans.

Multisensor and extrinsic constraints. To accomplish this challenge, robots will have to utilize information from a variety of sensors to ***improve the understanding of the state of the manipulated object*** and the robot own contact interaction state at each step of an action. This will enable a robot to react appropriately to unexpected interactions. Advances in skin-like tactile sensors will contribute to providing robots with proprioceptive and kinesthesia sense to assess the interaction state of the robot with all their surroundings, including humans, objects and environment. Planning will consider ***explainable intelligent systems to understand all the possible contact interactions for a task.*** In the next years, full exploitation of extrinsic constraints will allow improving robot dexterity both with hard and soft grippers. Explainable physical intelligence will consider ***user preferences*** to develop a task in collaboration with users, exploiting extrinsic and human contacts and also contacts with the full robot body.

Data generation for learning. The advances in hardware devices to capture reliable manipulation data will push the proliferation of databases with ***real interaction data from humans,*** enabling the application of DL techniques to achieve dexterous manipulation skills from real experiences and explore human-like possibilities in re-grasp planning and highly dynamic manipulations.

Common benchmark. Finally, an important effort will be put by the community during the next decade to define a common benchmark to ***compare both hardware and manipulation skills*** (Garcia-Camacho, 2020). This will include the definition of quality measures that may be related to the ability of explainable decision making about manipulation. Benchmarking in manipulation is a crucial challenge that the community faces in order the enable science progress, including efficient ways of sharing new solutions both for software and hardware.

4.4.2 Dependable Autonomous Navigation for Cities

AVs have the potential to be a major disruption element in society and the industry in the upcoming years. However, to massively deploy this technology, a breakthrough is required in their navigation capabilities. *Autonomous navigation* refers to the ability of a robot to move from one place to another and therefore implies understanding the environment where it has to evolve, deciding the most appropriate motion plan while facing unforeseen situations, and executing the resulting plan without human intervention.

This technology is a ***critical component in emerging applications*** such as automated transportation of people or goods, underwater and aerial robotics-based solutions, smart farming or autonomous search and rescue. For these applications to become a dayto-day reality, ***much more dependable robots are needed,*** thus requiring significant progress beyond the state of the art in the three main pillars of autonomous navigation presented in the following.

Perception. Although substantial enhancements in objects detection have been achieved in the last decade, mainly supported by the uncontestable performance of DL techniques, significant progress is still required in the understanding of objects nature and their motion prediction. These features will become particularly relevant in novel approaches where the ***processing and learning stages have an intrinsically predictable nature.*** In addition to that, although multi-sensor architectures intend to cope with the intrinsic limitations of each sensing technology, there is still a long way to reach ***an effective immunity to natural variations*** (weather, night/daylight, harsh environments...). Finally, scene attention strategies is a field where a significant effort has to be addressed. Indeed, a wide range of applications —including crops fields, mining, industrial infrastructures, search& rescue environment and of course on-road driving— ***need specific perception schemes*** (e.g. (Santamaria-Navarro et al., 2018)) where events may trigger different learning-enabled processing and fusion considerations. To this end, the development of semantic digital maps (and in case they are not available, SLAM techniques), where availability and constant updates can be guaranteed, will be a key issue.

Decisional autonomy. Making safe, efficient and human-aware decisions in complex environments (like streets for vehicles or crowded areas for mobile robots) can be significantly challenging since the knowledge about the environment is generally incomplete and the associated uncertainty is high (Godoy et al., 2015). With the aim of reaching human-level abstract reasoning and reacting safely even in complex urban situations, ***autonomous navigation requires methods to generalize unpredictable situations and reason in a timely manner.*** The trend in the last years is to feed with huge amounts of data learning strategies with intrinsic imitation mechanisms. However, the underlying black-boxes do not allow to design and deploy coherent decision validation mechanisms. In this connection, a key challenge in this research field is the design of novel approaches able ***to embed context-aware adaptability mechanisms*** while remaining able ***to monitor the evolvable behaviour of the learning-enabled decision-making strategy.*** This will allow identifying

novel research avenues where certification of decisions will become a reality, enabling thus accountable AI systems and therefore solving the most complex liability issues we are facing today.

Motion. Understanding the spatio-temporal relationship between the subject robot/vehicle and the relevant surrounding entities, while being constrained by the navigation area specificities may be a very difficult challenge (Artuñedo, Villagra and Godoy, 2019). Indeed, ***different kinds of robots*** (cars, drones, wheeled mobile robots) ***will need to interact with highly dynamic agents,*** whose behaviour may be extremely complex to predict and, in some cases, even to perceive.

As a result, progress beyond the state of the art is required to produce robust, reliable, consistent and acceptable paths and speed profiles in crowded environment (cities streets, museums, special events). User acceptability and engagement is key in robots that have to interact or even transport human beings in a world where AI-enabled and human-driven systems will cohabit. To that end, ***human-aware and human-like motion planners have to be designed,*** introducing context-dependent human-compatible and cognitively aware motion patterns. A trade-off between adaptability and behaviour trackability will have to be guaranteed also in this aspect of autonomous navigation. Finally, some specific robotic solutions will need not only ***a compliant motion pattern but also novel control designs to track the resulting targets*** with the best closed-loop performance (Milanés et al., 2011). This goal will be particularly challenging in under-actuated or singular dynamic systems, where several actuators have to cooperate to respond to the motion requirements.

4.4.3 Seamless Cooperative and Everywhere Localization Solutions

Location-aware applications and impact. The location of robots, persons or any objects in general (as in IoT) are key components in the new era where ***context-aware societies need to know the geo-spatial location*** in order to make better decisions, more efficient and sustainable for the world. Better awareness of location means better mobility and better interaction between actors (persons, robots, or IoT-connected objects). In order to cope with the fact that people spend most of the time indoors, we need reliable location technologies for indoor localization (Renaudin et al., 2019).

Localization technologies for indoor and outdoor spaces. Outdoors, under good visibility conditions, the use of *Global Navigation Satellite System* (GNSS,

satellite-based solutions) is the norm, and the most accurate solution, which can even be supported by cell-based positioning for a coarse outdoor position estimation if needed. On the other hand, at indoor environments, such as in buildings (hospitals, universities, museums), tunnels, subways or connections halls (such as airports, stations), the GNSS-based solution is not operative or its quality is very low. Additionally, higher demand for accuracy is required due to the finer granularity of indoor spaces (Salazar Gonzalez et al., 2019). ***Several technologies are being proposed*** by researchers (radio frequency, ultrasound, inertial, vision, ultra wide band (Jimenez, Seco, and Torres-Sospedra, 2019),...), ***but none of them can cover alone all indoor scenarios of interest.*** That fact demonstrates the difficulty of the indoor-location challenge. ***In emergencies, where an accurate location of victims and first-responders is essential, these technologies are still not available for robust operation in real conditions.*** Further research and innovation are needed to cope with this challenge.

Seamless everywhere location-awareness. The final goal must be to integrate different technologies in order to make possible the estimation both outdoors and indoors without any jump or dead zone when transferring from indoors to outdoors and vice-versa. The current trend is to use low-cost devices, such as those carried by persons every day (smartphone) to ***fuse sensor information coming from different sources,*** both devoted for location, and also signals not intended for that use (signals of opportunity, such as light intensity, atmospheric pressure, and so on), in order to achieve robust but reliable seamless location solutions. This is a challenge that needs a mid-term research effort.

Cooperative localization and update. A very important approach for the future success of these location-aware technologies, is the maintenance of the *beaconing infrastructure,* that it is normally needed in order to apply trilateration or fingerprinting techniques. Indoor infrastructures, especially its electronic components (WiFi access points, BLE tags, etc...) change continuously, making the location models to be out of date in a few years. A *cooperative localization,* which includes crowdsourcing, is needed in order to update the models to guarantee an accurate location along years. The sharing of information between different actors (e.g. smartphone signals from thousands of users) is needed to make location-awareness a robust and sustainable objective.

These are just a few examples of the challenges that researchers must face, in order to be able to create a knowledge, that can be transmitted to the industry for final integration in a commercial product for the final benefit of the society as a whole. The localization challenges cited require a decade to be solved

adequately, but ***the goal does not end there since the requirements and new fields of application for anywhere anything location,*** and its integration in the smart-spaces and smart-robotic fields, ***will grow with time.***

4.4.4 Towards Robots for Everyone: Easy Reprogramming and Continuous Learning

Future robots need to be easy to reprogram and need to change their actions to adapt to several unforeseen situations. Configurability refers to the mechanisms the robot has to ***understand human preferences and readjust the goals*** to fulfil these expectations. Adaptability is the ability of the robot to ***understand changes and adapt to them,*** including generalization (the use of known actions to tackle new objectives). We will need to develop new algorithms to let non-experts to interact with the robot and alter significant parts of their programming.

Learning user preferences. Classically, robots have been programmed to accomplish a given task as fast as possible with fewer errors. This paradigm is changing now with the inclusion of robots in human environments. The principal goal now is to ***satisfy the user while doing the task.*** But the robot needs methods to understand which are these user preferences (Canal, Alenyà and Torras, 2019), that may change from user to user. The objective of the robot is no more efficiency (in terms of time or energy); other considerations come into play (Andriella, Torras, and Alenyà, 2020), like adaptation to human preferences, adaptation to presumed human needs, human acceptability, or safer executions.

Continuous learning from experience. It is impossible to pre-program a complex and modern robot. Thus, robots will be sold with basic functions and will enlarge their abilities with the ***capacity of learning new skills.*** In robotics, new experiences may be difficult to obtain as real robot executions can potentially be disastrous. A very promising alternative is the inclusion of non-expert users in the learning loop. In this paradigm, the robot can reason about its own knowledge, can even discover that internal models are not accurate or new actions are needed, and can actively ask for help (Martínez, Alenyà and Torras, 2017). A precise interchange of information between the robot and the users is needed. In particular, the robot must be able to explain the internal states, the reason for the decisions and why believes some information is missing.

Understanding the semantics of each action. Current robots have the ability to understand, up to some extent, the consequences of their actions. Generalization is desirable to help robots to cope with unforeseen situations.

However, this capacity is now very limited. We will need new methods to help robots ***realize the effects of the actions*** at different levels, from the changes in the environment to the potential harm to the external elements or the robot itself. Moreover, safety, privacy, and ethical implications are also very important when considering human environments. This knowledge relates to different layers of abstraction, and effective mechanisms are required to fill with content these levels, to effectively relate and make connections through the layers, and to be able to use them to make better decisions.

4.4.5 Safe and Ethical Human-Robot Interaction

The move from industrial robotics to human-centred robotics has placed *human-robot interaction* (HRI) at the focus of robotics research. HRI poses very demanding challenges to ensure: (1) reliable and situated ***communication***, and (2) compliance with the strict ***physical contact*** requirements involved, for example, in assistive and co-working domains.

Interaction. In both communication and physical contact cases, interaction must be:

- natural/friendly (to enforce usability by non-experts)
- safe (both at the physical and cognitive levels),
- reliable (robust to disturbances),
- efficient (facilitating correct task performance),
- adaptable (to changing situations),
- ethical (respecting human values).

Cognitive and social robotics. Communication challenges are tackled within the broad field of cognitive and social robotics, and entail important developments in several interaction modalities:

- vision and depth sensing for human pose and motion estimation (Simo-Serra, Torras and Moreno-Noguer, 2017), intention prediction,
- gesture and emotion recognition,
- speech recognition and production,
- cognitive user modelling, including the learning of preferences.

Physical HRI. Interaction entailing physical contact has given rise to the subfield of physical HRI, which tackles new challenges such as:

- distinguishing intentional vs. unintended contacts,
- intrinsic safety,

- compliant manipulation (Colomé and Torras, 2020),
- bidirectional haptics,
- telemanipulation and kinesthetic teaching by demonstration,
- brain-machine interfaces,
- prosthetics, wearables and human augmentation.

Collaborative robotics. These two types of interaction need to be combined to develop fully collaborative robots, customizable to individual user needs and capable of shortterm adaptation and long-term learning from experience. ***Collaboration adds extra challenges*** related to autonomy levels, such as task distribution planning, role assignment (Andriella et al., 2018) and coordination, closed-loop interaction, and shared control.

4.6. Resilient Robot Teams

A multi-robot system is composed of several robots that can interact with each other to complete a set of goals. Multi-robot systems can be classified in several ways based on the motives behind their design. One of these taxonomies (Farinelli, Iocchi and Nardi, 2004) considers two types of specific features grouped: *coordination dimensions* to characterize the type of coordination, and *system dimensions* referring to the system features that influence the development of the multi-robot system. The main challenge in a multi-robot system is to provide ***a robust and intelligent control system*** so that the robots ***can communicate and coordinate among them to complete the task.*** Hence, it has been found that designing the control architecture, communication, and planning system are the major issues discussed among researchers. A few examples of cooperative multi-agent robots applications are soccer robot, unmanned guided vehicles and unmanned aerial vehicles, micro chain, and paralyzed robot.

Coordination dimensions. They are related to levels such as cooperation, coordination and knowledge. The organization of a multi-robot system can vary from a *fully centralized* system, where a leader oversees the organization of the work of the other robots, to an *entirely distributed* system, where the team robots are completely autonomous in the decision process and there is no leader. In real-world scenarios, one of the main concern is to find ***the effective cooperation and coordination among autonomous robots to perform the task in order to achieve a high quality of overall performance.***

System dimensions. System features, such as communication, team composition, system architecture and team size —which are especially relevant in

the system development— must be taken into account in the design. *Communication* among robots can be observed as a cooperation mechanism, which can be direct and indirect. Although there are still several open research aspects around it, the latter type of communication can reduce the complexity of large-scale system design and avoid the need for synchronization between robots by providing a shared communication structure to be accessed by each robot in a distributed concurrent manner. Regarding the *team composition*, advances are required to determine what is the most relevant choice for each application: (i) homogeneous teams composed of members with exactly the same features, or (ii) heterogeneous teams comprising robots that differ either in hardware or control software. In relation to the *system architecture,* three main groups of different approaches can be considered: deliberative, reactive or even hybrid (García-Pérez et al., 2008). In a deliberative architecture, the solution involves a longterm plan that considers all the available resources to accomplish a global goal collectively, whereas, in a reactive architecture, the plan affects only the robots involved in the problem. The proper combination of both logic conceptions to the specific problem is still a challenging point in multi-robot systems Finally, *team size* is an important issue that can be considered explicitly as a design choice, exploiting strategies to adapt team size to the complexity of the problem.

Path planning. There are various approaches to path planning in multi-robot systems, however, finding the optimal solution is *NP-hard.* All feasible approaches to path planning are a compromise between efficiency and accuracy of the result, and can be classified into three groups: central, decoupled and coordinated approaches. *Central* approaches consider individual robots as part of a single, coupled system, and uses a single, central planning unit for calculating the plans. *Decoupled* approaches consider each robot on its own and every robot should calculate its own path. Finally, *coordinated* approaches support time efficiency over completeness, a complete path planning is omitted, and traffic rules, highly-reactive or swarm behaviour are used. ***Adopting the right strategy for every application,*** where robustness and adaptability properties can be significantly different, ***is still a long-term research topic.***

4.7. Dexterity and Efficiency Through Bio-inspired and Parallel Mechanisms

The design of mechanisms has been and will continue to be one of the most important lines of research within robotics. In general, most of these devices have traditionally exhibited a lower performance than that achieved by

biological systems, especially in terms of adaptability, dexterity and energy efficiency. However, in the last years, with the irruption of softer polymers and bio-compatible materials, the development of new manufacturing techniques (e.g. the additive manufacturing, the shape deposition manufacturing, and the soft lithography), and the emergence of novel design paradigms, a disruptive shift is expected in the future robotic mechanisms, with a significant impact in many fields of applications.

Bio-inspired robots. Bio-mechanics of animals that walk, run, fly, swim, climb, manipulate, crawl and roll have inspired the design of many robotic mechanisms that are still far from achieving lifelike capabilities (Gonzalez de Santos et al., 2009). To improve the current performances, ***new actuator concepts that provide energy recovery*** (elastic or regenerative) ***mechanisms, higher power to weight ratios, high torques at low speeds, appropriate impedance for interactive tasks, and better speed and energy efficiency are required.*** Low cost, modular drive systems with integrated self-sensing capabilities are also highly desirable. Some specific applications would also benefit from high powered miniaturised actuators, energy-efficient propulsion systems in multiple environments, and self-stabilizing mechanisms.

Researches devoted to studying the plasticity of plants in their morphology and physiology in response to the environmental conditions can also offer new design rules for materials and bio-inspired systems.

Soft robotics. From the knowledge gained in bio-inspired robotics developments, it is clear that to achieve robustness and adaptability in complex unstructured environments one possible solution could be that future robots include soft technologies into their designs. In this context, one of the greatest challenges ahead is the ***design of deformable synthetic actuators*** (soft muscle-like actuation), ***capable of providing high forces, power density, active stress, active strain, inherent compliance and variable stiffness.*** The prototypes implemented today still exhibit low speeds, low precision, and lack of repeatability. Therefore, methodologies for controlling these soft devices also need to be further investigated. Advances in this area will contribute to human-robot interactions and capabilities of robots to adapt to different operations and environments. Other challenges are the development of entirely soft robot body structures endowed with flexible and stretchable circuits (i.e. embedded sensors) that can adjust to their shape and motion and the future convergence of soft robotics with tissue engineering.

Mechanisms for manipulation. Current robotic mechanisms for manipulation are comparable to those of humans in terms of speed precision and strength, but not in dexterity. Therefore, the main challenge ahead is to design mechanisms that ***enable robust whole-hand grasp, dexterous manipulation and in-hand manipulation*** of a wide range of objects. These mechanisms should have sufficient degrees of freedom and be energy-efficient and compact enough to operate in restricted spaces. Of particular interest are anthropomorphic adaptive fingers, modular systems, differential mechanisms that enable the hands to conform to unknown object shapes without feedback, and intelligent mechanical designs that incorporate functionalities traditionally accomplished through explicit control. Biological inspired variable compliance and variable impedance actuators should be further investigated to achieve manipulations mechanisms that result dynamic, robust and safe for interaction and that exhibit human properties and performances (e.g. weight, torque, stiffness, power and efficiency). The design of novel mechanisms for manipulation should also go hand-in-hand with the advancement in perception capabilities that enable a better understanding of the working scenarios (Fernández et al., 2018; Fernández et al., 2013).

Parallel robots. Parallel robots have received a lot of attention in the last decades due to their capacity to offer simultaneously high speeds and high accuracy in many applications. However, ***further research should be accomplished to exploit their whole potential and capabilities.*** The study of reconfigurable and modular robots in which the position of the joints can be varied, obtaining different kinematics characteristics and dynamic behaviours is still an open problem. Efficient numerical optimization algorithms and new methods for dimensional synthesis should be also explored. Of particular interest is the development of qualified calibration methodologies for cable-driven parallel robots. Their highly nonlinear dynamics, which increases notably when operating at high accelerations, tend to produce mechanical vibrations. Therefore, not only more efficient dynamic analyses that take into consideration the effects of structural elasticity but also mechanisms to avoid or suppress these vibrations, should be addressed. Novel designs combining cable-driven parallel robots with other mechanisms based on new soft materials could contribute to improving stiffness and load capacity of parallel robots. More advanced controllers are also required for higher trajectory tracking performance.

CHALLENGE 5

COMPUTATIONAL COGNITIVE MODELS

Coordinators
M. D. del Castillo (CAR, CSIC)
M. Schorlemmer (IIIA, CSIC)

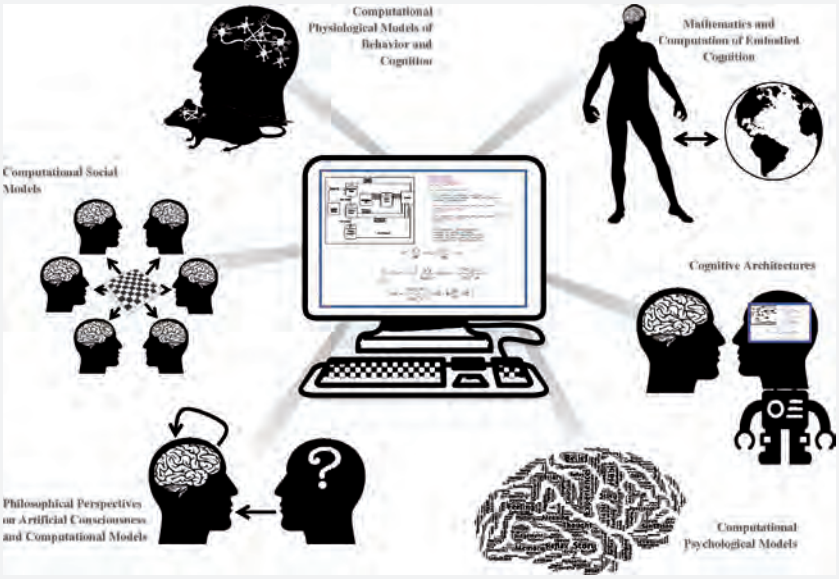
Participant researchers and centers
J. I. Serrano (CAR, CSIC)
L. Menéndez de la Prida (IC, CSIC)
V. Gallego (ICMAT, CSIC-UAM-UC3M-UCM)
D. Rios (ICMAT, CSIC-UAM-UC3M-UCM)
T. Galla (IFISC, CSIC – UIB)
C. Mirasso (IFISC, CSIC – UIB)
M. San Miguel (IFISC, CSIC – UIB)
R. Toral (IFISC, CSIC – UIB)
T. Ausín Díez (IFS, CSIC)
A. Serrano de Haro (IFS, CSIC)
Ma. Toboso Martín (IFS, CSIC)
A. Wagner (IFS, CSIC)
J. Sabater-Mir (IIIA, CSIC)
S. Canals (IN, CSIC - UMH)

1. EXECUTIVE SUMMARY

A model is a simplified and abstract version of a system or phenomenon that keeps its main features while leaving out other unnecessary issues. Working and probing with the model can bring insights about the original system and help to better understand it. The Chapter Computational Cognitive Models deals with attempts to ***describe and simulate human cognition and behaviour by computational models and algorithms*** focusing on different aspects of them.

This chapter brings together researchers from disciplines like AI, computational neuroscience, mathematics, statistics, physics, and philosophy in order to present major challenges in relation to the provision of computational models of cognition and behaviour from different complexity and observational perspectives. The challenges aim at both the understanding of the principles of intelligent behaviour of living beings endowed with nervous system and of the human mind, by building computational models of cognitive processes at different abstraction levels, and applying the learned knowledge to develop intelligent devices to simulate, boost, expand and recover living beings'

FIGURE 5.1–Computational cognitive models.



cognitive abilities. Computational modeling allows to tackle hard to reach problems, including stochastic issues, large scale problems due to a huge number of parameters and variables and complex forms of dynamic interactions, based on experiments.

The first section of this chapter describes models of neural biology and physiology that support complex cognitive functions including memory, language or consciousness. The second section describes key features of different psychologically plausible approaches to modeling human mind through procedures of general AI. We then focus on the integrated modeling of effective elements and emotion recognition with human decision making abilities. The next section emphasises intelligent behaviour and cognitive abilities as the result of embodied, situated systems, coupled with an environment through sensory and motor processes. The fifth section deals with modeling social phenomena by detecting patterns from collected social data. Finally, we present an analysis and discussion about AI and computational modeling, highlighting the limitations of underlying theories, from a philosophical viewpoint.

After a general overview of the research field and its approaches, the chapter highlights the impact of this field on basic science, its potential applications and current key challenges for each of the different perspectives.

2. INTRODUCTION AND GENERAL DESCRIPTION

Herbert Simon and Allen Newell can be considered the parents of the interdisciplinary *cognitive science* (CS), since they pioneered proposing a mechanistic definition and explanation of cognitive phenomena, thus linking psychology and computer science for the very first time.

There exist many different definitions of cognition coming from computer scientists, physicists and engineers that take into account issues like whether cognition is inherently human or can be extended to other animals or even artificial systems; whether cognition is innate or acquired through development; whether it is subdivided into different tasks or modules and how these modules may interact; the ultimate function of cognition; or the voluntary character of cognition and the implications of consciousness in it. In turn, disciplines like psychology, neuroscience or anthropology consider cognition to be related to the mental processes underlying human behaviour either in individual or social contexts and agree that perception, memory and thinking or reasoning are fundamentally cognitive processes.

With the publication of the first issue of the *Cognitive Science Journal* in 1977, CS was defined as a discipline created from shared interests among the people who study cognition from different points of view highlighting that the crucial issue for CS is the understanding of cognition, either real or abstract, either human or artificial. By the early 1990s, CS was established as a field that studies cognition using resources from several disciplines, including cognitive psychology, AI, linguistics, philosophy, neuroscience, statistics and cognitive anthropology. This chapter addresses this subject from different scientific points of view.

2.1. Computational Physiological Models of Behaviour and Cognition

Computational neuroscience aims at describing neural structures and processes through computational modeling and mathematical analysis. These processes range from the simplest computations at the cellular level, to the interaction between large groups of neurons that support the balance between

segregation and integration in neural circuits. These computations support complex cognitive functions including memory, language or consciousness (Kriegeskorte and Douglas, 2018). In a system as complex as the brain, made up of myriads of distributed elements interacting in parallel, *the properties that emerge cannot be inferred from the simple sum of its parts.* Therefore, when building computational models based on the physiology of the system and with the aim of reproducing some of its cognitive abilities into artificial systems, we need qualitative and quantitative neurobiological information at different spatial and temporal scales. A major challenge today is to distill those rules and principles to build next generation brain-inspired computational solutions.

Traditionally, computational and physiological models aim to decompose the complexity of behaviour and cognition into simplified representations and processes. For example, single-compartment models describing the time evolution of the membrane potential, and some ionic currents, at the axon level. Dynamical hierarchies from single neurons to brainwide networks are critical for behaviour and cognition. Over the last years, neuroscience research has inspired models of neural activity at the micro- and mesoscopic levels. These models expand from very basic processes at the level of single neurons (e.g. Hodgkin- Huxley multi-compartmental) and neural masses (e.g. Wilson-Cowan model, neural-mass models, neural networks) to more abstract representation of actions (e.g. agent-based models; dynamic causal modeling; reinforcement learning models; causal learning models). While this strategy has been fruitful, models still lack the ability to generalise to new situations beyond what they were fitted or trained for.

More elaborated models were usually discarded due to (1) the lack of computational power to tackle realistic structures, but also (2) a primitive neurobiological understanding of the actual underlying mechanisms, and (3) weak theoretical frameworks in which the many physiological details gain a functional sense. At present and in the years to come, *computational power will disappear as a limiting factor,* opening the possibility to incorporate more advanced dynamical models and a richer repertoire of neural elements (cellular compartments, neuromodulation, dynamic wiring diagrams) (London and Häusser, 2005).

More recently, leading discoveries are shifting our concepts to model behaviour and cognition more efficiently (Love, 2016). For instance, understanding how we learn from experience has made evident the major role of internal

driving forces (Steinmetz et al., 2019). This has resulted in conceptualising new reinforcement learning models guided by predictions instead of relying on external instructions. Such approaches might lead to new architectures in which artificial agents are able to compare immediate rewards with internal predictions in a distributed network (Dabney et al., 2020). Another example comes from the realisation that in some experiments the observed brain activity organises in stable low-dimensional manifolds to drive learned behaviours and actions (Gallego et al., 2017). The idea that major latent dynamical structures are underlying simple motor control and more elaborated cognition is currently a hot topic in the field. Establishing bio-inspired models of these latent variables and states are therefore a fundamental direction in developing new architectures in AI.

2.2. Cognitive Architectures

Cognitive architectures (CAs) represent an approach to modeling human mind through procedures of general AI for reasoning across different domains and adapting to new situations. CAs are influenced by diverse disciplines like computer science, cognitive psychology, philosophy and neuroscience. The key feature of CAs is their psychological plausibility by modeling human behaviour and underlying cognitive processes, which can lead to a scientific understanding of human mind (Sun, 2007). CAs depict approaches about the mental representations and computational procedures that process these representations. The intelligence of a cognitive architecture can be seen as the set of capabilities and behaviours it exhibits. CAs and intelligent software systems share the use of structures and pattern matching mechanisms. Most intelligent systems take an engineering approach creating systems simulating or even emulating partial facets of intelligence while the research on CAs pursue a *unified theory of cognition* in order to tackle a complete spectrum of human intelligence and processes of the mind.

A general criterion to categorise CAs may be the type of representation and processing procedures used. There are three major classes: symbolic, emergent and hybrid. *Symbolic or cognitivist* architectures represent information and knowledge as intelligible symbols that can be handled by production rules. This is the most common representation for dealing with planning and problem solving cognitive areas. *Emergent or connectionist* architectures build parallel models of nodes where processing consists in spreading data among nodes like neuronal models and connectionist logic systems. Symbolic unlike emergent architectures need an initial knowledge but the emergent ones require a training

phase and besides, the behaviour shown is more opaque to understand. *Hybrid* architectures integrate components of both classes of paradigms and are the most numerous in the literature. Among the most popular cognitive architectures are Soar, ACT-R, CLARION and ICARUS (Langley, 2017).

The main elements of human cognition to model in CAs include perception, memory, attention, actuation, social interaction, and problem solving, among others (Adams et al., 2012). There are many architectures that implement only a subset of them or even some architectures are specialised in concrete areas of general AI. Currently, most effort in CAs has been centered on high level abilities as perception, action selection, memory and learning. Perception symbolises the process that translates the input information into the representation used by the CA to perform its cognitive tasks. Input information can be presented in different modalities: vision by physical sensors or simulated, audition, symbolic, proprioception or a combination of several of them, depending usually on the specific application planned. Attention represents the mechanism of selecting relevant information from the input. Action selection is the procedure devoted to choose the goal to achieve and the way to do it according to diverse criteria. Memory plays an essential role in CAs since it is the storage for intermediate results of the computing process and makes learning possible. The features of memory systems like duration (short and long-term) and type (procedural, declarative, episodic,...) are borrowed from psychology. Finally, learning is the capability of a CA to improve its performance over time and can be divided into declarative and non-declarative depending on the nature of the knowledge (facts, skills).

2.3. Computational Psychological Models

The development and implementation of computational psychological models has a long tradition in AI. Take for instance, as a paradigmatic example, ACT-R (Anderson, Matessa, and Lebiere, 1997), the CA developed by Anderson. Using this architecture as a base, many researchers have developed computational models that try to reproduce different aspects of the human mind and human behaviour (ACT-R Research Group, 2013): language processing, perception and attention, problem solving and decision making or learning and memory are just some examples of the aspects these models try to cover. These *computational models are used as tools* (simulation tools) *to provide support to specific psychological theories.* So, from this perspective, computational psychological models are at the service of psychology.

There is, however, another point of view that recently is taking more and more relevance: *the use of these models at the core of autonomous entities* so they can exhibit some kind of realistic human behaviour. In this case, the underlying psychological theory is accepted as a valid model for certain human behaviour and the computational models are used to provide to the artificial autonomous entity (be it software or hardware) the capacity to reproduce that behaviour. This capacity can be very useful in domains like social robotics or agent-based social simulation, where simulating human behaviour realistically is essential.

Two major issues in the development of intelligent systems refer to incorporating decision making capabilities and affective features through appropriate models. This would increase the usefulness and acceptability of such systems.

2.4. Mathematics and Computation of Embodied Cognition

In the last decades there has been *substantial empirical research providing evidence that human cognition is ultimately grounded on our physical bodies* and sensorimotor experience with the environment (Barsalou, 2008; Clark, 1999; Shapiro, 2011). Proponents of this view of CS have explored the embodied character of the cognitive functions of the mind and its fundamental relevance for human conceptualisation and reasoning (Lakoff and Johnson, 1999; Varela et al., 1991).

Despite of this, computational approaches to conceptualisation and reasoning as carried out in AI research and development continue to be largely based on *traditional understandings of CS, which consider the human body as secondary and peripheral* to understanding the nature of mind and cognition. This is the case for both, traditional knowledge-based, symbolic approaches founded on logical and analytic philosophy, and also current data-driven, statistical methods based on artificial neural networks (Goodfellow, Bengio and Courville, 2016).

However, there is increasing evidence that the cognitive principles that arose from analytic philosophy are insufficient as starting point for the modelling of conceptual systems, which must include at least the notions of prototype, image schemas, and also conceptual metaphor and blending. There is now an *ongoing effort to establish bridges between the embodied view of cognition and computational realisations of conceptualisation and reasoning* so as to foster the symbiotic collaboration of human and AI. These efforts include the formalisation of the embodied cognitive principles and their implementation in computational environments (Anderson, 2003).

2.5. Computational Social Models

Our ability to collect large data sets and to carry out computer simulations has transformed fields such as physics or biology. Entire new disciplines have emerged, including computational physics and computational biology. Similar developments are also seen in the social sciences [1], albeit at a slower pace. The new field of *computational social science* is emerging. It sits at the intersection of multiple disciplines, including the social sciences, computer science, and physics. The extent of interdisciplinarity is such that modelling social phenomena has become a well-established activity among physicists [4]. The aim of computational social science is to study, understand and predict social phenomena using computational tools. It is not coincidence that this field is emerging at a time immediately after the advent of many of modern tools of *information and communication technology* (ICT) which we use in day-to-day life. A large fraction of the population has access to portable smart devices with continuous online connection. As a consequence, the analysis of social systems is facing a flood of data. Against this context, the field of computational social science can be characterised as follows: "A new field of science in which new type[s] of data [...] can be used to produce large-scale computational models of social phenomena." [2]

Big data (BD) is having a critical impact in all areas of human economy and society. Standard approaches focus on statistics and decision theory founded problems, which only support a single decision maker against his or her environment. However, multiple agents collide in many applied areas in business and policy. Thus, it is of customary importance to adopt mathematical models for conflict and collaboration in these systems. The traditional approach has been game theory with two distinguishing features: a traditional divide between cooperative and competitive situations; competitive situations pervaded by game theoretic concepts in turn dominated by common knowledge concepts.

Computational social science relies on techniques to collect data, efficient algorithms to detect patterns in this data, and on the construction of what has been described as 'generative models' of social phenomena (Conte et al., 2012). These models are linked to, but different from *machine learning* (ML) approaches to large datasets (Wallach, 2018). In particular, generative models systematically link cause and effect in social dynamics, and their aim is to be able to answer 'what-if' type of questions.

2.6. Philosophical Perspectives on Artificial Consciousness and Computational Models

Cognition and rationality have been in the center of philosophical investigation since ancient times. And AI and consciousness belong to the key topics in both, theoretical and practical philosophy. Contributions to this challenge have been made in philosophy of mind, epistemology and philosophy of cognition, philosophy of language, phenomenology and ethics.

A wide range of issues in the field of AI and computational cognitive models draw back to traditional ontological and epistemological questions: on the nature of cognitive processes like perception, memory, learning, thinking, feeling, language, reasoning and consciousness; on the correlation of concepts and intuitions, words and meanings; the mind-body problem, etc. Along the question of how to conceive the psycho-physical nexus, we distinguish between *dualist* and *monist* perspectives. There are three basic kinds of monist approaches: mentalism, neutral monism and physicalism – the latter being either eliminative or reductive (behaviourism, identity-theory) or non-reductive (anomolous monism, realisation physicalism and supervenience theories). Dualist approaches, on the other hand, can be classified in substance dualisms and dual attribute theories, be they organismic (emergentism, epiphenomenalism) or nonorganic. Apart from that, there are structuralist and functionalist approaches (e.g. the classical machine state functionalism, psycho-functionalism or analytic functionalism).

While the above mentioned topics concern theoretical philosophy, practical philosophy addresses the ethical implications of living with today's and future generations of intelligent systems. The central questions in this field depend on what kind of AI we address:

- With regard to already existing problem-solving AI (or *weak* AI), ethical questions focus on how to use the intelligent devices in a responsible way, i.e. enabling human self-realisation, enhancing human agency, increasing societal capabilities and cultivating societal cohesion.
- With regard to projects dedicated to develop AI similar to human intelligence, including artificial consciousness (*strong* AI), ethics focus on whether we have to treat such systems like autonomous persons, holders of rights, moral subjects with duties and responsibilities.
- Trans- and posthumanist approaches committed to developing silicon-based *superintelligence* triggering a singularity raise questions about the proper ethics of such intelligent systems, about responsibilities, loss of control/autonomy and the desirableness of certain scenarios.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

3.1. Computational Physiological Models of Behaviour and Cognition

The field of AI has faced a spectacular growth in the last decade. Originally inspired by neuroscience its application is now becoming more *interdisciplinary* than ever. A research plan based on the above challenges is expected to impact both, computational science and neuroscience (Hamrick et al., 2019; Guggenmos et al., 2020). The use of advanced mathematical models combined with extensive computer simulations and supported by experimental observations will result in great advance in the understanding of cognitive functions. By better understating basic principles of intelligent behaviour a full range of applications are expected for both basic science and technology. Modeling behaviour and cognition is at the forefront of this endeavour. The field today is strategical in that it will change not only the way we study the brain, but how we use brain-inspired solutions to make AI better, thus transforming available technological solutions for unmet needs. We envisage major impact of developing new computational models of behaviour and cognition in a range of basic science applications like computer vision, face identification, speech and language application, self-controlled navigation, etc. At the same time, this effort will be the basis to develop new technologies, ranging from efficient communication strategies in complex networks, to humanise autonomous agents implementing cognitive capabilities and to better predict and explain specific individual behaviour from population dynamics to transform healthcare, policy-making, e-commerce, etc.

3.2. Cognitive Architectures

Most CAs are being used as research tools. However, there are practical applications in different domains that are a benchmark to assess the underlying theory of every CA (Kotseruba and Tsotsos, 2020). Robotics, *natural language processing*(NLP), psychological experiments and *human-machine interaction* (HMI) are a few of them. In robotics, the major applications are navigation and obstacle avoidance, object manipulation and industrial tasks. In NLP, most applications focuses on syntactic and semantic processing of textual data, reading process and speech recognition. HMI is mainly devoted to develop CAs for supporting decision making, which can learn and collaborate with humans in problem solving. Since CAs model human behaviour, psychological testing present the largest number applications, focusing on experiments about memory, perception, attention and decision making.

An interesting and useful potential application is to apply CAs to model behaviour from psychological and physiological data of diseases that exhibit cognitive deficits in order to make a differential identification or diagnosis of the diseases. Besides, the resulting models can bring into light plausible neuroscientific hypotheses about the diseases that can be investigated from a multidisciplinary viewpoint.

3.3. Computational Psychological Models

Psychological theories tend to be descriptive so the specification from a formal perspective and the level of detail is usually quite limited. More than often, the level of detail is not enough to allow a straight computational implementation and requires further development of the original psychological theory. In this sense, the computational model, apart from its own intrinsic utility once finished, serves during the creation process as a catalyst to improve the original psychological theory. It forces the researchers to delve into aspects that in the initial version of the theory were ignored. Given that, the creation of computational psychological models will change the way social sciences (in this case psychology) develop new theories or revise those already accepted. There are several potential applications:

- Computational psychological models *as a tool to support the development and validation of psychological theories.*
- Computational psychological models *as a generator of human behaviour for an artificial entity.* Any application where realistic human behaviour is important can benefit from these kind of models: social robotics, social simulation or training environments are good examples.
- Computational psychological models *as a tool for the analysis of social data.*
- From the basic science point of view, major challenges include encompassing parametric models for individual and group decision making impacted by affective elements; integrating emotion recognition with decision making and calibrating the models with users with different personalities. There are many potential applications in relation with affective group decision making models of which we mention two important ones:
- *Affective social robots.* These may be used for many purposes including robots as teaching assistants, robots as assistant therapists for autism, accompanying the elderly and the like.
- *Internet of the affective things.* These entails delivering decision making and affective capabilities to connected devices typical of Internet of things so as to improve interactions with users.

3.4. Mathematics and Computation of Embodied Cognition

Proper mathematical models of embodied cognitive principles such as image schemas, analogy, conceptual metaphor and blending are prerequisites for a computational realisation thereof. Furthermore, such cognitively-inspired computational models are also necessary for the effective deployment of computational systems in human societies. Consequently, the impact of this research goes well beyond the scope of AI or computer science in general, and addresses also important societal issues.

From an AI perspective, though, the potential application domains in which the mathematics and computation of embodied cognition will most likely have a significant impact include:

- *Computational ontology.* As used in computer and information science, ontologies proved to be useful for addressing the semantic heterogeneity of highly specialised domains such as medicine, biology, genetics, anatomy, or geography. However, they proved to be inadequate for targeting more experiential and aesthetic domains. Computational ontologies founded on embodied cognitive principles may prove to be useful to address the conceptual diversity and dynamism in these domains, capturing the embodied, mainly non-conscious, and largely metaphorical and imaginative dimension of human conceptualisation.
- *E-learning.* The view that cognition is ultimately grounded on our bodily sensorimotor interaction with the environment is also relevant to how abstract concepts and ideas such as those of the STEM disciplines are taught and acquired. Computational frameworks that are based on embodied cognitive principles might lead to computational tools for e-learning that enhance the pedagogy of STEM (Weisberg, 2017).
- *Computational creativity.* Embodied cognition has proved to provide important principles on which to explore creativity in the arts and also in problem-solving. Analogy, metaphor, blending, have been advocated as important constituents of human creativity, and any advancement in the formal and computational understanding of embodied cognition will have a direct impact on the development of computational systems that assist and enhance human creative thinking and artistic expression.
- *Explainability for AI.* Trust and social acceptance of AI is closely linked with transparency and explainability of AI-based computational systems. Consequently, advances in computational approaches to the embodied view of cognition may prove significant for how AI-based

systems need to communicate with humans, taking into account the importance of image schemas, analogy, conceptual metaphor and conceptual blending in generating the narratives needed for effective explainability (Westberg, Zelvelder and Najjar, 2019).

3.5. Computational Social Models

The constructions of models of social phenomena is all but certain to deliver impact for other fields of basic science. Social phenomena at the macro-level (society as a whole) arise from the collective interaction of individuals at the micro-level (Conte et al., 2012). The development of methods to describe this link will make a difference for other areas of science, such as computational biology. The simulation of social systems with thousands of individuals requires computational capabilities and fast algorithms, transferable to other areas of science involving systems with many interacting components.

From the basic science perspective, two main issues are foreseen:

- Providing a model for competitive decision making which *overcomes common knowledge assumptions traditional of game theory.*
- Providing parametric models that allow *for transitions from cooperation to competition* and different degrees of cooperativeness among its members.
- Potential applications abound in many multi-agent, interconnected systems that are becoming part of the society. Here we mention but a few.
- *Smart-grid optimisation.* Consider a grid compounded of several electrical generators. In order to meet current power demands, they should be able to reach a consensus between all the generators in the system. See e.g. (Milton, 2015).
- *Autonomous driving.* Coordination of groups of autonomous driving systems is relatively simple with coordinated multiagent systems. However, for a long while we shall have the coexistence of autonomous and non-autonomous driving systems. Thus methods to coordinate such heterogeneous systems is essential.
- *Decentralized autonomous organisations.* Recent developments in cryptography and decentralized systems have been crucial in the development of technologies such as blockchain, which have made possible the beginning of these kind of organisations (see *https://en.wikipedia.org/wiki/Decentralized_autonomous_organisation*). Computational models that explain their interactions are thus in need.

- Application areas also include processes of cultural globalisation and polarisation (Centola, 2007), cooperation (Eguíluz, 2005), language contact and bilingualism (Castelló, Eguíluz, and San Miguel, 2006), as well as electoral processes (Fernández-Gracia et al., 2014), social contagion (Czaplicka, Toral, and San Miguel, 2016), opinion formation (Peralta et al., 2018)].

3.6. Philosophical Perspectives on Artificial Consciousness and Computational Models

AI studies is a realm of cognitive science at the interface of computational disciplines, philosophy, psychology, cognitive anthropology, neuroscience and linguistics. The contribution of philosophy to cognitive science can be classified as follows:

- *Conceptual* issues: clarify key concepts like intelligence, consciousness, experience, intentionality, perception, emotion, rationality, meaning, information, data etc.
- *Meta-theoretic* issues: concerning the practice of cognitive science and its foundational assumptions.
- *Basic epistemological* issues: reflections on AI shed a new light on traditional philosophical issues about the mind and basic ideas about the nature of cognitive processes.
- *Ethical* issues: AI raises new ethical questions that defy and go beyond the classical approaches.

With regard to computational models of the mind, philosophy serves to identify advantages and problems. Philosophy helps to define some conditions that models must meet and to disclose unresolved issues. To give some examples: models have to take into account the differences of ordinary and formal language, understanding language as a practice (including flexibility, vagueness etc.); implement the correlation of semantic, pragmatic and syntactic aspects; include cognitive features like intersubjectivity, imagination, creativity, emotion, etc.

4. KEY CHALLENGING POINTS

In the following we list the current key challenges in computational cognitive modelling as tackled from the different approaches and perspectives discussed above.

4.1. Next-Generation Bio-inspired Algorithms

In order to improve existing technologies, we need to ***better understand how the brain solves complex tasks.*** Inspiration should come from biology, neuroscience, psychology and social behavioural studies. This includes building computational models incorporating important physiological details as dendritic computations, short and long-term synaptic dynamics and neuromodulation, implementing wiring diagrams inspired in the last connectomic findings (*db.humanconnectome.org, sites.google.com/site/bctnet*) and using the developed functioning principles to implement cognitive capacities in artificial systems (Vernon, 2014). Our goal here is to identify and modeling latent state variables and dynamics underlying a subset of prototypical basic behaviours such as: decision-making, pattern separation/completion, transfer learning, domain-adaptation and causal reasoning. We will work to develop new sets of models of these basic intelligent behaviours with the major goal of fostering bio-inspired ML algorithms. Also, we aim at incorporating affective abilities to recognise, express and communicate emotions by designing versatile internal meta-controllers incorporating external and internal emotional-like drives to handle with the agent's predictions and expectations of some of the basic intelligent behaviour detailed above. Addressing this challenge will result in some basic prototypes (models and robots) able to process emotion-like behaviour.

4.2. A Behavioural Clone

A behavioural clone that simulates the behaviour of every human being in any existential stage in a personalised way. From childhood to adulthood, ***the clone will model ecologically developmental processes in cognitive skill acquisition in a human-like manner*** while providing a natural and reliable human-system interaction by developing a more realistic multimodal perception and bidirectional verbal and non-verbal communication. This challenge needs to answer the question about what initial data structure, knowledge and algorithms must initially have the clone (*phylogenetic*) and how these components are developed through everyday experience with the surrounding environment (*ontogenetic*) in order to build a biologically and psychologically

plausible clone. First, this clone will be able to early detect behavioural disorders in the cloned individual and to simulate the response to prescribed therapies allowing to optimise them before their real application. Second, the clone's behaviour can be altered in early developmental stages to follow up its temporal evolution helping to prevent future diseases in the cloned individual. Third, the clone can contribute to enhance the capabilities of the cloned individual by facing new circumstances and noticing the resulting behaviour. Fourth, the average behaviour of a set of clones can be useful to propose the best policies adapted to the group (for example, a pandemic situation). And finally, the clone of any one will be timeless ensuring his/her persistence beyond the death.

4.3. A Comprehensive Computational Model for Social Interaction

The necessity of advanced social interaction between artificial entities and humans is already a reality. Robots helping disabled and the elderly, home-assistants like Amazon Echo or Google Home, or avatars performing as digital teachers are just a few examples of artificial entities that need to interact with people in a natural, interpersonal manner. Unfortunately, the computational models these artificial entities implement nowadays are still far from providing a good experience from the perspective of the human being. These kind of ***artificial entities need a good interface with the human*** (NLP, computer vision, emotion recognition and expression) ***but also an internal psychological model that tells the artificial entity how to make decisions*** based on all these inputs and decide how the outputs need to be according to its final purpose. Sub-challenges that such a computational psychological model should overcome are, for example, how to embed affective elements within a coherent decision making structure or how norm compliance and values should drive the artificial entity behaviour.

4.4. A Collaborative Intelligence Based on Human Embodied Cognition

A key challenge is to ***implement and deploy computational systems that*** —in order to effectively collaborate with persons— ***take into account the way humans make sense of the world.*** This requires to define uniform mathematical models of embodied cognition and to develop computational realisations thereof. There are currently many different proposals of mathematical models for a variety of embodied cognitive principles, and also many computational systems that aim at capturing some aspects of embodied cognition. But they have all followed an ad hoc approach, relative to the particular application

domain, without putting the emphasis on general principles and integration. To achieve an effective collaborative intelligence based on principles of humanly-embodied cognition will also need to be linked with more traditional, knowledge-based approaches and current state-of-the-art data-driven, statistical methods. This is relevant for both building upon legacy logic-based conceptual models such as databases and ontologies and upon evolving learning-based systems for which the cognitive layer is not explicitly specified.

4.5. Two-Way Interaction Between Social Dynamics and Technology

The interaction of humans with communication technology is constantly evolving. Humans use technology for social interaction, and technology advances in response to the way it is used, leading to the formation of so-called *socio-technical systems* (Lighthill, 1973). Studying and understanding these is fundamental, in particular since this technology and its use have direct impact on our quality of life. A particularly relevant aspect in these systems is the one of information processing. Successes would for example consists of an ability to characterise how a particular technology changes social dynamics and, in reverse, what types of social dynamics drive the development of new technologies.

4.6. Validation of Models, Understanding Cause-and-Effect Relations

The validation of models against observational data or in targeted experiments (e.g. *behavioural game theory*) still requires continuous and iterated refinement of computational modelling approaches. Models often based in standard game theory (Osborne and Rubinstein, 1994) need to be extended to include games with incomplete information, as in Bayes Nash equilibria, concepts like adversarial risk analysis (Rios Insua, Rios, and Banks, 2009) or mechanisms to bridge the gap between cooperation and competition, (Esteban et al., 2020). Beyond validation for a specific case, models need to be able to answer what-if questions and disprove common-sense statements (Watts, 2011) by establishing cause-and-effect relations which are generally not obtained in data analysis by ML approaches. Success on this challenge consists of the delivery of models that can be useful for decision and policy makers: they have to be confronted with systematic testing and exploration of different realistic and hypothetical scenarios in real-word situations and to be able of reliable forecasting.

4.7. Understanding Embodied, Situated Cognition and the Role of Tacit Knowledge, Emotion and Cultural Factors in Human Rationality

This challenge is needed to enrich and foster dynamic, enactivist, and phenomenological approaches. One of the main tenets of cognitive science is that cognition is information-processing. According to the physical symbol hypothesis information-processing has to be understood as manipulation and transformation of physical symbols according to rules. The corresponding model of the mind as a computer (Newell and Simon, 1976) and its linguistic counterpart, the idea of a language of thought, the so-called mentalese (Fodor, 1975), gave rise to early computational theory of mind, but are no longer the only theoretical framework. Critics have been extensive: on the leading concept of representation, on not capturing intentionality, etc. Over time, dynamical and probabilistic models and non-cognitive positions, such as the enactivist perspective, embodied or situated cognition, have gained importance. The dynamical systems approach (Tim and Van Gelder, 1995) tries to explain how agents are embedded in their environments without using representation. Instead of conceiving intelligent behaviour as representational, computational and sequential, it is understood as interdependent, in constant-equilibrium and time-sensitive. This approach is intimately related to the situated cognition movement, according to which embodiment is the key to generate intelligent behaviour: the dynamically coherent coupling of the agent with its environment (Varela, Thompson and Rosch, 2016). The enactivist and phenomenological proposals on embodied cognition –with the focus on scaffolded, context-bound cognition, intersubjectivity, structures of attention and intentionality, and the feeling body– are considered extremely important by us. We hold the lived-body experience to be a crucial point, facing thus the hard problem of consciousness: the problem of explaining why and how some internal states are felt states (Chalmers, 2018). We pay special attention to temporality as primordial structure of consciousness and consider a study of the elements that constitute the human experience of time to be necessary in order to advance in the conceptual field of AI. Therefore, we will focus on temporal aspects of cognition (Arstila and Lloyd, 2014); on discrete vs. continuous time (Port and van Gelder, 1995); the past-future distinction (Prosser, 2016) as well as self-awareness, personal identity and action in time (Brook and DeVidi, 2001).

New advances in computational neuroscience consider the brain to be a multilevel predictive processor whose rationality consists in the ability to reason

about uncertainty. The main thesis of this Bayesian perspective of the mind and the predictive brain is that *human learning follows the principles of probabilistic inference* or at least that human behaviour is highly consistent with probabilistic reasoning. Despite the multitude of approaches, *main features of cognition resist formalisation.* Common sense reasoning, the understanding of causality or counterfactual reasoning –all three essential for morality, language knowledge and the understanding of the social, physical and biological world– cannot be programmed by rules. ***The whole realm of tacit knowledge, intuitive physics, folk biology and folk psychology***—a functional, encapsulated knowledge that we don't cognitively penetrate—***remains a key challenge to AI. Models of rationality provided by rational choice and expected utility theory fail within the empirical tests. Humans are no pure rational agents, and rationality has to be conceived as plural, linked to emotions and cultural-bound.*** With regard to the issue of learning, there is a desideratum of philosophical reflection on the research dedicated to the neural plasticity of the brain, which inspired the cybernetical approach of deep learning, also known as connectionism, with its basic idea of the parallel and distributed nature of brain processes.

The aforementioned ***epistemological topics have to be complemented by an ethical perspective.*** In the field of ethics of AI, we sustain that technological and ethical progress must go hand in hand. One of the new fields here is that of ethics by design. Furthermore we will address issues like: moral cognition; consciousness and conscience; emotional intelligence and the cognitive relevance of emotions; as well as inequalities and injustice in a technological world.

CHALLENGE 6

ETHICAL, LEGAL, ECONOMIC, AND SOCIAL IMPLICATIONS

Coordinators
P. Noriega (IIIA, CSIC)
T. Ausín (IFS, CSIC)

Participant researchers and centers
H. Mueller (IAE, CSIC)
M. Ellman (IAE, CSIC)
A. R. Cardoso (IAE, CSIC)
F. Aguiar González (IFS, CSIC)
D. López (IFS, CSIC)
M. Toboso (IFS, CSIC)
A. Wagner (IFS, CSIC)
A. García Aracil (INGENIO, CSIC – UPV)
S. Degli Esposti (IPP, CSIC)
L. M. Miller Moya (IPP, CSIC)
L. P. Jiménez (IRI, CSIC- UPV)
J. Pareto (IRI, CSIC- UPV)
C. Torras (IRI, CSIC- UPV)

1. EXECUTIVE SUMMARY

In six decades of history, AI has become a mature and strategic discipline, successfully embedded in mainstream ICT and powering innumerable online applications and platforms. Several official documents stating specific AI policies have been produced by international organisations (like the OCDE), regional bodies (EU), several countries (US, China, Spain, Germany, UK, Sweden, Brazil, Mexico...) as well as major AI-powered firms (Google, Facebook, Amazon). These examples demonstrate ***public interest and awareness of the economic and societal value of AI*** and the urgency of discussing the ethical, legal, economic and social implications of deploying AI systems on a massive scale. There is widespread agreement about the relevancy of addressing *ethical aspects of AI*, an urgency to demonstrate ***AI is used for the common good,*** and the need for better ***training, education and regulation*** to foster responsible research and innovation in AI.

This chapter is organised around four main areas: *ethics, law, economics and society (ELES)*. These areas shape the development of AI research and innovation, which in turn, influence these four areas of human activity. This interplay opens questions and demands new methods, objectives and ways to design future technologies. This chapter identifies the main impacts and salient challenges in each of these four areas.

In *ethics* the widespread consensus of the ethical aspects of AI has raised several AI- related questions for practical ethics about a responsible practice of AI, and therefore the challenge of ***embedding ethics in engineering through education and ethics by design.*** In addition, the concern about the unfettered autonomy of AI systems brought about the insight that it is possible to ***deploy artificial entities that behave ethically*** rising at the same time a fundamental debate about the feasibility and desirability of ***moral agency*** of artificially intelligent entities and systems.

The area of *law* faces the impact of new forms of social behaviour induced by AI that eludes current terminology and regulatory frameworks —like explainability of AI decisions and hidden agency in AI-powered applications— and therefore faces the challenge of ***adapting law to the uses of AI,*** including in this case the *legal personality of artificial systems*. Simultaneously, legal practice is adopting AI technologies for *automating compliance* in a rapidly evolving online environment, hence faces the challenge of ***developing new technologies for governance in augmented reality.***

Economics has received a notable *methodological impact* from AI technologies — like *deep learning* (DL) and agent-based modelling— and also has a profound influence in AI notions and methods. This creates the challenge to ***find synergies to fill the gap (epistemic as well as methodological) between economics and AI research agendas.*** Moreover, the significance of AI in the economy, in the role of stakeholders and its relevance for social well-being in general rises the need to ***identify and explain the repercussions of IA in the economy and in political economy policy.***

Finally, regarding *social science and society*, the emergence of new social phenomena linked to the digitally augmented reality where humans interact with intelligent systems has undeniable effects on how *socio-cognitive rationality* is built and *collective action* is taken. In fact, these hybrid space is creating a and new environment where human kind will co-evolve with AI entities. The challenge is to ***anticipate how that environment may be developed*** in order to foster human flourishing. The formation of public opinion, the evolution of social practices and sharing economy via AI-powered social networks are some of the phenomena induced by those interactions and hence justify the ***need for systematic empirical research of their consequences.***

CSIC is in a privileged position for a successful interdisciplinary approach to the challenges of AI in the ELES area. An interdisciplinary framework is

enabled by the confluence and synergies of the expertise provided by the following institutes: IAE (Institute for Economic Analysis), IFS (Institute of Philosophy), IIIA (AI Research Institute), IPP (Institute for Public Policies) and IRI (Robotics and Industrial Informatics Institute). Confluence and synergies would improve with the involvement of other CSIC institutes, in areas like anthropology, archaeology and music.

Some organisational actions may have a very positive impact on synergic interactions among ELES relevant CSIC institutions. In particular the creation of informal laboratories on topics related with the challenges described in this document; the fostering off joint participation in projects; the facilitation of joint contracts and research supervision; and the recruitment of personnel with expertise in psychology/cognitive science and law. One handicap that deserves attention is the lack of a strong group on law and technology in the CSIC structure.

2. INTRODUCTION AND GENERAL DESCRIPTION

There is a growing consensus about the strategic value of AI. The consensus is based on the awareness of the role AI plays in numerous and assorted applications in robotics, self-driving vehicles, e-commerce, health, security and advertisement. The acknowledged strategic value of AI comes as the result of the undeniable success of the embedding of AI in mainstream ICT. However, this success of AI can only be explained by two reasons. First, a maturation process of the discipline, the development of general purpose artefacts and the consolidation of a critical mass of experts, professionals and firms. And second, a timely combination of powerful IT infrastructure and the massive adoption of internet provided the fertile substrate for the mature discipline.

Two remarks as a matter of clarification. First, in this chapter we use the term AI as a liberal. Thus, AI includes all of classical AI, the technologies that come from it —including robotics, *machine learning* (ML) and *big data* (BD)— and, in general, AI entities and systems. Moreover since, as mentioned above, AI is embedded in ICT and is pervasive online, we sometimes call AI what, strictly speaking, would be an AI-enabled or AI-powered system or application.

Not unfrequently the press, and even some of the strategic documents, present an ambivalent perspective of the development of AI: sometimes unfounded optimism of the benefits of AI, sometimes dire predictions of its dangers. A likely explanation of the ambivalence is that AI, as most disruptive

scientific and technological innovations, faces the Collingridge dilemma; which postulates that when the discipline is emerging it is difficult to foresee its consequences but by the time one understands it, it may be too late to prevent the unwanted outcomes. We presume one may adopt a cautions but proactive attitude, striving to elucidate where AI is influencing society most, in order to anticipate its consequences and act in accordance. We follow that path by choosing the standpoints of the disciplines that are more salient with respect to the dilemma.

In this chapter we discuss the *social impact* of AI. We focus on four aspects: ethics, law, economics and society. In all four cases we explore how the uses of AI have social effects that pertain to these areas; we also explore how the development of AI poses questions and influences the four areas and how these four areas, in turn, influence the development of AI. Specifically, in the next paragraphs we outline, for each area, what that interplay with AI is about.

2.1. Ethics

Because of its object of study, AI research concerns itself with basic questions about the human mind and the purposeful activity of humans and society. Hence, AI applications impinge sensitive human and social activity by automating, enhancing, or taking over some tasks and roles that rely on traits of human intelligence. Thus, the ethical import of AI is two-fold. First, by posing some *classical ethical questions (introspection, responsibility, autonomy, rights and wrongs, justice)* in a new context where moral reasoning is to be formally and empirically modelled into an artefact. Second, in the ethical implications associated with the deployment and use of AI technologies in *sensitive applications (autonomous weapons, massive face recognition systems, medical diagnosis and prognosis, affective and care robotics)* in the framework of responsible research and innovation.

The acknowledgement of the potential benefits of AI together with its undesirable consequences has permeated into the AI research agenda recognising the opportunity, and the need, of research and innovation that puts the well-being of humans at the centre. This focus has been made explicit through terminology that reflects subtle shadings in the understanding of how AI research and innovation should be pursued: *responsible AI; trust-worthy AI; AI for-good; human-aware or human-centric AI.*

This shift of attention has found its way into policy at regional, national and institutional levels. A good example of how this approach is articulated can be

found in the European Approach to Artificial Intelligence, where ethics (together with a regulatory framework) plays a key role as postulated in its Ethics Guidelines for Trustworthy Artificial Intelligence[1].

2.2. Law

The intended development of a human-centric AI acknowledges the need for a *proper regulatory framework*. One that fosters responsible research and innovation in AI, facilitates a productive adoption of AI technology and protects the rights and needs of individuals, enterprises and society in general. Rather than a short-sighted approach to legal groundwork, most policy documents, and the EU one in particular, invite an effort to elucidate the rights that need to be protected, the directives, guidelines, standards and regulations that enable the protection of those rights, while fostering the best potential adoption of AI technologies by society.

In addition to this instrumental role of law in the development and adoption of AI, legal notions and legal philosophy provide valuable notions, intuitions and problems to the foundation and practice of AI. For example, the need to organise interactions among artificial (and natural) entities brings about a need for governance that may be articulated in economic terms as a mechanism or an institution, but in the legal tradition AI draws inspiration form the notions of norms, norm enforcement and compliance in general. On the other hand, AI by its own developments and by the artefacts it produces poses questions to legal foundations and practices that may be as mundane as the protection of the industrial property of a convolutional neural network that is trained for a specif type of medical diagnosis, to the subtle questions of moral agency and the dispute of the pertinence of the "legal persona" of artificial entities.

2.3. Economics

The rapid rise of AI poses a challenge both to the field of economics itself and will have a lasting, profound impact on the organisation of the economy and the political economy of society. It is important that research institutions provide solid support for embracing of the huge economic opportunities that the adoption of AI has to offer. This requires research which is informed by debates within economics and other social sciences and at the same time understands the new tools that the AI revolution has brought. There is otherwise a

1 The policy document and guidelines are available in https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625 and https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

very real danger that AI will be seen as aiding the concentration of political and economic power in the hands of a few and destabilising the political economy.

On a methodological side the cooperation of researchers in AI and economics should be especially fruitful as both are used to work with *quantitative data* and think in statistical models. Economics has already delivered inspiration for AI research and stands to benefit considerably from research which is currently ongoing in AI. However, methods are developed very quickly in both fields and new research will need to build *hybrid models* to facilitate communication between the fields. There are currently huge opportunities for an organisation like CSIC if research across centres and interchanges could be better coordinated.

2.4. Society

The adoption of AI technologies has also a profound *impact in society*. However, as what happens with other technologies, its uses may have unexpected consequences. The distinguishing trait of AI technologies, though, is that they mirror or potentiate cognitive and social behaviour of individuals and groups and by so doing modify society. In fact, one of the most significant outcomes of AI is the invention of an *augmented reality* where natural and artificial entities share a digital, and physical, interaction space. The upshot of such interactions is the emergence of new social phenomena and, eventually, some sort of co-evolution of society and natural autonomous entities.

With such prospects in view, a responsible attitude is to dedicate analytic attention to the ethical and societal impacts of AI, educate professionals and the public on those effects and help develop expertise in AI within the government and public institutions. The content of this chapter responds to this spirit.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

3.1. Ethics

A widespread acknowledgement of the ethical signiftcance of AI. Over the last few years there has been an increasing awareness of the ethical aspects of AI. It is grounded, mainly, on the realisation of the important *ethical implications* of the uses of AI (impact on labour, autonomous vehicles, face-recognition software, AI-powered e-commerce, recommender systems, etc.).

This realisation has four main effects: (i) It motivates the design of policy agendas and industry charters with a strong ethical component (e.g. the EU). (ii) It has prompted several initiatives and specific actions in favour of an ethical AI (manifestos, endowments, creation of institutes and research programmes, funding for research and support of coordinated actions, public pronouncements of major AI empowered stakeholders). (iii) It originates a pressing need to articulate directives and good practices for responsible AI research and Innovation. (iv) It induces a sharp recognition of the need for education in ethics and AI at all levels of society.

The debate on the autonomy of AI systems. From its very beginnings, the point of AI was to model rationality and construct artefacts that exhibit rational behaviour. After almost sixty years, the range of behaviour that has been implemented is quite broad and the performance is rather proficient in many cases. Now, if you "encapsulate" those forms of rationality into a system that can take its own decisions based on its encapsulated rationality, you have a certain type of "autonomy" that is specific to AI. In fact, such autonomy is the fundamental feature that differentiates AI from other disruptive technologies. The issue of autonomy becomes more sophisticated with the development of *autonomous agents*, systems that "act on their own" (see Chapter 2 in this book). These are systems that are autonomous in the previous sense but in addition, they are *situated* in a changing context where they are meant to interact with the environment and with other agents (human or not). The key point is that they decide how to interact based on their own internal decision model, that is their encapsulated rationality. These ideas apply not only to software systems, but also to robots that exhibit behaviour that is situated and reactive and is thus also referred to as autonomous. Common examples of these autonomous artificial entities are web information harvesters used for Google searches, automated online "trader bots", online airline ticketing services and conversational personal aides (like Siri and Alexa). But in the same category one can place autonomous agents whose behaviour is more complex and in a fundamental way, like self-driving cars and killer drones.

As agency of this type becomes more frequent and the interactions become also more consequential, the ascription of *responsibility* and *accountability* of those actions arises (which is not only an ethical issue but a practical and legal one). It is the problem of many hands and many things connected. They include engineering ethics of designers, manufacturers, and maintenance

systems; ethical aspects of the artefacts themselves (characterisation of moral agency of artificial entities); and ethical attitudes of the users (hybrid socio-technical systems of humans and intelligent autonomous artefacts). Hence, the debate is served: What does it mean when someone claims that an artificial system is autonomous? In what sense one may claim that an artificial entity has moral agency? What are the features of human agency that one may contend are essential to moral agency? Are they implementable in an autonomous agent? While one side of the debate is to determine to what extent an artificial entity may be autonomous, the other side is to decide whether it is even desirable that truly autonomous artificial intelligent entities be deployed.

The insight that it is possible to develop artiftcially intelligent systems that behave ethically. It has been postulated that one can use AI to control the unwanted side-effects of AI. The underlying insight is that ethical behaviour can be imbued in artificial systems as a form of control. The idea is to imbue values in the architecture of the autonomous entity or imbue values in the governance of the environment where such autonomy is present. The ideal situation is when one may prove that a given degree or extent of alignment of the autonomy with the values is achieved (the so called Value Alignment Problem).

3.2. Law

Addressing the pervasiveness, speed of innovation and the opacity of the uses of AI technologies. Legal practice is the result of a long evolving tradition forged on experience and guided by ideals like justice, common good and dignity of human beings. Its institutions are stable and effective for a vast majority of situations, however AI brings two major disruptive elements: the *speed* at which technology (and its pervasive and assorted use) changes, and its *opacity.* The first component creates loopholes and uncharted spaces for activities and actions that circumvent or are beyond the reach of current normative frameworks and by the time these frameworks adapt to the unwanted effects of the uses of that technology, the activities and actions have changed or use technologies that evade the adapted frameworks. Such is the case, for instance in the use of biometric identification technologies that, by the time they get standardised in a context of use industry has already deployed new deceiving devices. The second problem is that AI-powered activities and practices are often based on design components whose complexity makes it hard to determine compliance with a given regulation or include black boxes that

limit the disclosure of functionalities or the assessment of the form or extent to which certain compliance is achieved or evaded. Such is the case, for example, of the explainability requirement in legal appeal and redress of automated decisions (Article 22 of the GDPR).

AI enabled forms of agency and autonomy. It is not always clear who the principal of some actions is or should be. For instance, in the case of self-driving vehicles and automated diagnosis where the "many-hands problem" *dilutes responsibility* and the degree of autonomy afforded to artificially intelligent entities by sophisticated use of AI technologies *dilutes accountability*. Cloning of AI processes or agents may produce compounded effects that are difficult to foresee or contend with–like the stock exchange crises produced by fast-speed trading. Automatic micro contracting questions also the notion of collateral and auditable transactions, in as much as by the time a contract is agreed upon, it has already produced its intended effect and ceases to exist. In the wide picture, there is the debate of legal persona of autonomous artificially intelligent entities.

The automation of governance and compliance. While "legal expert sytems" promised to automate legal reasoning, machine readable regulation promises to make compliance a matter of AI processing and thus eliminate the need for legal experts and interpretation.

Reality is far more complex. In both cases. Underneath the challenging ideal of *compliance by design* there is a strong substrate of legal theory and a substantial collection of formal and AI artefacts that afford some automation in the process of governance and compliance within a legal system: legal ontologies, normative logics, norm specification languages, norm-based automated reasoning, online institutions, normative multiagent systems, to name a few. In addition, there are juice debates on enforcement, punishment and reparation, principle and value-based argumentation and, not surprisingly, a large amount of prototypes and actual applications of these ideas. And there are challenging opportunities, as seen in the next section of this chapter.

3.3. Economics

Methodological and theoretical impact. The field of economics has, in the last five decades, developed a formal language to describe society and analyse the data it generates. The combination of formal models to describe individual and aggregate behaviour has been a big advantage to the field when dealing with data. A rich formal toolkit has developed to tackle issues of causality

and integrate theoretical ideas regarding the function of society with empirical work. The Nobel prize 2019 was, for example, awarded to a team of three economists for their experimental approach to alleviating global poverty.

However, *this toolkit is challenged by the wide adoption of AI methods* like ML and causal inference and the increased availability of large quantities of data which are often unstructured, i.e. text, sound and images. The use of *artificial agents in multiagent systems* could provide new simulation-based models which take seriously the heterogeneity of individual preferences and information sets to model the role of governance, values, culture, path analysis and other emerging collective phenomena.

But AI has also adopted concepts, methods and paradigmatic examples from economics into its mainstream toolbox. The most significant examples are *game theory, social choice and voting, together with mechanism design*. These disciplines have not only had a profound impact on the foundations for AI models of rationality and social coordination, but have also provided standards of rigour and test-cases. Moreover, these disciplines have had an enormous impact because of their *systematic use in AI-based applications*. Work on causal inference in econometrics could provide a useful input into the debates around this key issue in ML.

The economic impact of AI. AI affects the economy in many ways. The increased use of AI leads to *productivity gains* in economic activity in general (through the automation of processes and a reshaping workforce capabilities). We will see a wave of *innovation in products and services* that foster labour productivity (like robots, software, data services). But these positive gains will also be linked to shifts in economic power. Automation will change labour market dynamics by destroying some routine jobs and creating new jobs in what is known as job polarisation. Administrative, manual, or repetitive jobs disappear as they are replaced by machines. Second, AI raises the demand for well-educated workers with technical degrees, while reducing the demand for low education workers creating a skill gap. Third, the presence of people and firms with *data science* (DS) capabilities has an impact on national competitiveness. The importance of data for supervised ML means that those companies and governments with access to large amounts of data will advance faster in the development of AI. These considerations have important implications in terms of market regulation, anti-trust and trade.

The Impact of AI in Political Economy: The economic changes resulting from the deployment of AI solutions will increase social and economic inequalities and an imbalance of power between those individuals, countries and regions active in that deployment compared to those lagging behind. Society will need to be able to compensate the losers of the possible economic changes implied by the adoption of AI. Political institutions will need to adapt the rising importance of recommender systems for politics in social media and the further concentration of opportunities, wealth and income. Higher education institutions have a critical role to play in equipping students with the personal and professional competencies (kind of knowledge, skills and attitudes) and capabilities (relevant to active students' engagement in society) that will be in demand in the face of the proliferation of AI and DS. The role of social media for political dialogue needs to be understood particularly well in Spain - a country with a long history of political polarisation (dos Españas).

3.4. Social Science and Society

New metaphors and tools for socio-cognitive rationality and collective action. With the evolution of multiagent systems,there was a gradual recognition of the need to differentiate agents and the social environment where agents interact as two different first class entities (see Chapter 2). This distinction feeds theory and intuitions from the social sciences into AI and, in turn, AI research in these topics has enabled several developments that are relevant for social psychology and sociology. For example:

- Agent models and architectures that account for socio-cognitive rationality (self image, mental models, awareness of the environment, attention).
- Governance devices that articulate interactions in the social space including norms and norm enforcement, but also social norms, organisations, team-work.
- Tools and artefacts. Social coordination conceptual frameworks, methodologies and support tools for design and deployment of *socio-cognitive systems* and *hybrid* online social systems (human and artificial agents).
- Methodologies and tools for network activity analysis, sentiment analysis, text mining, semantic searches...
- Crowd-based technologies: for collective epistemology —like knowledge aggregators, recommender systems, opinion markets—, for problem

decomposition-integration —web surveys, crisis mapping—, for collaborative work —education, constitutional reforms—, and so on.
- Formal treatment of collective action and coordination.
- Agent-based simulation of social phenomena.

Research and development domains in massive on-line AI-powered social interactions. For example:

- The emergence of a research space in social life mediated or supported by social networks and online communities (learning, travelling, gaming, gambling, dating).
- A new character of public opinion: modes of polarisation, information silos; new rhetorics (fake news, false authority, search filtering); "influencer" roles and devices; profiling and micro-targeting, cascading messages through and across social networks and platforms (video and post sharing, re-tweets, whatsapp...)
- Evolution of social practices like entertainment, education, shopping, dating; and evolution of individual skills and habits associated with those web-based AI-powered activities (reading skills, summary representation, memory and information search, friendship networks, antisocial behaviour...)
- The sharing economy: peer-to-peer retail, lodging, transportation, time-banks, lending.

Algorithmic bias, risks of manipulation, social control, hidden agency and other unintended consequences. Despite all the advantages that the large scale application of ML is bringing to digital services, concerns have emerged around the unintended or perilous consequences of automation on online platforms and the opacity of AI systems (Pasquale, 2015). Recommendation systems can reinforce established opinions or reinforce polarised views by presenting to the users information consistent or similar to their preferences and preventing them from being exposed to dissonant information. Search engines, for example, have been criticised for enabling the creation of filter bubbles around users. This type of phenomenon—also known as echo chambers—can be exploited by malicious actors as part of disinformation campaigns. Face recognition systems may be quite useful for security purposes but may rise privacy concerns and abusive control from authoritarian governments. The use of traces of personal web activity (searches, purchases, mobility...) is used to draw a digital profile of individuals that may be used for many different purposes, not always beneficial for the individual lawful.

4. KEY CHALLENGING POINTS IN ETHICS, LAW, ECONOMICS AND SOCIETY

As suggested in the introduction to this chapter, from a social perspective, the ***ultimate challenge*** AI brings about is to solve its version of the Collingridge's dilemma:

How to anticipate the repercussions of AI and act timely in order to achieve the best and avoid the unwanted. In fact, this dilemma may be unfolded in two components: one epistemic, the other ethical, as follows.

The ***epistemic*** component is:

C_1: *To determine what are the repercussions —actual and potential— of the developments of AI.*

The ***ethical*** component (C_2) is:

C_2: *To strive for the best outcomes from AI.*

The first component (C_1) is essentially linked with the different technologies involved in AI. Much of what is discussed in the rest of this book provides indication of what the outcomes of AI research and development may be. However, the repercussions of current and future AI, though, is a matter more of the humanities and the social sciences and addressed in the challenges of this section. The ethical component (C_2) involves two aspects. First, one has to postulate a notion of "good" with respect to which repercussions are valued. As mentioned in the introduction of this chapter, "good AI" has been characterised as trust-worthy, human-centred, responsible. They all share the main concern of caring and protecting human beings. Minimisation of harm, protection of vulnerable individuals and groups and the achievement of sustainable development goals and public well-being, for example, fall within the scope of this concern. But there are several other values. The point is that several values —often conflicting— may be involved in assessing whether a repercussion of AI is good or not, and to what degree.

The second aspect is how to go about in order to achieve good repercussions. In this case two complementary perspectives are involved: First, choosing among potential courses of action —a matter of value-driven behaviour— which can either be learned or imbued. Second, promoting some courses of action and preventing others —a matter of normative systems or policy. This decomposition of the ultimate challenge still hides a fundamental question:

What differentiates AI from other disruptive technologies in a fundamental way? We argue that the answer is ***artificial autonomy.***

It is usually agreed that the ostensible objective of AI is to model human rationality and embed it into artefacts. Sometimes this objective is met by automating specific tasks that typically involve human intelligence in some form, like playing chess, trading stock, face recognition, medical diagnosis. Some other times the goal is not just to automate some activities but to mimic the cognitive processes involved, like natural language understanding, generalised pattern recognition, learning, planning or automated inference. And in both approaches, AI has been quite successful in its six decades of existence. However, some AI researchers discern a more elusive goal: to build artefacts endowed with *general intelligence*. That is, artefacts whose competence is not limited to specific tasks, traits or skills but that they possess an intelligence that, like human intelligence, allows them to act successfully within the ever-changing circumstances of a rich environment. But this capabilities entail that the artefacts must be purposeful, self-conscious, creative and, ultimately, *autonomous*.

It is debatable whether general intelligence, and therefore "true autonomy" will or should ever be feasible in artificial systems. Nevertheless, from the skill gap induced by sophisticated robots, the perverse use of micro-targeting in political campaigns, the need for automated self-censoring of twitter, to the advent of self-driving vehicles and killer drones, it is undeniable that artificially intelligent entities exhibit some form of autonomous behaviour. Hence, a third key component of the ultimate challenge for AI is:

C_3: *To develop means for harnessing artificial autonomy*

The next paragraphs translate these three components into concrete challenges for each of the four areas.

4.1. Challenges in Ethics

1. **Embedding ethics in AI systems** (*The value alignment problem*). How to build autonomous artificial entities that are probably aligned with human values. How to embed social values in the design and development of technologies? What kind of new human rights are related to AI development? How could one ensure a safe human-AI interaction? sub-challenges
   a. ***Development of guidelines and standards.*** Soft approach to imbue values in autonomous artificial entities.

b. ***Value-driven behaviour in artificial entities.*** Conceptual frameworks and tools. (i) *Individual behaviour:* agent architectures; reasoning about values, individual values, context-dependent values. (ii) *Social enforcement*: the role of norms and incentives; social norms and social environment; paradigms and framing. (iii) *Value assessment*.

c. ***Provably reliable ethically-aligned behaviour.*** Formal techniques and devices for characterising, testing and ensuring that the behaviour of artificial entities is aligned with given values.

d. ***Paradigmatic cases.*** (i) *Building socially responsive artefacts*. Evolution and adaptation of value-driven systems (see Chapter 2). (ii) *Policy-assessment systems.* Interplay between individual value-based behaviour and value-alignment of emergent social behaviour. (iii) *Value-sensitive applications* like self-driving vehicles, public surveillance, uses of health records, screening and micro-targeting.

2. Embedding ethics in engineering. What issues in AI deserve an ethical approach? How to identify realistic expectations about AI achievements and their effects? Who should be aware of ethical issues in AI? What are the key contents and the means for fostering a realistic view AI? How to develop awareness to the ethical issues of AI? What tools are needed for ethics education for AI? How to expand the ethical literacy of engineers and society? What kind of narratives can we use? The point is, then, to design an agenda for developing critical —ethical, objective, realistic— literacy of AI in society and research.

There is an acute need for education on the ethical aspects of AI. As stated in the IEEE document (IEEE, 2019)"*...the key is to embed ethics into engineering in a way that does not make ethics a servant, but instead a partner in the process. In addition to an ethics-in-practice approach, providing students and engineers with the tools necessary to build a similar orientation into their inventions further entrenches ethical design practices*". The challenge is in fact that engineers and practitioners come together with social scientists and philosophers to develop contents for such literacy, for example, case studies, interactive virtual reality gaming, narratives (science-fiction), and additional course interventions that are relevant to students, professional and society at large.

Core outcomes of the agenda should be:

a. ***Guidelines and teaching syllabus for ethically aware AI R&D.*** It is essential to offer ethical background to students and professionals in technological fields. The ethical dimension of AI must be part of the formative curricula of engineers, computer scientists, and other

specialists involved in the design and development of AI and intelligent robots. The standard curricula and delivery should be complemented with non-classical materials like quality science-fiction, serious games and the like.

b. Unless this education is effective, the abuse of ethics expertise and research by Big Tech Companies is likely to effectively stifle government regulations. Thus, it is crucial to preserve independent and public ethical research and education on AI.

c. ***Guidelines and curriculum development to foster critical awareness of AI.*** For IT AI experts, IT professionals and schools that includes ethical awareness as well as critical understanding of AI risks and promises.

d. ***Establishing an observatory of ethics in AI.*** A systematic analysis of the ethical impact of AI and the means to foster ethically desirable outcomes.

3. Characterising "moral agency" in artiftcial entities. Could one qualify artificial entities as agents with consciousness, intentional mental states, or indeed moral agents, with ability to engage in moral judgements. Is it feasible to provably imbue values in artificial systems and make those values operational?

The goal is to develop the theories, formalisms and behavioural standards that clarify the extent and shapes to which moral agency can be predicated of an artificial entity.

Such development involves:

a. ***General AI.*** What type of autonomy in artificial entities can be achieved without presuming general AI? What are the constitutive assumptions of general AI? Is general AI an achievable quality for an artificial entity?

b. ***Values.*** In what sense one may claim that ethical theories apply to AI systems. How to make values operational —observable and commensurable, as well as an implementable cognitive construct— in artefacts. Value taxonomies and their relevance. Means to imbue values: norms, incentives, codes of conduct, education and internalisation of norms.

c. ***Artificial autonomy.*** Examining —under the light of artificial agency— moral issues that are meant to be made operational. For example, damage, accountability, responsibility and moral judgement.

4.2. Challenges in Law

1. **Developing "robot laws".** Adapting legal conventions in order to make them applicable to artificial entities as well as to systems and situations that involve them.
   a. ***To develop a strategic taxonomy*** of relevant legal issues, subjects and domains for regulating activity involving artificial entities.
   b. ***To explore the "legal personality" of artificial autonomous entities*** and to identify consequent responsibilities, and thus liabilities, associated with that personality.
   c. ***To re-cast crime, tort and misdemeanour in AI.*** Specifically, the key assessment of features—accountability, responsibility, guilt and liability—and the key handling of processes—infraction, detection, charging, judging, blame assignment, reparation—, need to be performed in AI-assisted socio-technical systems.
   d. ***To address AI-enabled misconduct on digital platforms and its policy implications.*** In order to facilitate *legal appeal and redress in automated decision-making* and *third-party evaluation of AI systems performance and accuracy*, an analysis of instances of misrepresentation, unfair micro-targeting, automated and AI-driven forms of social control and nudging, or the improper delegation of entitlements and responsibilities, should be carried out.
2. **Developing governance technologies for social interactions in augmented reality.**
   a. ***Normative frameworks*** that apply to artificial entities, and to hybrid populations.
   b. ***Normative languages*** that allow back-and-forth rewriting processes from natural language, to legally precise expressions, to machine readable representation. And the machine readable representation of a normative framework is amenable to automated compliance, enforcement and evolution.
   c. ***Institutional frameworks*** for hybrid systems involving artificial autonomous behaviour.
   d. ***Anchoring*** hybrid online social-coordination systems on the actual socio-legaltechnological environment.

4.3. Challenges in Economics

1. **To develop a methodology to fill the gap between economics and AI methodologies for research.** There are three clear areas of development:

a. ***Integrating unstructured data into economic models.*** It will be relevant to explore the development of AI technologies for *feature extraction* from unstructured data available through *text, sound and images*. ML has led to a revolution in this regard and has made it possible to explore entirely new sets of data like TV shows, social media content or satellite imagery. However, it is unclear how to integrate these methods for data extraction into economics models and help address socially important questions.

b. ***Agent-based modelling.*** There are three main areas of potential co-development and some obvious research topics of common interest: (i) *Agents* models of rationality; incentives and motivations; agency and delegation; deductive rationality vs social rationality. (ii) *Multiagent systems* including topics like mechanism design, collective decision making, negotiation and contracting, online institutions, organisational theory, governance. (iii) *Agent-based simulation*. Microbehaviour and collective outcomes, population dynamics, governance, values, culture.

c. ***Experimental economics.*** Design methodologies and best practices for experiments involving a mixed population of humans and artificial agents. Develop protocol standards and specification languages for the definition, implementation and analysis of online experiments with hybrid subject populations.

d. ***Developing a common frameworks for causality and policy interventions.*** Economics has a well-developed econometric toolkit for the identification of causal effects. Currently there is a rival system under development outside economics and, while it is clear that the new causal inference has big formal advantages, there is a danger that methods will be re-invented. On the other hand, multiagent systems and online institutions have approached notions of collective action, governance and policy-making with a perspective that has not permeated into economic theory yet.

e. ***Improving forecasting accuracy.*** Forecasting economic outcomes has a long tradition in economics and is mostly model-based in the field. This approach clashes with the use of ML as the latter provides little scope for integrating policy decisions. However, important organisations like central banks and financial institutions rely on forecast models to make decisions. A new kind of hybrid modelling is needed to accommodate these uses.

2. **To develop a theory for the new AI-enabled commons.** Notions like moral hazard, social choice, behavioural economics play an increasingly important role in AI research. The outcomes of such use should influence economics research but have not been taken up by economists yet.
3. **To identify and explain the economic and political economy repercussions of AI.** In three main directions:
   a. ***Shaping the economic impact of AI and compensate losers.*** Research into the labour market repercussions of the adoption of AI is now common. However, there needs to be a better understanding of the economic impact before technologies are broadly adopted to understand where and which adoption is desirable and who will be the losers from it. This requires broad thinking which combines both research on the labour market impact, allocation of capital across firms and the social and cultural shifts that this brings.
   b. ***Exploring the unintended consequences and externalities of the digital information economy.*** The rise of social media systems has given tools like recommender systems and micro-targeting enormous influence on individual behaviour and welfare. Especially the younger generation is increasingly dependent on social media. The welfare effects of this and the scope for policy interventions can only be understood through multidisciplinary research.
   c. ***Understanding the impact of AI technologies and platforms on the social contract and the functioning of political institutions.*** The concentration of political and economic power in the firms controlling AI systems and the rise of social media as a way to gain information provides a challenge to political institutions. News get filtered through automatised training and the algorithms employed by powerful firms have a direct impact on political debates. Social media and AI-enabled online interactions are transforming the notions of *accountability, representation and responsiveness* of public officials and institutions, on one side and, on the other, the role of individuals and organisations in democratic deliberation, democratic agreement, and public opinion.

4.4. Challenges in Society

1. **Drawing a road-map to develop the use of AI for human flourishing and co-evolution with artiftcial systems.** From teleworking to distant learning, social activity takes place in virtual environments. The

coexistence of humans and AI supported systems and agents require the mutual adaptation of humans and machines toward co-evolution. Artificial entities have capabilities, skills and entitlements that may be designed and deployed with human flouring in mind. To ensure that the large scale adoption of artificial agents and robots create positive effects on the economy and on society, machines and humans need to grow together and learn from each other.

2. **Developing a conceptual framework for the implementation of socio-cognitive agents and socio-cognitive technical systems.** On one side, the challenge requires to address the ***psychological, sociological and computational aspects of artificial rationality social behaviour,*** which involves individual rationality (introspection and the interplay between beliefs, desires and intentions) as well as social rationality (that is, expectations about the behaviour of other agents or about the impact of one's behaviour on the behaviour of others and the possibility of affecting and being affected by other agents or by the social environment where interactions take place). On the other side, the challenge demands the ***construction of online systems including socio-cognitive agents,*** which may be both artificial and human.
3. **Identifying criteria and specifying indicators to evaluate emerging effects of AI on society.** There is a growing need for empirical research to better understand users' demands, values and needs, and also to unveil unexpected consequences of AI-powered human and human-machine activity. To promote a positive collaboration between humans and AI machines and their healthy coexistence, we need to further study how humans interpret and react to AI-assisted systems, and assess the extent to which they may be vulnerable to be deceived or misled by autonomous machines and AI-enabled social interactions.
   The topics to study may be organised around three main lines:
   a. ***Evolution of individual cognitive practices:*** attention, search, recall, collaboration, motivation, locus of control, self-perception, dignity.
   b. ***Evolution of AI-enhanced social practices:*** epistemology, collective agency and coordination, social care —for risk and disadvantaged population, like disabled, elderly, minorities, children, migrants, displaced—; education, health, gaming and gambling; sex and group entertainment; political activism; social control.
   c. ***Unintended consequences of AI:*** biases, discrimination, abusive microtargeting, hidden agency, political manipulation, unfair profiling, misappropriation of personal web traces, and so on.

This challenge requires the combination of small scale qualitative research with large scale quantitative studies. Through both large-scale field experiments exploring the use and adoption of AI-based recommendation systems we can better understand users' requirements and pitfalls. Small-scale laboratory experiment, in contrast, may help shed light on perceptions of vulnerable groups, such as kids, elderly or disabled persons interacting with robots or other autonomous machines.

CHALLENGE 7

LOW-POWER SUSTAINABLE HARDWARE FOR AI

Coordinators
T. Serrano (IMSE-CNM, CSIC – US)
A. Oyanguren (IFIC, CSIC - UV)

Participant researchers and centers
J. Villagrá (CAR, CSIC – UPM)
J. J. Hernández Rey (IFIC, CSIC – UV)
L. Fiorini (IFIC, CSIC – UV)
A. Argyris (IFISC, CSIC – UIB)
C. Mirasso (IFISC, CSIC – UIB)
F. Campabadal (IMB-CNM, CSIC)
J. M. Margarit (IMB-CNM, CSIC)
L. Terés (IMB-CNM, CSIC)
M. Delgado-Restituto (IMSE-CNM, CSIC – US)
B. Linares-Barranco (IMSE-CNM, CSIC – US)

1. EXECUTIVE SUMMARY

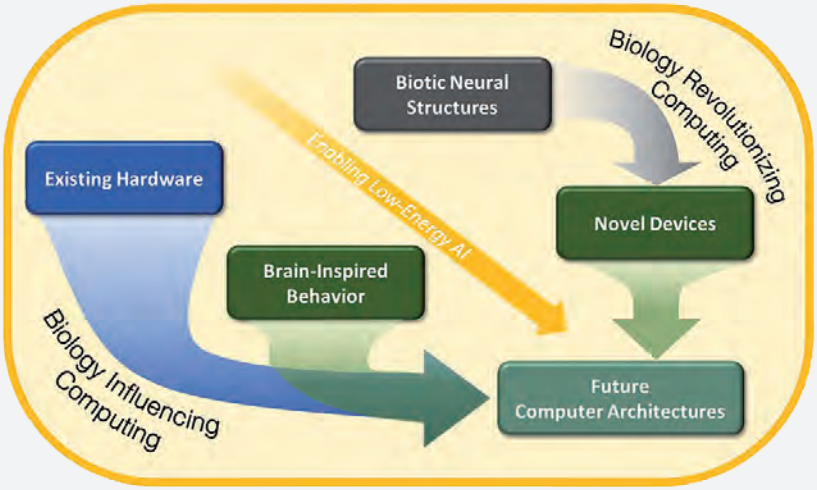
The R&D of high-performance unconventional hardware aims to implement efficient low power high-speed AI systems. This line of work is required to achieve sustainable AI systems and to develop new applications that due to the requirements of high-speed response and power constraints cannot be implemented with the current hardware solutions.

Current AI systems are typically based on running heavy computation algorithms on big powerful remote servers. Both the communication of data to and from the servers and the specific computation consume a considerable amount of energy. To make AI part of our future daily life in a sustainable way, research should be done to move the computation to edge power efficient processors.

Biological brains are a model of natural processors exhibiting an astonishing ability to implement cognitive algorithms with *low energy resources* and *high speed response*.

New bioinspired hardware architectures for AI emulating the computational principles and architectures of biological brains should be developed. The next generation AI hardware should be specifically tailored to run the AI algorithms at the edge in real time and in an energy efficient way.

FIGURE 1.1–Falta



The high-speed sustainable hardware should be based on advanced technology combining existing CMOS technology with new emerging devices as memristors and photonic devices. This beyond CMOS technology will make possible to achieve efficient highly parallel architectures of neural computing units highly interconnected though adaptable synaptic devices with long-term memory. The close integration of memory and computation saves data communication energy and improves the computation speed.

The high-speed sustainable hardware for AI have the potential to make a real breakthrough in a plethora of everyday applications requiring high speed cognitive processing of sensor data and decision taking with an affordable energy budget such as: autonomous cars, autonomous drones, robotics... Other application environments such as wearable medical devices, industrial production, visual inspection of production lines, or surveillance can benefit from this technology. High-speed analysis of *big data* (BD) can also impact and may allow scientific discoveries in basic fields like high-energy physics or astronomy.

CSIC combines experts in designing brain interfacing systems, unconventional neuromorphic hardware architectures, neural processing photonic systems, beyond CMOS technologies, and real-time processing of massive data ***that***

can jointly work in an effort to achieve beyond state-of-the-art AI hardware systems and AI applications specially suited for fast BD analysis.

2. INTRODUCTION AND GENERAL DESCRIPTION

The achievements of AI systems have been outstanding, outperforming humans in some cognition tasks such as the well-known Jeopardy contest where the IBM Watson computer (Ferrucci et al., 2010) defeated its human competitors with almost 40-fold difference in reaction time. However, *human brains still largely outperform the most advanced supercomputers when we compare physical size and energy efficiency.* Coming back to Watson's competition, Watson had 2880 computing cores, occupying the equivalent physical size of 10 refrigerators, and consumed 80 kW, while human brain occupies less than 2 liters and consumes 10-25 W.

In recent years, we have witnessed the growth of AI applications and algorithms. Typically, these algorithms require a heavy computational load and huge data storage which is performed in remote servers on the cloud. Consequently, these systems require an intense data communication between the individual front-end devices running the application and the cloud servers performing the computations and storing BD. These internet communications consume a great amount of energy. As an example, just two google searches emit the same CO_2 amount than boiling a kettle (Warman, 2009). It has been predicted than following the current trend, the internet communications will consume 20% of the world's electricity and will generate 5.5% of the world's carbon emissions by 2025 (Vidal, 2017), and by 2030 the communication technology will consume 51% of total electricity and will contribute up to 23% of the global gas emissions (Andrae and Edler, 2015). Currently, there are already important efforts to move part of the computation from the cloud to embedded AI processors to be used as 'edge' computing devices (Chen et al., 2019).

Furthermore, the big super computation servers are based on traditional Von Neumann computer architectures where a high-speed computation unit sequentially performs operations on data that are read from a separated memory unit. Data have to be transferred to and from the memory at high rates, consuming a high amount of energy in the data transference. This Von Neumann bottleneck, also known as memory bottleneck, is accepted to be one of the main limitations of the performance of conventional processors when

running artificial neural networks algorithms (Backus, 1978) . In this sense, the use of dedicated AI low power parallel computing hardware accelerators (Edge TPC, 2019), (Servethehome, 2019), (Extremetech, 2019), (Kim et al., 2019), (Lee et al., 2019), (Cho et al., 2019), (Sano et al., 2020), processing in memory architectures (Li et al., 2016), or commercially available parallel hardware platforms such as *graphic processing units* (GPUs) and *field programmable gate arrays* (FGPAs) have been proposed to implement more efficient AI processors.

The astonishing ability of *biological brains* to solve cognitive problems using low-power low-speed noisy computational neurons has motivated that engineers had looked for inspiration in biology to look into ways of achieving efficient cognitive systems. The brain is very different from conventional computing systems in terms of technology, architecture, as well as signal processing and coding. It is composed of slow components (neurons time constants and synaptic delays are in the order of milliseconds [$10^{-3}s$]) compared to CMOS technology (with operating frequencies in the gigahertz range [that is, $10^{-9}s$]). The brain components suffer from high noise and high variability among them, whereas digital computers operate on high precision digital numbers. However, the on-line plasticity of the devices allows compensating variability and faulty components. The implementation of on-line adaptation in CMOS technology consumes many resources. Alternative beyond CMOS technologies like memristive nanodevices exhibit the capability of on-line adaptation following biologically inspired learning rules while are able to implement very compact memory with no leakage (Linares-Barranco and Serrano-Gotarredona, 2009; Camuñas et al., 2019). Thus, dedicated AI processors combining CMOS technologies with emerging devices (memristors, spintronics, etc.) are being investigated.

Opposed to the high-speed sequential computing paradigm of digital computers, human brain is a highly parallel architecture composed of a high number of parallel processing units or neurons (in the order of 10^{10} 10^{11} neurons) operating at lower speed. Furthermore, the neural processing units are highly interconnected through synaptic weight connections (in the order of 10^{14} synapses) which have an on-line adaptive memory. So that, opposed to what happens in conventional computer architectures, memory and computation are closely interleaved in biological brains.

Not only are there architectural differences between brains and conventional computers, but the signal coding and processing principles are also quite

different. In standard computers as well as in conventional AI systems, data are sampled at periodic time intervals and the recognition algorithms are applied sequentially to these sampled data. For example, in conventional AI vision systems, the visual input is a video sequence which is composed of a sequence of static representations of the visual scene or frames which are acquired at periodic time intervals (typically 20-30*ms*). In this conventional AI systems, also known as *artificial neural networks* (ANNs), time is not an explicit variable but it is implicitly contained in the sampled data (Ghosh-Dastidar and Hojjat, 2009). However, in biological brains data are communicated and processed as asynchronous electrical pulses or spikes. In the retina, there are no frames; sensed data are represented as the asynchronous occurrence in time of a flow of spikes that are communicated through the optic nerve along time. Spikes generated in retina travel through the optic nerve and are processed and propagated asynchronously by all the layers in the visual cortex. It has been demonstrated (Thorpe, Fize, and Marlot, 1996) that recognition of a familiar object occurs in the upper layer of the cortex with just the delay of a single spike front traveling through all the cortical layers. A class of AI systems or neuromorphic systems where signals are represented by spikes have emerged and are nowadays also known as *spiking neural networks* (SNNs) or the third generation of neural networks (Maass, 1996; Ghosh-Dastidar and Hojjat, 2009). In this kind of systems, computation is no longer driven by a periodic sampling clock, but computation is driven by the occurrence of spikes resembling the spike-driven computation of biological neural systems (Farabet et al., 2012). It is believed that the parallel computation, the close interaction between computation and memory, an extremely efficient information coding, and on-line adaptation are some of the clues of the high efficiency of biological brain computation.

Different approaches at different levels of abstraction have emerged to design bioinspired low power dedicated hardware to implement more efficiently different aspects of the complex perception, cognition and actuation abilities of the biological neural systems. The different specific low power AI dedicated hardware approaches can be complementary with conventional digital AI systems to optimize the efficiency and abilities of any specific AI system depending on the particular application.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

The development of low-power edge computing dedicated smart sensors and new processors will potentially impact the emergence of many new applications and more efficient solutions of intelligent systems in many application areas. Some of the areas where low power AI hardware can have an important impact and where CSIC can contribute based on the expertise of its research groups are presented in the following.

3.1. High-Performance Energy-Efficient Smart Systems: Autonomous Driving, Drones or Robots

The development of intelligent autonomous systems such as autonomous vehicles or robots demands high-speed low power acquisition and processing of massively parallel input data and high-speed real-time generation of decision and control signals. The high-speed demand is specially challenging in applications such as drones where high-speed reaction is a must. Low-power operation is specially demanding in drones or robots that must operate in remote or inaccessible areas.

This demanding specifications cannot be covered with conventional computing systems. Neuromorphic engineering sensors and processors are being demonstrated able to efficiently acquire massive data and perform high-speed extraction of features using low-power budget. Furthermore, neuromorphic engineering systems implementing spatiotemporal data coding and learning algorithms may lead to compact real-time efficient implementations of recurrent neural networks which are currently very computationally demanding. To obtain energy efficient and high-speed systems, the neuromorphic engineering computation principles can be combined with the use of novel emerging technologies to further improve the system performance. A possible approach is to combine the electronic computation technologies with photonic technologies.

Despite the improvements in the communication networks and the supporting dense computing technology, data acquisition and signal processing are currently firmly dependent on the electronic infrastructure. Even though most of the global data traffic flow is routed nowadays through fiber-optic channels as optical signals, the intelligence that is applied in the optical domain to make any kind of processing is minor. Transferring even a small part of intelligent computations to the optical domain can reduce the unsustainable energy footprint of currently implemented concepts, training methods and AI chips. Moreover,

the computational speed has the potential to be increased dramatically, limited only by electrical bottlenecks introduced by optoelectronic conversions. In addition, this may impact a plethora of applications through:

- the increase of the *transfer rates* and *response times* from and within data centers and the incorporation of AI concepts and techniques on-the-fly in the optical domain,
- the *incorporation of dedicated hardware* AI implementations to the next generation of low power photonic integrated circuits for information processing.

In the case of autonomous driving, the biggest challenge to massively deploy autonomous vehicles is finding solutions that address a high degree of uncertainty in highly complex environments. Both artificial perception and decision-making aspire to respond to this challenge and have a critical cross-domain element for the appropriate management of a wide variety of situations: *machine learning* (ML). However, to embed this kind of software approaches, often too computational-intensive, into on-board safety-critical computing devices, which have to face challenging energy and space requirements, motivates the rise of new hardware paradigms. Indeed, ***the advent of powerful multi processors system-on-chip, combining different processing technologies, such as GPUs, FPGA or neuromorphic chips, will enable affordable fault-tolerant hardware architectures.*** As a result, more dependable systems will be available, thus inspiring user trust and engagement, the most difficult hurdle to make this new mobility paradigm a reality.

3.2. Real Time Analysis of Big Data Scientific Systems

Basic science systems where a huge amount of data hit detectors have often to be quickly analyzed for possible storage or for alert. Nuclear and high energy physics, astrophysics and cosmology can greatly benefit from the development of algorithms and ML tools in specific hardware for fast event reconstruction, discrimination and selection. The availability of dedicated hardware able to perform real-time ML on BD may allow scientific discoveries in the cosmology field, discovery of rare and new events in particle physics or even alert to the presence of hazardous events (asteroids). In a similar way, we foresee potential applications in many other basic scientific fields facing the problem of real time BD analysis.

3.3. Monitoring Systems for Health and Security

The development of edge *smart multisensor-fusion* systems can be applied to miniaturized, adaptive instrumentation for real-time industrial/environmental monitoring in order to reduce maintenance and analytical costs, offer long-term stability in front of continuously changing environmental conditions (e.g. temperature, interference effects, sensor drifts), and provide immediate assessment in critical situations like terrorist threats or contamination. This technology will also impact the development of wearable, cognitivebiomonitoring devices for personalized diagnosis and preventive health care, providing continuous multimodal monitoring of individuals in natural environments and therefore allowing to study how dynamically-changing markers can be correlated with their physiological states. Brain mapping and brain-machine interfaces closed-loop neural implants are a promising solution for the treatment of neurological diseases (Bergey, 2015; Zimmermann and Jackson, 2014), as well as for the implementation of sensorimotor brain-machine interfaces. Neural signals captured by the implant are internally processed by means of AI techniques and, based on the outcome, stimulation patterns are triggered either for ameliorating the impact of the disease (e.g., by stopping uncontrolled epileptic seizures) or for restoring lost senses after a neural injury. Signal processing in neural implants should, on the one hand, reduce dimensionality and extract features able to provide clinically relevant information (Yoo et al., 2013) and, on the other, exhibit low power consumptions to not exhaust the presumably short energy resources available. Among the many different operators which have been proposed for the extraction of neural features, those suitable for the quantification of brain functional connectivity are gaining growing interest (Sakkalis, 2011). *Functional connectivity* refers to the statistical evaluation of coupling strengths between brain regions to identify those involved in sensory responses, motor activity or intellectual/emotional processing. Functional connectivity can be assessed on neural signals captured at different spatial resolutions: from singleunit responses acquired by micro-electrode arrays to aggregated signals obtained through electroencephalography or functional magnetic resonance imaging measurements. Besides providing means for brain mapping, clinical studies have demonstrated that functional connectivity also gives relevant information to distinguish between normal and pathological brain states (Bruña, Maestú and Pereda, 2017). Indeed, it has been shown that abnormal synchronization patterns are associated with different neurological disorders, such as, epilepsy (Jiruska et al., 2012), Alzheimer's disease

(Knyazeva et al., 2012), Parkinson's disease (Schnitzler and Gross, 2005) or schizophrenia (Uhlhaas and Singer, 2010).

The referred applications in neurophysiology, disease monitoring and therapy, and brain-machine interfaces suggest the development of dedicated integrated circuits for AI processing based on functional connectivity (Delgado-Restituto, Romaine, and Rodriguez-Vazquez, 2019); ***these applications would be especially valuable in scenarios demanding power efficient solutions*** (e.g., implantable neural prostheses), fast on-site operation to avoid neural activity transfer, and off-line computations by a host computer (e.g., neurostimulators for seizure abortion).

3.4. Emerging Memristive Technologies

As previously mentioned, one shortcoming of conventional computing systems is the separate implementation of the memory and computing units and the time and power consumption required to interchange the data between the memory and the computation part. This makes this conventional systems particularly unsuitable to implement neural networks systems.

In neural networks systems, the synaptic elements are in-memory computing elements, which compute the modulated signal transferred between neural units and at the same time store the knowledge of the system. Furthermore, the synaptic devices have to be massively implemented to achieve a global distributed knowledge of the system. The electronic implementation of synapses is still a challenge for the scientific community.

Memristors are two terminal devices whose resistance changes as a function of the current or voltage applied to their terminals. Memristors devices are probably the best placed to replicate neuronal synapses. This is due to two reasons; the first is the capability of these devices of modulating their electrical resistance thus emulating adaptable connectivity between neurons. The second reason is their small size, few nanometers, which allows many devices to be integrated in a very small area, allowing the number of connections between neurons to be comparable to that of a biological system.

However, as the *memristive technology is still emerging* and several issues concerning reduced variability, low energy operation, device endurance, and current limitation to avoid device damaging have to be addressed to push them into a mature technology that can be densely integrated with CMOS to achieve tightly coupled dense in-memory computing devices. To address this

challenge different approaches are being explored, either in terms of device operating conditions or in the optimization and design of new materials combinations. The memristive technology will have impact on:

- *Electronic synapses in neuromorphic systems.* The capability of tuning the memristor resistance together with the potential device scaling of metal-insulator-metal structures in a crossbar configuration are features that are promising for these applications.
- *Non-volatile memories* The scalability, endurance and CMOS compatibility of resistive random access memories based on filamentary resistive switching devices are being considered as potential alternatives to current flash memory technology.
- *Security applications.* The cycle-to-cycle and the device-to-device variabilities are being explored for physical unclonable functions for hardware security applications.

4. KEY CHALLENGING POINTS

4.1. High-Speed Sustainable Hardware for AI

Recently, the availability of large amount of data, the increment of the computing capabilities of CPU processors and the use of high-performance parallel GPU units have motivated a rapid increase in the performance of automatic ML methods. Nowadays, conventional ANNs achieve impressive classification accuracy in tasks such as image recognition.

However, the realization of these classification tasks is very computationally and power demanding even for the most powerful CPUs. For this reason, the use of dedicated AI low power parallel computing hardware accelerators (Edge TPC, 2019), (Servethehome, 2019), (Extremetech, 2019), (Kim et al., 2019), (Lee et al., 2019), (Cho et al., 2019), (Sano et al., 2020), and processing-in-memory architectures (Li et al., 2016), have been proposed.

Furthermore, the interaction with the environment in real time is still very limited even when using hardware accelerators what limits its application in environments where high reaction speed is required like automatic driving or drone control.

On the other hand, SNNs that use a spatio-temporal coding of the signals while still achieving lower recognition rates compared with their sampled or frame-based ANN counterparts have demonstrated their ability to achieve much

lower recognition latency and energy than using conventional frame-based vision recognition architectures (Pérez-Carrasco et al., 2013). This high-speed capability and power efficiency performance has motivated emergence of vlsi implementation of complex large scale spiking processing and learning SNN architectures (Furber, 2016).

In parallel, novel neuromorphic sensors for vision, tactile, olfactory and audition have been developed. In particular, novel vision sensors based on sensing the illumination temporal contrast (commonly known today as *dynamic vision sensors* (DVS)) have reached a development maturity that has arisen the interest of the industry in the neuromorphic technology (Posch et al., 2014). These sensors have reached an integration density level (megapixel DVSs are commercially available), low interpixel variability, high-dynamic range (higher than 120 dB), and very low latency (lower than 1ms). Furthermore, the coding of the temporal changes compresses the information reducing the computing requirements of the subsequent processing system.

These developments of *neuromorphic sensors and processors* have risen the interest of the field in the last years as they are promising candidates to implement AI systems that require interacting with the environment in real time such as robotic applications, real time surveillance, autonomous driving... in a more efficient way. Recently, many neuromorphic technology spin-off companies have been launched (such as CelexPixel, Insightness, IniVation, GraiMatterLabs, Prophesee, BrainChip, CortexAI...) and other big companies such as Samsung, Sony, IBM or Intel have begun researching and developing the field.

One of the reasons why SNNs still achieve lower recognition performance than conventional ANNs is the lack of training methods as effective as *backpropagation*. The nondifferential nature of the spikes prevents the backpropagation rule do be directly applicable to SNNs. Other biologically inspired learning rules such as *spike-timing-dependent plasticity* (STDP) can be used for on-line training SNN systems but with lower performance results. However, backpropagation based trained ANNs are not appropriate to incorporate novel learning during operation and suffer from what is known as *catastrophic forgetting*. That is, when an ANN previously trained to recognize a set of objects is re-trained for recognition of a new object set, it gets a serious degradation in recognition of the former training set and one has to retrain it for the complete dataset (although sometimes it is sufficient to retrain the higher layers only). Recently, it has been demonstrated that combining a

supervised back-propagation trained ANN with an unsupervised STDP learning SNN classifier, robust learning avoiding catastrophic forgetting is achieved (Muñoz-Martín et al., 2019).

In this context, the development of ***hybrid ANN/SNN architectures exploiting and combining the best capabilities of each approach*** is a key challenging point to achieve high-speed reaction high-accuracy recognition systems with robust learning capabilities. The developed systems should be optimally partitioned depending on the target application to exploit the advantage of high-speed, natural processing of dynamic contents and temporal redundancy suppression of neuromorphic SNN systems with the advantages of more mature learning and computational algorithms of conventional ML systems.

In particular, this challenge requires research on:

- Development of system partition rules for SNN and conventional ANN systems. SNN are more efficient for massive data acquisition and feature extraction processing.
- Study of signal optimal conversion and interaction between SNN and conventional ANN systems.
- Learning algorithms for SNN exploiting spatio-temporal correlations.
- Development of efficient signal temporal coding in SNN to minimize redundancy in signal representations.
- Implementation of efficient hardware friendly on-line learning for SNN.
- Development of combined DNN and SNN learning systems.

One of the possible fields where ***hybrid SNN/ANN systems may have a breakthrough is in real time analysis of massive parallel data.*** The real-time analysis of massive parallel data is a challenging task for conventional sequential processing units even for the more advanced CPUs with many core architectures. This is the case for data coming from highly parallel sensors such as visual data coming from high-resolution vision sensors, data coming from large deployments of IoT sensors, or data for *high energy physics* (HEP) experiments of the *Large Hadron Collider* (LHC) at CERN. Nowadays, the parallelism of GPUs and FPGAs can provide a higher computational power than traditional CPUs and these platforms are currently used to implement ANN processing systems of highly parallel input data. Furthemore, in this kind of massive parallel input data systems there is a large data redundancy in space and time resulting in that only a small fraction corresponds to interesting signals. Trigger systems to decide which data have to be persisted making use of new advanced

technologies on computing accelerator platforms such as GPUs and FPGAs are currently used. SNN data processing techniques exploiting efficient spatio-temporal coding of the signal eliminating temporal redundancy in the data and exhibiting low-latency and low-power offer unique capabilities for trigger systems which have to do fast selection decisions on the data of interest. Conventional ML and ANN architectures and methods have been demonstrated to be very efficient in data selection and they can be deployed in the final stages of data processing. Nevertheless, despite the complexity, they are potentially very interesting to be implemented in specific hardware such as FPGAs. This work is starting *to be developed in the HEP field, where ML algorithms are being implemented in FPGAs to improve the performance of the current state of the art reconstruction* (Ortiz Arciniega, Carrió, and Valero, 2019). The experiments of the future high luminosity upgrade of the LHC will use ML algorithms in real time reconstruction accelerator systems such as FPGAs to achieve the expected performance of the experiments, whose trigger systems require the highest possible processing rate and bandwidth and can be implemented using SNN dedicated hardware or SNN architectures programmed in FPGA hardware. In this experiments, the large number of interactions per bunch crossing, happening at a rate of 40 MHz, will make very challenging the distinction of key signatures, which are usually recognized using topological characteristics of the event and kinematic properties of the reconstructed objects. Real-time particle track reconstruction will be crucial to perform fast selection decisions and to record potentially interesting data events for higher level of processing. This poses a major challenge because of ***the large combinatorial and the size of the associated information flow, requiring unprecedented massively parallel pattern-recognition algorithms.*** For this purpose, SNN neurobiology-inspired algorithms, such as Retina (Abba et al., 2016) are of great interest for this field. These SNN bioinspired algorithms can be programmed on FPGAs to offload the most repetitive and logically simple tasks of the trigger systems such as detector decoding, clustering and tracking reconstruction. In particular, GPUs with a many-core architecture for fast particle reconstruction, is also being proposed as a solution for real-time data analysis at some of the decision software stages (Halyo et al., 2013; Aaij et al., 2020)].

The following research should be developed to achieve the real-time analysis and reconstruction of this big physics data:

- Fast object reconstruction and pattern recognition for alignment and calibration of detectors in real time.

- Fast, highly parallelized, seeding algorithms which fulfil timing requirements in trigger systems.
- Fast and efficient data-driven discrimination for data storage.
- Fast and accurate sample classification for selection and online analysis.

Another application field than can benefit from the use of high-speed low power SNN pre-processing algorithms is the sensor fusion. *Multisensor integration* exploits the extended coverage of multiple detectors to increase perceptual confidence in smart systems (Hall et al., 2009; Margarit-Taule et al., 2019), but embedded implementations are yet in their infancy due to the lack of hardware able to infer from the multivariate, nonlinear, time-dependent and noisy signals supplied by modern sensors. Current commercial instrumentation is only able to analyze a small fraction of the markers targeted in everyday applications, generally employs one selective sensor per parameter, and requires periodical calibration in front of environmental changes and sensor non-idealities. Most analyses are still performed in laboratories, resulting in increased costs, hindered logistics, and delayed detection. By using principles of how biological systems promptly combine multisensory information and generate meaningful features in dynamic and uncontrolled real-world conditions, spiking neuromorphic networks are emerging as a powerful, VLSI-amenable computing paradigm to accelerate sensor fusion and enable continuous learning and context awareness under these constraints (O'Connor et al., 2013; Diehl et al., 2015; Li et al., 2019). Such networks can be embedded in smart systems to fuse multivariate data from a set of Si-based microsensors on the edge. The resulting cognitive multisensor system is targeted to be faster, smaller, energy efficient, and highly resilient to noise, nonlinearities, matrix effects, and drifts associated with the sensors. ***To achieve the target edge smart multisensor fusion system, the following research points should be addressed:***

- Accurate sensor modeling to adjust readout/processing circuit design to their static and dynamic characteristics.
- Definition of sensor fusion algorithms tolerant to sensor variability, drift, and crosssensitivities.
- Development of energy-efficient analog front-ends and analog/digital encoding-decoding architectures.
- Implementation of algorithms for incremental and local learning in embedded neuromorphic networks.
- Joint integration of different sensing technologies and VLSI readout/neural processing circuits.

4.2. Development of Next Generation Technology for AI

For efficiently implementing high-speed low-power systems that can perform intelligent tasks in a similar manner as human do and at the same time consume an affordable amount of energy and computing resources, ***hardware parallel architectures alternative to sequential Von Neumann architectures and overcoming the scaling problems of CMOS technology have to be devised.***

The development of low power highly parallel in-memory computing hardware platforms specifically tailored for the class of neural network computation and based on the hybrid combination of conventional CMOS technology and novel computing technologies such as nano memristive devices or photonic computing systems is a key challenge to achieve artificial systems that can emulate the real-time highly parallel computing capabilities of natural cognitive systems.

To develop such a system requires a jointly research effort addressing both architectural and technological questions.

At the ***technology level,*** one of the most promising paradigms to obtain dense neural networks with massive synaptic interconnections among the neural units is the combination of CMOS technologies with new memristive devices. Their existence was theoretically hypothesized by Leon Chua in the 1970s based on circuit theory and it was in 2008 when memristive behaviour was first demonstrated in nanoscale devices by William's group at HP Labs (Camuñas et al., 2019). They are right now the most promising candidates to implement low power in-memory computation and hybrid CMOS-memristive architectures have already been demonstrated using them as binary memory devices. In the neuromorphic engineering field, memristors have been considered as artificial synapses as it has been demonstrated that *when stimulated with electric pulses at their terminals, they exhibit a learning rule of their resistive state which resembles the spike-timing-dependent (STDP) learning rule observed in biological synapses* (Camuñas et al., 2019). Several on-line learning neural networks parallel hardware architectures have been proposed combining CMOS neurons interconnected with massive memristive devices.

However, most of the proposed systems are still based on software simulations whereas only very small demonstrators have been built and mostly based on a binary storage capability of the synaptic devices.

Another challenge that limits the dense integration between arrays of synaptic memristive devices fabricated on top of CMOS neurons and densely interconnected with them is the need of placing a current limiting CMOS transistor in series with every memristor (the so called 1T-1R structures) to avoid damaging of the memristive devices. The implementation of current limitation techniques using the same nanotechnology is series with each device is a must to achieve *high synaptic integration density*.

The are other processing steps in the memristive technology that should be optimized in order to increase the device yield and to improve device performance, such as retention time, resolution of memory states, noise, endurance and variability. In the case of filamentary resistive switching devices, set and reset switching processes are of a stochastic nature leading to a large device-to-device and cycle-to-cycle variability. These processes need to be fully understood for reliable applications. These emerging devices call for new characterization techniques to assess the device electrical behavior. The development of these new techniques will allow to gain a deeper understanding of the physical mechanisms responsible for the device behavior, and to extract parameters that will be necessary in the development of simulation tools for circuit designers.

Photonics is another beyond CMOS technology emerging in the field of neural network hardware implementation. The concepts of optical information processing and computing were introduced several decades ago, with a vision for solving image (pattern) recognition problems (Stroke, 1972; Abu-Mostafa and Psaltis, 1987). Since then, optical processing and computation have spread to different communities, providing revolutionary solutions to their problems, as well as leading to novel applications.

Key factor for this success has been the diversity of available building blocks that introduce computation functionalities in optics and photonics, with low-energy consumption and small footprint. Such attributes have been based lately on semiconductor devices, photonic crystals, fiber-optics, photonic integrated circuits or photonic nano-devices, providing great flexibility to select the most compatible and incorporate it in the various platforms of technological infrastructures (Minzioni et al., 2019). On the one hand, such building blocks can be used in on-demand designs of dedicated photonic neural network configurations that scale up computational capabilities. On the other hand, fiber-optic topologies—supported by a mature technology from the telecom industry—provide a huge flexibility to experiment and test numerous, but relatively simple configurations that addressed diverse classification and time-series prediction

problems. For example, the transfer of reservoir computing concepts in all-optical, electro-optical and photonic integration hardware has been a successful paradigm of the last decade, for fast and efficient task-solving computation.

Regarding the development of photonic neural processing systems the following research points should be addressed:

- Introduce versatile concepts of optical computing as an emerging technology for AI hardware accelerated computing with low power consumption.
- Incorporate novel photonic solutions in the optical interfaces of fiber-optic communication infrastructure (data centers / supercomputing centers).
- Implement standalone ML photonic substrates as fundamental ML building blocks (such as photonic perceptrons).
- Design, test and validate photonic integrated circuits with a minimum energy footprint that offer parallel advanced capabilities through AI implementations.
- Design and implement photonic accelerators with AI capabilities for ultra-fast data pre-processing and volume reduction of big data obtained in real-time (eg. from optical sensing and communications networks) or offline.

At an ***architectural level,*** research should be conducted to develop new system architectures tailored for the specifities of the neural network systems. Based on those specifities new in-memory computing basic building blocks, routing and communication blocks and parallelization strategies should be designed. The architecture should look for a ***compromise between versatility and efficiency.*** A versatile hardware is desired to be able to implement different signal codings, different learning algorithms, different neuron models and different neural architectures. However, adding versatility results in a more complex architecture less area and power efficient. At the same time, the architecture should divide the system functionalities among the different technological platforms (CMOS, memristive and photonics) to optimize the system performance.

CHALLENGE 8

SMART CYBERSECURITY

Coordinators

D. Arroyo Guardeño (ITEFI, CSIC)

P. Brox Jiménez (IMSE-CNM, CSIC – US)

Participant researchers and centers

J. Godoy (CAR, CSIC – UPM)

J. Villagra (CAR, CSIC – UPM)

H. Mueller (IAE, CSIC)

V. Gallego (ICMAT, CSIC-UAM-UC3M-UCM)

A. Kosgodagan (ICMAT, CSIC-UAM-UC3M-UCM)

R. Naveiro (ICMAT, CSIC-UAM-UC3M-UCM)

D. Rios Insua (ICMAT, CSIC-UAM-UC3M-UCM)

D. Rodríguez González (IFCA, CSIC – UC)

S. Hidalgo Villena (IMB-CNM)

S. Degli Esposti (IPP, CSIC)

P. Noheda Marín (IQOG, CSIC)

1. EXECUTIVE SUMMARY

The relationship between AI and cybersecurity is of conflicting nature. While threat detection and containment can be perfected by the use of new AI tools and methodologies, the current data deluge and the increasing complexity of *information and communication technologies* (ICT) make almost impossible to properly protect them without the guidance of automatic decision making solutions. However, an excess of confidence in such solutions can compromise security and enclose safety risks for information systems. Moreover, neglecting the treatment of these risks could undermine the different modalities of governance that configure some pillars of our democracy. Fundamental citizens' rights such as privacy or accountability in public procurement are major components of the conundrum related to the alignment of technological possibilities with ethical, legal and normative regulations. This chapter summarizes the CSIC approach to tackle the above challenges in the crossed domain of AI and cybersecurity, ***underlining the need for an integral strategy for the deployment of secure and safe AI systems.*** This approach encompasses the whole stack associated with the design and implementation of AI solutions, ranging from the hardware to the application layer, and considering the

theoretical underpinnings to adequately bridge AI functionality and cybersecurity requirements.

2. INTRODUCTION AND GENERAL DESCRIPTION

ICT systems need to integrate approaches to minimise security risks. In order to forecast, monitor, and update the *security of ICT systems, techniques based on AI and other automated tools contribute to threat detection* in massive data processing with minimal human intervention in order to protect national critical infrastructures. According to the EU NIS Directive (2016/1148), the 2019 Spanish Cybersecurity Strategy and global standards (e.g. ISO27001, ISO27005), cyberspace needs to be protected from malicious and illicit activities of all kinds. The resilience and operational continuity of critical infrastructures (water, energy, transport, financial and health sectors) need also to be ensured. Computer Security Incident Response Team must be equipped with the appropriate advanced cybersecurity tools to adequately respond to attacks or system failures.

All private and public actors must contribute to achieve the following objectives. First, the protection of computer systems, networks between computer systems, networks of networks (social networks included) against eventual attacks to their hardware, software and electronic data, as well as the disruptions or failures of the services they provide (computer technology). For instance, *Global Navigation Satellite Systems* (GNSS) are strategic to provide geo-spatial positioning. Second, in order to ensure the trustworthy development of interconnected digital technologies, we need two things. A significant improvement in the quality and safety of AI-assisted services facilitating the interaction between humans and between humans and bots. The safe, proportionate and ethical processing of massive amounts of data through a robust infrastructure to enable extensive and intensive digital communication activity both terrestrial and through satellites.

One of the added problems that advanced cybersecurity has to handle is that there is no overarching government ruling in the cyber-physical domain. It has generated and generates numerous conflicts over boundaries, hierarchies of rights and priorities over how the technology should be designed and used. Additionally, the irruption of available efficient *machine learning* (ML) algorithms (in special, *deep learning* [DL] algorithms) and their ubiquitous implementations in public and private services (including, health, education, justice, governance, public and administration, citizen security, wellness, mobility,

business management and optimization, marketing and demography), together with autonomous computer security bots that automatically repair security vulnerabilities without human intervention (cybersecurity reasoning systems), increase the need for advanced cybersecurity research with capability of facing new kinds of global safety vulnerabilities, risks and threats.

Governments cannot only rely on private companies to develop next-generation cybersecurity solutions, but need to have their own experts to ensure common security criteria are respected in product development, cryptographic protocols are robust, backdoors are not built into critical systems. Starting from a vision of *cybersecurity as a public good,* public-private partnerships on cybersecurity need to maximise and advance theoretical and operational knowledge on risks and vulnerabilities of those digitally automated systems on which society increasingly relies to function.

As a huge variety of applications is being automated through ML algorithms, it is essential that these techniques are robust and reliable as many decisions are based on their outputs. State-of-the-art ML algorithms perform extraordinarily well on standard data but, recently, have been shown to be vulnerable to adversarial examples, data instances targeted at fooling them (Goodfellow, Shlens and Szegedy, 2015). Algorithms designer should take into account the possible presence of adversaries to be robust against such data manipulations. The work in (Comiter, 2019) provides a review from the policy perspective showing how many AI societal systems, including content filters, military systems, law enforcement systems and autonomous vehicles (AV), to name but a few, are susceptible and vulnerable to attacks. The proper alignment between the benefits of the new cyberphysical reality and the ***protection against cyberthreats using AI techniques is a major challenge.*** To face this properly, cybersecurity solutions have to be evaluated.

Traditionally, information security has been structured around the protection of the *confidentiality, integrity and availability* (CIA) of information. The complexity of living in an hyper-connected environment demands going beyond the CIA triad (Vacca, 2013) to include security objectives such as *authentication, authorization and auditability* (3Au scheme) procedures (Krutz and Vines, 2002). This implies the deployment of security policies focused on the delimitation of a security perimeter supported by access control policies (e.g. traffic routers, firewalls) that enable network traffic analysis and control. The correct application of policies and procedures to protect the CIA and 3Au properties guarantee the proper functioning of information systems and communication

channels. Thus, information assets are protected through the preservation of the CIA+3Au, and the adequate creation of a secure perimeter. Security can only be achieved if there is a coherent and consistent articulation of such paradigms in relation to the solutions for achieving a good security perimeter. Moreover, the integration of AI as part of an integral cybersecurity strategy should acknowledge the contributions from the conventional information security framework, but it has to be extended to treat other aspects arising form (cyber)safety, physical harms, cyberattacks targeted at hindering assets as reputation and trust in institutions, organizations and governments, as it occurs in the case of hybrid warfare. Data-driven methodologies and tools should be devised to further extend the CIA+3Au and construct a smart perimeter to articulate cybersecurity and cybersafety more coherent and consistent manner.

The variable and dynamic nature of various modes of information exchange makes it necessary to reinterpret the *security perimeter,* in which the concept of digital identity plays a central role. A digital identity is nothing more than an application that assigns a user a place in the cyberspace. That application is an operation performed on the basis of something users know (password) and something that identify them (a biometric feature or an ID). These digital identity management mechanisms require the involvement of some sort of central authority, which is responsible for checking the authentication information provided by the user and authorizes access to the system/service if it is valid. In practice this means that this central authority has stored information that enables the verification of digital identities. If the central authority loses this information (e.g., credential theft in cloud services such as Dropbox) there is a serious security problem. In addition, the central authority has the ability to record all the activity of a digital identity and, therefore, of the corresponding user. The latter infringes on users' privacy, something that may have legal and/or regulatory consequences, in addition to a possible deterioration in users' confidence in ICTs. This impact is of major relevance if we take into consideration AI applications based on the automatic treatment of personal data to sustain or deliver automatic decision making. In other words, the outsourcing of data storage and computing encloses a challenge in terms of governance, which should be properly considered and managed to adequately include AI as part of any integral cybersecurity strategy. As an important component of the CSIC smart cybersecurity plan, blockchain and distributed ledger technologies are discussed as a means to monitor AI activity, foster accountability by means of advanced digital evidence recording and treatment, and promote transparency in the context of e-governance.

3. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Information security in general, and cybersecurity in particular, are built upon the intertwining of multiple disciplines and knowledge domains. Ranging from communication systems to networks, any software application cannot be considered neither safe nor secure unless the underlying hardware is properly analyzed and validated, which also calls for an exhaustive pondering of each single functional and non-functional requirement. In the very case of AI and cybersecurity, the dual nature of their connection demands the articulation of holistic approaches. Advances in AI disciplines as *natural language processing* (NLP), recommender systems, people analytics, DL, and chatbots have increased the effectiveness of different types of autonomous agents in scenarios as digital on-boarding in banking systems, AI-powered messaging in e-commerce and targeted advertising, robotic autonomous systems, autonomous vehicles, unmanned aerial vehicles, cyber weaponry, and automatic writing. In all these contexts, information is not only captured, transformed and exchanged. Certainly, automation is imbricated in the very generation of new pieces of information that could be further fed into (automatic) decision making/taking processes. Security and safety in all these areas demands knowledge construction and sharing between professionals and researchers with very different background and interests. Therefore, there is an urge to foster interdisciplinary theoretical innovation in cybersecurity, which has a special relevance in the case of the crossed domain of AI and cybersecurity.

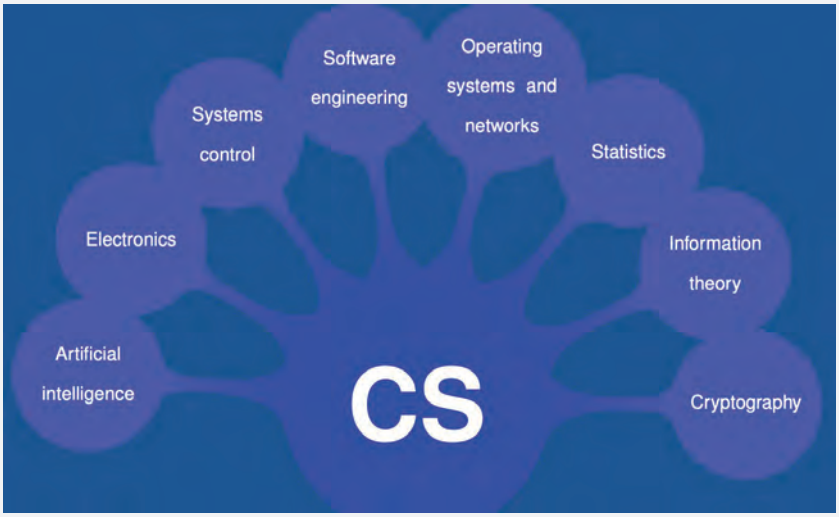
3.1. Fundamentals

Mathematical, logical and physical foundations

Many mathematical sciences aspects are relevant at the AI-cybersecurity interface. To name but a few, statistical and ML methods are essential for properly uncovering patterns and trends in attacks; risk analysis provides methodologies for the proper management of threats; game-theoretic methods facilitate modelling the presence of intelligent adversaries. Advances in these areas facilitate progress in cybersecurity analytics; in turn, complex problems in the cybersecurity domain motivate advances in mathematical sciences.

Smart cybersecurity is going to be articulated in cyberphysical enviroments where any mathematical framework should be properly enhanced and adapted by physical underpinnings. The application of formal methods, ϖ calculus and communication sequential processes, is of major relevance to bridge

FIGURE 8.1–Knowledge domains involved in the deployment of an holistic cybersecurity policy.



mathematical tools and the requirements and limitations of different operational contexts. The latter has been successfully applied in cybernetics and control systems to verify safety and liveness properties. Nonetheless, there should be a roadmap to better bridge AI, cybersecurity and control theory.

Threat modeling

AI can be applied to the analysis of malicious activity in information systems and networks. This could contribute to extend and enhance current methods to characterize computer security threats, as STRIDE. Indeed, AI can aid in grasping more adaptively the tactics and strategies of potential attackers. It is worth noting the difficulties associated with modeling insider threats and byzantine faults, the threats associated with users with privileged information about an organization, but also the problems associated to spurious activity and unintentional misuse of information systems. Along with *data mining* (DM) and ML, game theory and formal methods based on ϖ calculus should be integrated into general frameworks for the characterization of attack vectors and their impact in information systems (Diaz et al., 2019).

With the purpose of *enforcing cybersecurity,* there exist several security threats whose subsequent analysis can be very helpful to protect the systems at the

end. *Big data* (BD) tools should be applied in the design, implementation and validation of adaptive systems for the protection of cyberspace and, therefore, of our physical world (Curry et al., 2013). To this end, the first step is to carry out an in-depth review of the uses of DM, ML and AI, both in the characterization of abuses in the use of ICT and in the establishment of safeguards and other proactive mechanisms for a safer and more satisfactory experience in the ICT field (Stamp, 2017). Successful cases of DM-ML-AI can be found in the context of cryptography (Rivest, 1991; Bost et al., 2015), the identification of malicious software (Ucci, Aniello and Baldoni, 2017) and intruders (Huang, He and Dai, 2015) or the attribution of responsibility in the diffusion of software (Caliskan-Islam et al., 2015). Based on the recognition of the work in each of the cybersecurity domains where some element of the DM-ML-AI triplet has been applied, we also need to realize the need for comprehensive approaches to analyze heterogeneous data sources to identify security threats quickly and effectively (de la Torre, Lago and Arroyo, 2019). Likewise, we have to take into consideration that such an integral approach must be continuously adjusted, to react in a convenient way to short-term or structural modifications of the operation context.

In this context, DL is presented as a key element in providing such a comprehensive approach. In recent years, deep neural network models have broken records in many ML application areas, from NLP to machine vision, and have recently begun to be applied in the context of cybersecurity. The main hypothesis of the medium-term objectives consists of considering DL techniques as the fundamental basis of the integral scheme for early detection of threats in information security management systems. This approach takes as a reference, in the first place, the advantages of deep networks as mechanisms for the automatic codification of features and their capacity to take advantage of existing structural and spatial dependencies in data to build reliable models. Secondly, it takes into account the positive results of DL in the framework of cybersecurity, in applications such as intrusion detection(Kim and Aminanto, 2017), privacy protection assessment (Rimmer et al., 2017), malware identification (Bisio et al., 2017; Yu et al., 2017; Lison and Mavroeidis, 2017; Su et al., 2018), continuous user authentication (Shoshitaishvili et al., 2017), etc. Finally, the comprehensive scheme for dealing with security events and incidents must be supported by an adequate articulation of its life cycle. In this respect, a methodology for the correction of the hyperparameter space of DL models will be implemented based on recent results in the field of *adversarial ML* (AML) and algorithmic fairness (Lepri, 2018).

Adversarial machine learning

A basic hypothesis of ML theory is that their systems rely on *independent and identically distributed* data for both the training and testing phases. However, security aspects, which conform the field of AML, question such hypothesis given the presence of adaptive adversaries ready to modify the data and obtain a benefit, potentially degrading the performance of ML algorithms with important consequences. Practically all ML methodologies have been touched upon from an adversarial perspective including, to quote just a few, naive Bayes; logistic regression; support vector machines; latent Dirichlet allocations; or deep neural networks.

AML is a difficult area rapidly evolving through an arms race in which the community alternates a cycle of proposing attacks and implementing defences. Thus, it is crucial to develop sound techniques. Note though that, stemming from the pioneering (Dalvi et al., 2004), most of this research has been framed within a standard game-theoretic approach pervaded by Nash equilibrium and refinements. However, these entail common knowledge assumptions which are hard to maintain in the security realm. We could argue that *common knowledge* is too commonly assumed. A new paradigm seems to be required.

Disintermediation and decentralization: beyond blockchain

A technology with great potential in AI governance is the blockchain (Werbach, 2018), the base technology of bitcoin. In the case of bitcoin, blockchain is oriented towards creating a record of financial transactions based on a distributed consensus protocol and, therefore, without a central authority (the database administrator or backing services). Today, these distributed networks solve the problem of the custody of information assets, without relying on a trusted third party. *Distributed ledger technologies* (DLT) offer a whole range of solutions for disintermediation in the processes of information management and value exchange (ISO). These solutions lead to new ways of generating and negotiating trust and tracing and linking operations. These facilitate new forms of managing digital identities that are consistent with the requirements of the *General Data Protection Regulation* (GDPR) (Voigt and Von dem Bussche, 2017) and the *Payment Services Directive* PSD-2 (Cortet, Rijks and Nijland, 2016), and with the specifications of the regulation on electronic identification and trust services in electronic transactions in the internal market (electronic identity recognition scheme (European Union)).

The design of new consensus protocols in asynchronous distributed systems configures an essential step in the definition of new decentralized trust management schemes. DLT and, particularly, blockchain are built upon consensus algorithms that enable state validation and replication without the participation of any sort of central authority. This opens up the possibility of establishing data governance schemes with different levels of traceability, linkability, transparency, and accountability. Among the most adopted consensus algorithms, the *proof-of-work* (PoW) is the most popular public permissionless blockchain of the kind applied in the case of bitcoin and ethereum. In addition, there exist plenty of alternatives to the PoW, such as the proof of stake or the practical byzantine fault tolerant. From a theoretical point of view, it must be established the relation between the safety and security of blockchain ecosystems and the properties of the underlying consensus protocols. Moreover, consensus protocols along with the peer-to-peer communication network and the cryptographic layer of a blockchain determine the on-chain governance of this technology. However, it is important to notice that the blockchain technology does not impose restrictions by itself on the behaviour of external agents. As a result, it is necessary to take into consideration different off-chain governance schemes to ensure the right balance between conflicting interests, business dynamics, and legal and normative requirements. Thus, the study of consensus algorithms and protocols should be conducted not only from the perspective of the theory of asynchronous distributed systems, but should also encompass contributions from other fields, such as game theory, corporate governance and complex networks. This type of analysis is especially required in the case of public permissionless blockchains, since the efficient and effective collaboration among stakeholders cannot be achieved unless efficient and effective rewarding policies are clearly and robustly defined.

Blockchain can play a key role in IT governance. Taking as a starting point the *International Standard for Corporate Governance of Information Technology* (ISO/EIC 38500), while also acknowledging the fundamental role ICTs plays in e-government, there is a need to envision new ways of incorporating IT design decisions into the complexities of corporate, social and political governance. The design, implementation and validation of different technological means to foster citizens' participation in decision making, public procurement monitoring, and other expressions of digital transparency and accountability, is necessary to promote trust in e-governance schemes and ensure active verification of the trustworthiness of trustees.

The following areas of research can be identified within the blockchain domain:

- Blockchain protocols to bear data curation processes and protect the data lifecycle.
- Advanced blockchain governance schemes to monitor AI-based decision making.
- Development of schemes based on blockchain for the correct protection of information assets and the privacy of sensitive user data (Finck, 2018).
- Use of blockchain as a distributed storage system for event capture in highly dynamic environments, such as IoT networks (Iglesias García, Diaz, and Arroyo, 2019).

In addition, blockchain scalability and usability should be tackled in each of the above research objectives, otherwise the technology and, in specific, its integration in the AI ecosystem will be critically hindered.

Privacy-utility models for data sharing and minimisation

Privacy is a very elusive, context-dependent concept, difficult to be expressed in formal terms. Privacy can be interpreted as confidentiality, as control, and as practice. One major challenge in the data deluge era is the development of a solid theoretical framework to define the possibilities and limitations of data exploitation models. In conjunction with traditional sanitization and pseudo-anonymization procedures, the differential privacy approach seems promising in delivering privacy respectful ML and DM. Along the development of cryptographic tools, this kind of proposal can be further extended by means of, for instance, homomorphic encryption and multiparty computation. Furthermore, trust management, specifically the definition of *dis-joint trust domains* is very relevant to promote trustworthiness across different models of data outsourcing, storage and computation. A number of proposals have been made to protect people's and information privacy. Some examples are: group signatures; network traffic anonymization through Tor; CryptDB for querying encrypted data using homomorphic encryption; or the theoretical framework of differential privacy (widely used in biomedical data processing). Indeed, users' privacy can be protected using *privacy enhancing technologies* (PET). PETs include cryptographic tools for managing anonymity and processing encrypted information, as well as solutions for information obfuscation. While PET technologies enable the protection of users' identities, they can also be abused by malicious actors, as it happens with bitcoin and Tor, generally associated with ransomware attacks and

payments (e.g. WannaCry, Petya or Ryuk attacks). These technologies are also associated to illegal market activities on the dark web. As any technology, PETs can be abused. Anonymity, which be used as a camouflage by criminals, is a cornerstone of electronic voting systems (Querejeta Azurmendi et al., 2019) and of privacy-respectful e-commerce platforms (Diaz et al., 2019). Undertaking an analysis of privacy-security tradeoffs (Arroyo et al., 2015; Diaz et al., 2018), and of privacy-surveillance tradeoffs (Ball et al., 2019) is necessary to enable to new forms of interventions of security agencies into encrypted communications and services. Striking a balance between the need to fight malicious activities and the safeguard of people's data protection rights, as required by the GDPR (Voigt and Von dem Bussche, 2017), is a huge challenge that requires interdisciplinary research, technological innovation, and policy development.

3.2. Applications

Data-driven threat models and containment in information systems

Over the last two decades, DM and ML techniques have increasingly been used to gather and interpret digital evidence, and enhance auditing and forensics capabilities on both networks and devices. *The protection of the security perimeter has evolved:* from the simple use of routers, firewalls and other network devices, we have witnessed the proliferation of behavioural, data-driven operational models focused on intrusion and malicious activity detection. Some examples are: intrusion detection systems, web application firewalls, and security information and event management systems. These solutions rely on BD analytics, thus requiring further improvements of DM, NLP and other ML techniques. Moreover, the elusive nature of evolving cyberthreats, and the heterogeneity of agents operating in cyberphysical systems, demand the integration of information sources placed outside corporate or government control. In the case of disinformation campaigns and other information operations, the study and characterization of social media and open data is key to leverage open source intelligent. There is a widespread demand for effective procedures to monitor the quality of information travelling online. As in previous cases, information gathering is an essential element that require to improve sensorization and event recording/treatment along the entire data lifecycle. The proper integration of hardware and software tools and methodologies is needed to ensure system and AI trustworthiness in line with the security-by-design principle.

AI can foster the creation of robust authentication, authorization, audit and accountability solutions. In this vein, a line of research that needs to be

expanded is the use of DL architectures for the extraction, processing and modelling of information from system and security logs (Chuvakin, Schmidt and Phillips, 2012), but also from open sources on security alerts and social networks (Bair et al., 2017; Sabottke et al., 2015; Khandpur et al., 2017). Recurrent long short term memory networks (Hochreiter and Jürgen Schmidhuber, 1997) are currently used, which are very useful for modelling sequences, although other techniques such as variational autoencoders (Kingma and Welling, 2013) or generative adversarial networks (Goodfellow et al., 2014) should also be taken into consideration.

Misinformation, fake news and cyber-attribution

Users are bombarded by the flow of digital information coming not only from newspapers, TV and radio, but also from digital media and user generated content. The new media environment impacts the way people form their opinions as well as the mechanisms leading to democratic consensus. Increasingly divergence of opinions and polarization reinforced by echo chambers is fragmenting public opinions and dividing citizens even when discussing common-sense daily issues.

Hybrid threats include methods of warfare, such as propaganda, deception, sabotage and other non-military tactics which have long been used to destabilise adversaries. Foreign governments may use social media platforms to influence public views and polarise political opinions by targeting opposite political candidates during election campaigns. Spain has been the target of various disinformation campaigns and information operations sponsored by foreign states. Historical cleavages and local conflicts extend the vulnerability surface of society and offer opportunities to certain groups or to external attackers to easily exploit people's vulnerability to misinformation at their advantage. Nowadays, it is possible to monitor information operations thanks to Twitter's election transparency initiative and the publication of thousands of tweets associated with propaganda or disinformation.

To better understand the relationship between fake news, post-truth, misinformation and disinformation it would be useful to clarify the terminology. *Disinformation* refers to motivated faking of news or other information as part of cyberwarfare. To have disinformation we need an attacker or malicious entity suing inaccurate information to intentionally deceive the audience. *Misinformation,* in contrast, only refers to the presence of inaccurate information. Disinformation campaigns are also called 'information operations' and

are mostly conceived to affect the stability of political institutions in another country. A promising line of research at the crossroad between computer science and social science explores the role social media platforms play in building political polarisation starting from affective polarisation and ideology. Understanding the drivers of political divisions and opinion polarisation is a fundamental research priority to safeguard our democracies from hybrid threats and information operations (Westwood et al., 2018).

Besides understanding how people consume and interpret information, another important line of research in computer science focuses on misinformation spreading in complex networks through human accounts and bots, and also on assessing information accuracy and source trustworthiness. Evaluating information quality and reliability on digital platforms demands the development of automated AI tools for the early detection and reaction to disinformation campaigns. The coexistence of official and unofficial sources creating and spreading contents online complicates the task of assessing the veracity of a piece of information. As today, the detection of misinfornation is based on singling out malicious spreading strategies but not so much on the content itself. Another challenge to be faced in the fake news detection area, is the recent interest in text-generating AI systems, raised partially due to the creation of the GPT-2 system by OpenAI. This fact manifests the urgent need to introduce real time and automatic fake news detection methods.

News curation and cyber-attribution are major challenges in the digital media environment. Instances of misinformation (inaccurate or contested information) coexist with hybrid threats and cyberwarfare. Individual users may unintentionally amplify the effects of specific information operations by keeping sharing malicious or inaccurate contents. While some users are more vulnerable to phishing attacks, other users are more vulnerable to disinformation campaigns promoting forms of reasoning rooted in conspiracy theories. Filter bubbles and radicalization dynamics can also reinforce the effects of misleading communications thought the by polarisation of public views.

Mis/disinformation bring in specific challenges related to: (a) the reliability of information sources and the identification and forensic attribution of the attacker; (b) the classification of misleading information and its tracking through complex-network analysis; and, (c) an in-depth understanding of peoples' vulnerability to mis/disinformation and the deployment of accurate and effective counter-deception communications.

Addressing these challenges require analysis of psychological mechanisms and public opinion reactions, combined with the deployment of detection tools capable of identifying malicious sources and misinformation spreading on a variety of channels, from chatbots to encrypted communication systems. Social media and instant messaging applications are constantly under scrutiny as likely misinformation channels. However, the identification of misleading information on encrypted channels entails controversial privacy-security tradeoffs. Current solutions try to incorporate a human-in-the-loop logic into the automatic detection of misleading information through NLP and complex network analysis. Additional measures to improve security without compromising privacy envision user reporting mechanisms and cyberawareness.

Cyber threats and cyber insurance

As discussed above, all kinds of organisations are critically impacted by cyber threats. *Risk analysis* is a fundamental methodology to help manage such issues. With it, organizations can assess the risks affecting their assets and what security controls should they implement. Numerous frameworks support cybersecurity risk management. Similarly, several compliance and control assessment frameworks, provide guidance on the implementation of cybersecurity best practices. They have many virtues; however, much remains to be done regarding risk analysis from a methodological point of view: a detailed study of the main approaches to cybersecurity risk management reveals that they often rely on risk matrices, with well documented shortcomings, potentially inducing suboptimal cybersecurity resource allocations.

In this context, a complementary way for dealing with cyber risks through risk transfer is emerging: cyber insurance products have been introduced in recent years. However, cyber insurance has yet to take off. Thus, all advances proposed, addressed towards improving smart cybersecurity risk management should help enhancing better grounds for cyber insurance adoption. Moroever, dynamic AI approaches to cyber risk management may lead to enhanced dynamic cyber insurance products.

Security, transparency and accountability in the Fintech sector

The transition into a cashless society and the implementation of open bank infrastructures entail a series of challenges. On this concern, it is of major relevance the required effort to achieve adequate tradeoffs between the protection of citizens' rights (e.g., in the context of GDPR and PSD-2) and compliance with fraud detection and prevention requirements. The convenient implementation

of know your customer and *digital onboarding* solutions by banks and other financial institutions is a must to meet the requirements of EU laws on anti money laundering and counter terrorism financing. The proliferation of cryptocurrencie and, specifically, of those enabling financial transactions in a (pseudo) anonymous way. The analysis of public permissionless blockchains is critical to identify and prosecute criminal activities, as it is the case of drug trafficking or cyberattacks as ransomware campaigns or the implementation of command and control systems targeted to access privilege information from institutions and organization. Network analysis and sociophysics provide a set of methodologies that pave the way to monitor and trace cryptoassets, and eventually enable the identification of illegitimate financial flows. NLP is another useful framework to correlate dynamics in underground fora in the deep web and capital flow by means of cryptocurrencies as bitcoin. Moreover, the aggregated analysis of open data and other cryptosignals is demanded to scaffold adequate governance schemes and accountability solutions in cryptoeconomy.

Secure computing on dependable systems

Lemmas such as "design for test" or "design for manufacturability" have recently been replaced by the novel "design for trust" imperative. To ensure the dependability of systems, both software and hardware and increase Spanish and European technological independence, a set of new methods supported by open hardware and software modules have been proposed to guarantee full protection of personal and confidential information. Specific hardware modules can increase the security of digital processing systems, while trusted and open *systems-on-chip* (SoC) platforms can increase European technological self-determination. Dependable systems built from the root of trust to the interface level are suitable for a wide range of applications; from tiny and ultra-low power internet-of-things devices, to the most advanced high performance computing systems. The open RISC-V initiative is an example of trustworthy hardware development meant to improve security, while reducing risks, and allowing the exploration of new micro-architectural extensions for specific application domains also enabling sustainable energy-efficient tools and methodologies.

Embedded systems are major components of the physical layer of any information system, and thus of any AI system. The hybrid (hardware/software) nature of embedded processors can be exploited to increase the overall level security of AI systems. In certain scenarios, there is an urge to update the management of cryptograhic keys and remote attestation to enhance conventional secure and trusted hardware architectures based on trusted platform

modules. For instance, these hardware modules are not affordable to be integrated in devices with limited resources. One option is to build hardware root-of-trust, also known as hardware anchor, with the inclusion of hardware dedicated modules. The *intrinsic digital identity* (ID) derived by hardware is used as an anti-counterfeit mechanism to detect the impersonation of devices (Martínez-Rodríguez et al., 2018). The idea is that the unique ID is used to regenerate cryptographic keys as many times as necessary without having to store them. Therefore, the underlying hardware circuitry is used as the basics to build high level cryptographic protocols running on software programs on the processors.

Another way to protect electronic devices is to monitor the system during its normal operation to detect eventual anomalies. The design of hardware modules that collect micro architectural information of the system is a mechanism to diagnose it (Tang, Sethumadhavan and Stolfo, 2014). Modern processors and SoCs include the so-called *performance monitoring units* (PMUs). This idea can be extended with the design of hardware dedicated modules to measure execution times and events using hardware/software co-design methodologies. Therefore, PMUs provide real-time feedback to diagnose bugs, identify anomalies, or bottlenecks during program execution. This information can be used as training set for AI techniques. Additionally, other hardware modules to test the hardware platform can be included to reinforce the decision-making accuracy of AI techniques. For instance, the inclusion of aging sensors that measure the performance degradation of circuitry over time. Or the integration of on-line testing techniques to guarantee the reliability of the output response during the generation of digital ID.

Furthermore, AI-based penetration testing can be carried out to achieve resistance against *reverse engineering* (RE) techniques and attacks. Actually, RE techniques of *integrated circuits* (ICs) can be applied with different security purposes. For instance, the certification of cryptographic algorithms implemented in hardware, the identification of hardware trojans inserted during the fabrication process, and the hacking of ICs to access sensitive information for forensic purposes (Quijada et al., 2018). Whereas the non-invasive *side channel attacks* (SCAs) exploit vulnerabilities on software or hardware implementations of cryptographic algorithms to recover the secret key (Hettwer, Gehrer and Güneysu, 2019), the use of methodologies based on AI techniques reduce the time required to complete RE and SCA procedures. In the case of SCA attacks, published works have demonstrated that DL based SCAs are very

efficient when targeting cryptographic implementations even protected with the common side-channel countermeasures (Maghrebi, 2019). In short, AI can be used to conduct RE on ICs with a wide range of security purposes as:

- Pattern and object identification. Logical block and security mechanism identification on IC surface.
- IC segmentation. IC decomposition into basic logical blocks.
- Intellectual property modules detection. Proprietary logical blocks identification using graphical convolutional networks.
- Image super-resolution. Use of low-quality images to reconstruct high-quality ones.
- Image restoration. Damaged image reconstruction using *generative adversarial networks* GANs.
- Power consumption patterns. Functional patterns identification using recurrent neural networks.
- Voltage contrast IC signals identification. Application of NLP procedures to pattern identification and functionality inference from IC signals and video analysis.
- IC Synthesis tools. Total or partial IC generation using GANs.
- Style Transfer. IC design and technology identification in order to emulate its functionality using computer-aided libraries for IC design.
- Adversarial attacks. Use of evolved DL models in IC reconstruction.

Dependable autonomous vehicles

Although the dream of the SAE for a level 5 *autonomous vehicle* (AV) could be envisioned as completely independent entities without any connection with the outside world, every current player envisions the car of the future with multiple interactions with its environment (*vehicle to environment* [V2E]), including road agents, infrastructure elements and service providers. This V2E communication has to guarantee distributed end-to-end security to ensure robustness against all types of attack vectors. The standards and security practices used today are not adequate for AVs. Functional safety standards like ISO 26262 and its evolution into SOTIF, information sharing like Auto-ISAC, software coding guidelines like MISRA, or overall safety scores like EURO NCAP do not solve the security issues that AVs will have to face in the short term. Upcoming AVs will have to consider securing each AI mechanism system itself, and communications between edge computing devices or vehicles with encryption and authentication mechanisms against attacks.

AI mechanisms for AVs

Cognition-inspired mechanisms and AI techniques are used to learn about the environment, and must detect unpredictable and harmful behavior, including hacking. In this context, the actions of an AI strategy may be limited by how it learns from its environment, how the learning is reinforced and how the exploitation dilemma is addressed. AVs could be exposed to malicious actors trying to manipulate the artificial perception and decision-making systems based on adversarial learning mechanisms that influence the training data for abnormal traffic detection.

Secure V2X communication

The majority of V2X messages are safety-related broadcast, with no restriction on which vehicles within range are allowed to read them. A security-related requirement for safety-related messages, is that senders should be trustworthy and accountable. Therefore, the removal/revocation of detected misbehaving participants from the V2X system is key. As a result, the requirement needs to be supported by appropriate technical measures and certification procedures for the underlying software and hardware.

The two previous aspects need to be seamlessly considered in critical AV functions. A good example of this is the enhanced and collaborative perception, which draw from AI to interpret a larger environment in a more dependable way, and needs consistency checks using ad-hoc V2X C-ITS messages, such as the Collective Perception Message. In any case, the involved components should be secure (i) by design (to ensure CIA); (ii) by default (with the confirmed capability to support these security properties at installation); and (iii) throughout their lifecycle.

Data protection by design and by default

Appropriate business, legal, and technical infrastructures are needed to transform into practice GDPR principles such as *privacy by default* and *by design* (Bender et al., 2014). They require to adopt the data minimization principle from the design-phase of a system and as a default property. According to art. 25 of the EU GDPR, people's personal information should be protected by design and by default. This implies two things. First, proper data minimization procedures should be designed in any information management solution including AI applications. Second, that data minimization principles should represent default features. Because of the key role that data plays in ML, adequate schemes to evaluate the quality and reliability of data are needed. There is a

need to improve the theory of data minimisation to enable utility models for AI consistent with citizens' expectations and data protection rights. It is also necessary to establish adequate data classification procedures to comply with current data protection policies. In addition, specific organizational and technical measures must be established to protect data according to their level of sensitivity. For instance, highly sensitive data should be protected using data loss prevention systems by means of advanced data-driven solutions.

Biomedical research and privacy risks

The availability of healthcare data in digital formats has increased over the years with the introduction in hospitals and clinical practice of digital devices and equipment (e.g. computed tomography, MRI, digital microscopes). The storage and availability of health data bring huge opportunities for biomedical research, but also pose considerable privacy risks. According to GDPR Art. 9, additional organisational and technical measures need to be established for the processing of "personal data revealing racial or ethnic origin, [...] genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation". These special categories of data are highly sensitive but also very valuable in economic and instrumental terms as they can improve dramatically medical research and facilitating the personalisation of treatments. However, health, genetic and biomedical data posits serious anonymisation challenges, which are accentuated by data variety, that is, from the fact that these data come in a variety of formats, which further complicate data management and the de-identification process. For instance, a 3D reconstruction including the face of a person can be easily obtained from MRI acquisition of a person's head. While it is possible to mitigate the privacy risk of re-identification by using brain extraction software or other de-facing techniques, there are instances for which no appropriate de-identification tools are available. Furthermore, brain structures can be matched using expectation–maximization algorithms, so once an adversary is in possession of one MRI view of the brain of a person it can identify other images of the same brain. Genomics and DNA data also present specific privacy protection challenges. The popularisation of consumer genomics services has increased reidentification risks for individuals as well as for ethnic groups. It has been estimated that a large proportion of US population can be identified starting just from a DNA sample even if these people never used a DNA service.

4. KEY CHALLENGING POINTS

We have identified the following major areas at the crossroads of cybersecurity and AI to which CSIC researchers will contribute.

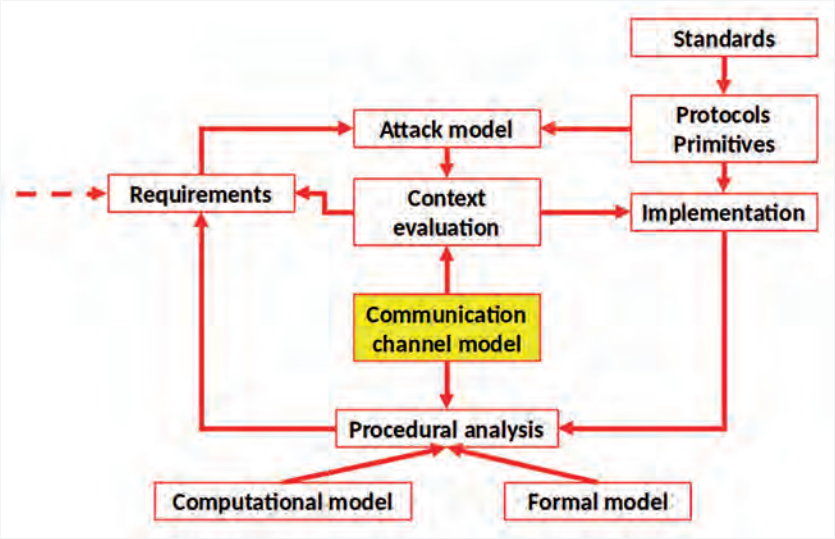
4.1. Fighting Misinformation About Science

The COVID-19 pandemic foregrounded the interconnection between the physical and the digital world and also demonstrated the extent to which the spread of biological and informational viruses bring new challenges to societies. The accelerated pace of scientific production and dissemination during the pandemic and the proliferation of fake news, rumors and hoaxes on those findings put institutions and official sources as well as fact-checking platforms under pressure. People in many countries were victims of misinformation about the contagion or the treatment of the disease. ***The COVID-19 pandemic made evident the need for better crisis management and preparedness strategies incorporating automatic end-to-end content verification tools.*** Acknowledging the capital role of human experts in information curation and verification, we are designing a solution incorporating a variety of data sources. We envision the integration of crowdsourcing approaches, based on knowledge extraction from open social networks and instant messaging apps, and expert knowledge obtained from scientific publications and technical standards. The definition of falsehood and truth is an epistemic and technical challenge, as well as establishing rules for drawing a line between freedom of expression and misinformation. Methodologies for assessing information quality and source reputation and trustworthiness also need to be established. Open questions remain on how to mitigate the damage that fake news and rumors can cause in society. There are several current examples of collective decisions taken under the pressure of misinformation. Collecting data from the different spreading media (online social networks, newspapers and traditional media) is fundamental to characterize and identify the spreading patterns. Modeling and AI can help to design new strategies to reduce the impact and to explore ways to expose the population to un-skewed information.

4.2. Imposing Security-by-Default Along the Computing System by Leveraging AI

Computing infrastructures are constant targets of cyber attacks aimed at gaining access to valuable assets such as data or computing power. To respond to advanced persistent threats, zero-day attacks, hybrid or other emerging threats, there is an ***urgent need to offer integral approaches to secure information***

FIGURE 8.2–Integral methodology for the secure software and hardware development life cycle.



assets and critical infrastructures. An important part of this holistic approach to threat modelling and containment is represented by methodologies to assess the reliability and trustworthiness of hardware and software components along their life cycle. Traditionally the *plan-do-check-act* (PDCA) methodology has been applied to ensure ICT systems quality. However, the complexity of hardware and software production, the dependence on third parties, and the variability of the application contexts make necessary to enlarge and perfect the PDCA methodology (Diaz et al., 2019) (see Figure 8.2). With such a goal, new AI tools are created to identify threats and evaluate security risks along the design and deployment of hardware and software products.

The expertise of CSIC researchers working on smart cybersecurity will be targeted at:

- ***The application of AI techniques*** to detect and avoid side-channel and fault injection attacks, through the adequate application of reverse engineering approaches;
- ***The deployment of computational and formal methods*** to characterise communication channels and attack models in information and communication systems;

- ***The development of procedures for comprehensive risk analysis,*** with a special consideration for automatic decision making and autonomous systems.

4.3. Creating a Formal Model for Adversarial Machine Learning

AML is an area of major importance in security and cybersecurity to protect systems which are increasingly being based on ML algorithms (Comiter, 2019). As we mentioned, there is a need for new paradigms going beyond standard game-theoretic approaches. One possibility is through *adversarial risk analysis* (ARA), which does not entail common knowledge conditions, being therefore much more realistic. As a byproduct, we obtain more robust algorithms. Here are some research challenges in this respect.

Robustifying ML algorithms through Bayesian ideas. Bayesian methods provide enhanced robustness in AML. (Ekin et al., 2019) shows how game theory solutions based on point estimates of preferences and beliefs may lead to unstable solutions, while ARA solutions tend to be more robust better acknowledging uncertainties. Thus, a promising challenge consists of developing efficient algorithms for approximate Bayesian inference with robustness guarantees. Indeed, there are several ways in which the Bayesian approach may increase security of ML systems. Regarding opponent modelling, an agent has uncertainty over her opponent's type initially; as information is gathered, she might be less uncertain about her model through Bayesian updating. Uncertainty over attacks in supervised models can also be considered to obtain a more robust version of adversarial training. Combining this approach with ARA opponent modelling may further increase robustness. Lastly, there are alternative approaches to achieve robustness in presence of outliers and perturbed observations, as through robust divergences for variational inference (Futami, Sato and Sugiyama, 2018). The robust Bayesian literature burgeoned in the period 1980-2000 (Rios Insua and Ruggeri, 2000). In particular, there has been relevant work in Bayesian likelihood robustness, referring to likelihood imprecision, reminiscent of the impact of attacks over data received. Note that Bayesian likelihood robustness focuses around random or imprecise perturbations and contaminations in contrast to the purposeful perturbations in AML.

Modelling and computational enhancements may enhance operational aspects. First, a core element in AML is the choice of the attacker perturbation domain. This is highly dependent on the nature of the data attacked. For example, in computer vision, a common choice is an A_p ball of certain radius

centered at the original input. These perturbations, imperceptible to the human eye, may not be representative of threats actually deployed. Thus, it is important to develop threat models that go beyond A_p norm assumptions. Moreover, it is important to deal with multiple agents in its variants (one defender vs. several attackers, several defenders vs. several attackers) including cases in which agents on one of the sides cooperate.

There is also a need for new algorithmic approaches. Exploring gradient-based techniques for bi-level optimization problems arising in AML is a fruitful line of research, (Naveiro and Ríos Insua, 2019). However their focus on white box attacks. On the other hand, Bayesian methods are also hard to scale to high dimensional problems or large datasets. Recent advances in accelerating stochastic gradient Markov chain Monte Carlo samplers are crucial to leverage the benefits of a Bayesian treatment. The ARA framework essentially goes through simulating from the attacker problem to forecast attacks and then optimize for the defender to find her optimal decision. This may be computationally demanding and we could explore single stage approaches to alleviate computations.

AML methods need to be developed. For example, there is a need to extend ACRA to to discriminative models and multi-class problems. Further research is also required in relation with attacks to unsupervised learning, for example through k-means clustering, autoregressive models and NLP which are just beginning to attract attention.

4.4. Safeguarding Privacy in the Era of Big Data and AI

Privacy-preserving techniques currently available have not resolved the usability-privacy tradeoff. In other words, ***preserving data privacy through obfuscation, de-indentification and encryption create some problems to data utility.*** These alternative techniques need to be complemented with strong communication security protocols and orchestrated into a comprehensive context-sensitive methodology. Identity management represents a fruitful strategy to foster privacy protection through adequate utility and privacy data models and the implementation of suitable anonymization and pseudoanonymization techniques. Starting from basic sanitization procedures (designed to eliminate quasi-identifiers), other methods based on attribute generalization can open the door to the deployment of forms of privacy-respectful querying rooted in *differential privacy* (Dwork, 2014). This can offer privacy protection, but could also reduce research data quality. Furthermore, it cannot be applied to unstructured

data types like medical images, although there exist some important contributions from the field of federated learning. To safeguard data quality, an alternative proposal is to build secure storage and analysis platforms for biomedical data, so called data safe havens, where researchers can analyse data, but cannot extract them. The problem this time is how to ensure platform security, user access control and auditing, network and physical security.

In the field of biomedical research, enabling GDPR-compliant data sharing of health records and genomics data is a key scientific and societal priority. Because of the ***intrinsic identifiable nature of genetic and biometric data, de-identification is an actual challenge.*** A mix of pseudonymisation techniques and access control procedures are usually adopted to protect this type of data, whose sensitivity is high both for individuals and groups. Racial and ethnic information presents risks to group privacy not just to individual privacy. BD adds further risks as it makes easier to identify specific groups.

VOLUME 11 REFERENCES

Aaij, R., Albrecht, J., Belous, M. et al. (2020). Allen: A High-Level Trigger on GPUs for LHCb. *Computing and Software for Big Science*, *4*(1), 7. DOI:10.1007/s41781-020-00039-7

Abba, A., Bedeschi, F., Citterio, M. et al. (2016). The artificial retina for track reconstruction at the LHC crossing rate. *Nuclear and Particle Physics Proceedings*, *273–275*, 2488–2490. DOI:10.1016/j.nuclphysbps.2015.09.434

Abu-Mostafa, Y. S., & Psaltis, D. (1987). Optical Neural Computers. *Scientific American*, *256*(3), 88–95. DOI:10.1038/scientificamerican0387-88

ACT-R Research Group. (2013). ACT-R Publications & Models. http://act-r.psy.cmu.edu/publication/, 2002-2013.

Adams, S. S., Arel, I., Bach, J. et al. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine*, *33*(1), 25–41. DOI:10.1609/aimag.v33i1.2322

Anderson, J. R. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462. DOI:10.1207/s15327051hci1204_5

Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence*, 149 (1), 91–130. DOI:10.1016/S0004-3702(03)00054-7

Andrae, A., & Edler, T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges*, *6*(1), 117–157. DOI:10.3390/challe6010117

Andrejczuk, E., Bistaffa, F., Blum, C. et al. (2019). Synergistic team composition: A computational approach to foster diversity in teams. *Knowledge-Based Systems*, *182*, 104799. DOI:10.1016/j.knosys.2019.06.007

Andriella, A., Alenyà, G., Hernández-Farigola, J., & Torras, C. (2018). Deciding the different robot roles for patient cognitive training. *International Journal of Human Computer Studies*, *117*, 20–29. DOI:10.1016/j.ijhcs.2018.03.004

Andriella, A., Torras, C., & Alenyà, G. (2020). Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction. *Cognitive Computation*, *12*(4), 793–810. DOI:10.1007/s12559-019-09696-2

Antonucci, A., Cholvy, L., & Papini, O. (2017). Symbolic and Quantitative Approaches to Reasoning with Uncertainty. *Springer International Publishing*. DOI:10.1007/978-3-319-61581-3

Arciniega, J. L. O., Carrió, F., & Valero, A. (2019). FPGA implementation of a deep learning algorithm for real-time signal reconstruction in particle detectors under high pile-up conditions. *Journal of Instrumentation*, *14*(9). DOI:10.1088/1748-0221/14/09/P09002

Arroyo, D., Diaz, J., & Gayoso, V. (2015). On the difficult tradeoff between security and privacy: Challenges for the management of digital identities. *Advances in Intelligent Systems and Computing*, *369*, 455–462. DOI:10.1007/978-3-319-19713-5_39

Artila, V., & Lloyd, D. (2014). Subjective time: The philosophy, psychology, and neuroscience of temporality.

Artunedo, A., Villagra, J., & Godoy, J. (2019). Real-Time Motion Planning Approach for Automated Driving in Urban Environments. *IEEE Access*, *7*, 180039–180053. DOI:10.1109/ACCESS.2019.2959432

Backus, J. (1978). Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs. *Communications of the ACM*, *21*(8), 613–641. DOI:10.1145/359576.359579

Bair, J., Bellovin, S., Manley, A. et al. (2017). That Was Close! Reward Reporting of Cybersecurity "Near Misses." *Colorado Technology Law, 16, 327*

Ball, K., Degli Esposti, S., Dibb, S. et al. (2019). Institutional trustworthiness and national security governance: Evidence from six European countries. *Governance*, *32*(1), 103–121. DOI:10.1111/gove.12353

Ball, N. M., & Brunner, R. J. (2009). Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, *19*(7), 1049–1106. DOI:10.1142/S0218271810017160

Banks, D. L., Rios, J., & Insua, D. R. (2015). Adversarial risk analysis. *Adversarial Risk Analysis*. DOI:10.1201/b18653

Bansal, S., & Tomlin, C. J. (2019). Control and Safety of Autonomous Vehicles with Learning-Enabled Components. *Springer, Cham*, 57-75. DOI:10.1007/978-3-319-97301-2_4

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. DOI:10.1146/annurev.psych.59.103006.093639

Batty, M. (2018). *Inventing Future Cities.* MIT Press. https://books.google.es/books/about/Inventing_Future_Cities.html?id=zs0StQEACAAJ&redir_esc=y

Bengio, Y., Lodi, A., & Prouvost, A. (2018). Machine Learning for Combinatorial Optimization: a Methodological Tour d'Horizon. *European Journal of Operational Research*, *290*(2), 405–421. http://arxiv.org/abs/1811.06128

Bent, R., & Van Hentenryck, P. (2018). Online Stochastic Combinatorial Optimization. *The MIT Press*. DOI:10.7551/mitpress/5140.001.0001

Bergen, K. J., Johnson, P. A., De Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363 (6433). DOI:10.1126/science.aau0323

Berger, J. O., Insua, D. R., & Ruggeri, F. (2000). Robust Bayesian Robustness. *Lecture Notes in Statistics*, 152. DOI:10.1007/978-1-4612-1306-2_1

Bergey, G. K., Morrell, M. J., Mizrahi, E. M. et al. (2015). Long-term treatment with responsive brain stimulation in adults with refractory partial seizures. *Neurology*, *84*(8), 810–817. DOI:10.1212/WNL.0000000000001280

Billari, F., & Zagheni, E. (2017). Big Data and Population Processes: A Revolution? DOI:10.31235/osf.io/f9vzp

Bing, Z., Meschede, C., Röhrbein, F. et al. (2019). A survey of robotics control based on learning-inspired spiking neural networks. *Frontiers in Neurorobotics*, *12*, 35. DOI:10.3389/fnbot.2018.00035

Bischoff, R., & Guhl, T. (2010). The strategic research agenda for robotics in Europe. *IEEE Robotics and Automation Magazine*, *17*(1), 15–16. DOI:10.1109/MRA.2010.935802

Bisio, F., Saeli, S., Lombardo, P. et al. (2017). Real-time behavioral DGA detection through machine learning. *Proceedings - International Carnahan Conference on Security Technology*, 1–6. DOI:10.1109/CCST.2017.8167790

Bistaffa, F., Blum, C., Cerquides, J. et al. (2019). A Computational Approach to Quantify the Benefits of Ridesharing for Policy Makers and Travellers. *IEEE Transactions on Intelligent Transportation Systems, 1-12.*

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. DOI:10.1080/01621459.2017.1285773

Blum, C., & Roli, A. (2003). Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys*, 35(3), 268–308. DOI:10.1145/937503.937505

Blum, C., Pinacho, P., López-Ibáñez, M., & Lozano, J. A. (2016). Construct, Merge, Solve & Adapt: A New General Algorithm For Combinatorial Optimization. *Computers & Operations Research*, 68, 75–88. DOI:10.1016/j.cor.2015.10.014

Bost, R., Ada Popa, R., Tu, S., & Goldwasser, S. (2015). Machine Learning Classification over Encrypted Data.

Bonet, B., Palacios, H., & Geffner, H. (2009). Automatic Derivation of Memoryless Policies and Finite-State Controllers Using Classical Planners. *ICAPS*.

Brachman, R. J., & Levesque, H. J. (2004). Knowledge Representation and Reasoning. *Elsevier Inc*. DOI:10.1016/B978-1-55860-932-7.X5083-3

Brook, A., & DeVidi, R. C. (2001). Self-reference and Self-awareness. *Benjamins*.

Bruña, R., Maestú, F., & Pereda, E. (2018). Phase locking value revisited: Teaching new tricks to an old dog. *Journal of Neural Engineering*, *15*(5). DOI:10.1088/1741-2552/aacfe4

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine*, 26 (4), 53. DOI: 10.1609/aimag.v26i4.1848

Buchanan, B., & Sutherland, G. (1969). Heuristic Dendral: A program for generating explanatory hypotheses in organic chemistry. *Machine Intelligence, 4*, 209–254.

Buchanan, B. G., & Shortliffe, E. H. (1984). Rule-based expert system. *Reading: Addison-Wesley*

Byrd, R. H., Hansen, S. L., Nocedal, J., & Singer, Y. (2016). A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, *26*(2), 1008–1031. DOI:10.1137/140954362

Caliskan-Islam, A., Greenstadt, R., Harang, R. et al. (2015). De-anonymizing Programmers via Code Stylometry. *24th {USENIX} Security Symposium,* 255–270.

Campbell, M., Hoane, A. J., & Hsu, F. H. (2002). Deep Blue. *Artificial Intelligence, 134*(1–2), 57–83. DOI:10.1016/S0004-3702(01)00129-1

Camuñas-Mesa, L. A., Linares-Barranco, B., & Serrano-Gotarredona, T. (2019). Neuromorphic spiking neural networks and their memristor-CMOS hardware implementations. *Materials*, 12 (17), 2745. DOI:10.3390/ma12172745

Canal, G., Alenyà, G., Torras, C. et al. (2019). Adapting robot task planning to user preferences: An assistive shoe dressing example. *Autonomous Robots*, 43 (6), 1343–1356.

Castelló, X., Eguíluz, V. M., & San Miguel, M. (2006). Ordering dynamics with two non-excluding options: Bilingualism in language competition. *New Journal of Physics*, *8*(12), 308. DOI:10.1088/1367-2630/8/12/308

Cegielski, W. H., & Rogers, J. D. (2016). Rethinking the role of Agent-Based Modeling in archaeology. *Journal of Anthropological Archaeology*, *41*, 283–298. DOI:10.1016/j.jaa.2016.01.009

Centola, D., González-Avella, J. C., Eguíluz, V. M., & San Miguel, M. (2007). Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups. *Journal of Conflict Resolution*, *51*(6), 905–929. DOI:10.1177/0022002707307632

Chatzilygeroudis, K., Vassiliades, V., Stulp, F. et al. (2020). A survey on policy search algorithms for learning robot controllers in a handful of trials. *IEEE Transactions on Robotics*, *36*(2). DOI:10.1109/TRO.2019.2958211ï

Chi, P., Li, S., Xu, C. et al. (2016). PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory. *Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016*, 27–39. DOI:10.1109/ISCA.2016.13

Christoph Posch, B., Member IEEE, S., Serrano-Gotarredona, T. et al. (2014). Retinomorphic Event-Based Vision Sensors: Bioinspired Cameras With Spiking Output. *Proceedings of the IEEE*, 102, 1470–1484 DOI:10.1109/JPROC.2014.2346153

Chuyakin, A., Schmidt, K., Phillips, C. (2012). Logging and log management: the authoritative guide to understanding the concepts surrounding logging and log management. *Newnes.*

Cintula, P, Fermüller, C., & Noguera, C. (2015). Handbook of mathematical fuzzy logic-volume 3. *College Publications.*

Cintula, P., Hájek, P., & Noguera, C. (2011). Handbook of Mathematical Fuzzy Logic - volume 1. *College Publications*.

Cintula, P., Hájek, P., & Noguera, C. (2011). Studies in Logic, Mathematical Logic and Foundations. *College Publications.*

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, 3(9), 345–351. DOI:10.1016/S1364-6613(99)01361-3

Collobert, R., Weston, J., Com, J. et al. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, 2493–2537.

Colomé, A., & Torras, C. (2020). Reinforcement Learning of Bimanual Robot Skills. *Springer International Publishing,* 134. DOI:10.1007/978-3-030-26326-3

Comiter, M. (2019). Attacking Artificial Intelligence AI's Security Vulnerability and What Policymakers Can Do About It. *Belfer Center Paper.*

Conte, R., Gilbert, N., Bonelli, G. et al. (2012). Manifesto of computational social science. *European Physical Journal: Special Topics*, *214*(1), 325–346. DOI:10.1140/epjst/e2012-01697-8

Confalonieri, R., Pease, A., Schorlemmer, M. et al. (2018). Concept Invention: Foundations, Implementation, Social Aspects and Applications. *Springer, series in Computational Synthesis and Creative Systems.* DOI:10.1007/978-3-319-65602-1

Coral (2019). Edge TPC. https://coral.withgoogle.com/docs/edgetpu/faq/

Cortes, J., & Egerstedt, M. (2017). Coordinated Control of Multi-Robot Systems: A Survey. *SICE Journal of Control, Measurement, and System Integration*, *10*(6), 495–503. DOI:10.9746/jcmsi.10.495

Cortet, M., Rijks, T., & Nijland, S. (2016). PSD2 : the digital transformation accelerator for banks. *Journal of Payments Strategy & Systems*, *10*(1).

Cowan, R. S. (1987). The Whale and the Reactor: A Search for Limits in an Age of High Technology. Langdon Winner. *Isis*, *78*(2), 296–297. DOI:10.1086/354443

Curry, S., Kirda, E., Schwarts, E. et al. (2013). Big Data Fuels Intelligence-Driven Security, *RSA Security Brief.*

Czaplicka, A., Toral, R., & San Miguel, M. (2016). Competition of simple and complex adoption on interdependent networks. *Physical Review E*, *94*(6). DOI:10.1103/PhysRevE.94.062301

Dabney, W., Kurth-Nelson, Z., Uchida, N. et al. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. DOI:10.1038/s41586-019-1924-6

Dalvi, N., Domingos, P., Mausam, S. et al. (2004). Adversarial classification. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 99–108. DOI:10.1145/1014052.1014066

Darwiche, A. (2018). Human-level intelligence or animal-like abilities? *Communications of the ACM*, *61*(10), 56–67. DOI:10.1145/3271625

Dastani, M., El Fallah Segrouchni, A., Leite, J., & Torroni, P. (2010). Languages, Methodologies, and Development Tools for Multi-Agent Systems. *Springer Berlin Heidelberg*, 6039. DOI:10.1007/978-3-642-13338-1

De la Torre-Abaitua, G., Lago-Fernández, L. F., & Arroyo, D. (2021). On the application of compression-based metrics to identifying anomalous behaviour in web traffic. *Logic Journal of the IGPL*, *28*(4), 546–557. DOI:10.1093/JIGPAL/JZZ062

De Santos, P. G., Garcia, E., Ponticelli, R., & Armada, M. (2009). Minimizing energy consumption in hexapod robots. *Advanced Robotics*, *23*(6), 681–704. DOI:10.1163/156855309X431677

Deb, K., & Myburgh, C. (2016). Breaking the Billion-Variable Barrier in real-world optimization using a customized evolutionary algorithm. *Proceedings of the 2016 Genetic and Evolutionary Computation Conference*, 653–660. DOI:10.1145/2908812.2908952

Delgado-Restituto, M., Romaine, J. B., & Rodríguez-Vázquez, Á. (2019). Phase Synchronization Operator for On-Chip Brain Functional Connectivity Computation. *IEEE Transactions on Biomedical Circuits and Systems*, *13*(5), 957–970. DOI:10.1109/TBCAS.2019.2931799

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 373(1755). DOI:10.1098/rstb.2017.0342

Diaz, J., Choi, S. G., Arroyo, D. et al. (2019). A methodology for retrofitting privacy and its application to e-shopping transactions. *Advances in Cyber Security: Principles, Techniques, and Applications*. 143–183. DOI:10.1007/978-981-13-1483-4_7

Diaz, J., Choi, S. G., Arroyo, D. et al. (2018). Privacy in e-Shopping Transactions: Exploring and Addressing the Trade-Offs. *International Symposium on Cyber Security Cryptography and Machine Learning,* 206–226.

Diehl, P. U., Neil, D., Binas, J. et al. (2015). Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. *Proceedings of the International Joint Conference on Neural Networks*. DOI:10.1109/IJCNN.2015.7280696

Dobchev, D., Pillai, G., & Karelson, M. (2014). In Silico Machine Learning Methods in Drug Development. *Current Topics in Medicinal Chemistry*, *14*(16), 1913–1922. DOI:10.2174/1568026614666140929124203

Dwork, C. (2013). Differential privacy: A cryptographic approach to private data analysis. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 296–322. DOI:10.1017/CBO9781107590205.018

D'Inverno, M., Luck, M., Noriega, P. et al. (2012). Communicating open systems. *Artificial Intelligence*, *186*, 38–94. DOI:10.1016/j.artint.2012.03.004

Eguíluz, V. M., Zimmermann, M. G., Cela-Conde, C. J., & San Miguel, M. (2005). Cooperation and the emergence of role differentiation in the dynamics of social networks. *American Journal of Sociology*, *110*(4), 977–1008. DOI:10.1086/428716

Ekin, T., Naveiro, R., Torres-Barrán, A., & Ríos-Insua, D. (2019). Augmented Probability Simulation Methods for Non-cooperative Games. *ArXiv*

Esteva, M., Rodríguez-Aguilar, J.-A., Sierra, C. et al. (2001). On the Formal Specification of Electronic Institutions. *Springer Berlin Heidelberg*. 126-147. DOI:10.1007/3-540-44682-6_8

EU (2014). Electronic identification, authentication and trust services. https://eurlex.europa.eu/legal-content/en/txt/pdf/?uri=celex:32014r0910&from=en.

ExtremeTech. (2019). Intel Details Its Nervana Inference and Training AI Cards. https://www.extremetech.com/computing/ 296990 -intel-nervana-nnp-i-nnp-t-a-training-inference.

Farabet, C., Paz, R., Pérez-Carrasco, J. et al. (2012). Comparison between frame-constrained fix-pixel-value and frame-free spiking-dynamic-pixel convNets for visual processing. *Frontiers in Neuroscience, APR*. DOI:10.3389/fnins.2012.00032

Faratin, P., Sierra, C., & Jennings, N. R. (1998). Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems, 24*(3–4), 159–182. DOI:10.1016/S0921-8890(98)00029-3

Farinelli, A., Iocchi, L., & Nardi, D. (2004). Multi-Robot Systems: A classification focused on coordination. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(5),2015–2028.

Feigenbaum, E., Feldman, J. (1963). Computers and Thought. *McGraw Hill.*

Fernández-Gracia, J., Suchecki, K., Ramasco, J. J. et al. (2014). Is the Voter Model a Model for Voters? *Physical Review Letters, 112*(15). DOI:10.1103/PhysRevLett.112.158701

Fernández, R., Montes, H., Salinas, C. et al. (2013). Combination of RGB and multispectral imagery for discrimination of Cabernet Sauvignon grapevine elements. *Sensor, 13*(6), 7838–7859. DOI:10.3390/s130607838

Fernández, R., Montes, H., Surdilovic, J. et al. (2018). Automatic detection of field-grown cucumbers for robotic harvesting. *IEEE Access, 6*, 35512–35526. DOI:10.1109/ACCESS.2018.2851376

Ferrucci, D., Brown, E., Chu-Carroll, J. et al. (2011). Building watson: An overview of the deepQA project. *AI Magazine, 31*(3), 59–79. DOI:10.1609/aimag.v31i3.2303

Finck, M. (2018). Blockchains and Data Protection in the European Union. *European Data Protection Law Review, 4*(1), 17–35. DOI:10.21552/edpl/2018/1/6

Flaminio, T., Godo, L., & Hosni, H. (2017). On boolean algebras of conditionals and their logical counterpart. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10369 LNAI*, 246–256. DOI:10.1007/978-3-319-61581-3_23

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review, 1*(1). DOI:10.1162/99608f92.8cd550d1

Fodor, J. A. (1975). *The Language of Thought.* Harvard University Press.

French, S., & Rios, D. (2007). Statistical Decision Theory, *Wiley.* DOI:10.1007/978-0-387-73194-0_3

Furber, S. (2016). Large-scale neuromorphic computing systems. *Journal of Neural Engineering*, 13(5). DOI:10.1088/1741-2560/13/5/051001

Futami, F., Sato, I., & Sugiyama, M. (2018). Variational Inference based on Robust Divergences. *International Conference on Artificial Intelligence and Statistics*, 813-822

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, 94(5), 978–984. DOI:10.1016/j.neuron.2017.05.025

Gallego, V., Naveiro, R., & Insua, D. R. (2019). Reinforcement learning under threats. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 9939–9940. DOI:10.1609/aaai.v33i01.33019939

Garcia-Camacho, I., Alenya, G., Kragic, D. et al. (2020). Benchmarking Bimanual Cloth Manipulation. *IEEE Robotics and Automation Letters, 5*(2), 1111–1118. DOI:10.1109/LRA.2020.2965891

García-Pérez, L., García-Alegre, M. C., Ribeiro, A., & Guinea, D. (2008). An agent of behaviour architecture for unmanned control of a farming vehicle. *Computers and Electronics in Agriculture, 60*(1), 39–48. DOI:10.1016/j.compag.2007.06.004

Geffner, H., & Bonet, B. (2013). A concise introduction to models and methods for automated planning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 22*, 1–141. DOI:10.2200/S00513ED1V01Y201306AIM022

Ghallab, M., Nau, D., & Traverso, P. (2016). Automated Planning and Acting. *Cambridge University Press*. DOI:10.1017/CBO9781139583923

Ghosh-Dastidar, S., & Adeli, H. (2009). Third generation neural networks: Spiking neural networks. *Advances in Intelligent and Soft Computing, 61*, 167–178. DOI:10.1007/978-3-642-03156-4_17

Godoy, J., Pérez, J., Onieva, E. et al. (2015). A driverless vehicle demonstration on motorways and in urban environments. *Transport, 30*(3), 253–263. DOI:10.3846/16484142.2014.1003406

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies, 10*(1), 1–311. DOI:10.2200/S00762ED1V01Y201703HLT037

Gómez Esteban, P., Liu, S., Ríos Insua, D., & González-Ortega, J. (2020). Competition and cooperation in a community of autonomous agents. *Autonomous Robots, 44*(3–4), 533–546. DOI:10.1007/s10514-019-09867-y

González-Ruibal, A., Chilton, E., Kristiansen, K., & Niklasson, E. (2014). Towards a new paradigm? The Third Science Revolution and its Possible Consequences in Archaeology. *Current Swedish Archaeology, 22*(4), 11-71.

Goodfellow, I., Bnegio, Y., & Courville, A. (2018). Deep learning. *Genetic Programming and Evolvable Machines, 19*(1–2), 305–307. DOI:10.1007/s10710-017-9314-z

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M. et al. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations-Conference Track Proceedings*.

Guggenmos, M., Schmack, K., Veer, I. M. et al. (2020). A multimodal neuroimaging classifier for alcohol dependence. *Scientific Reports, 10*(1). DOI:10.1038/s41598-019-56923-9

Guiochet, J., Machin, M., & Waeselynck, H. (2017). Safety-critical advanced robots: A survey. *Robotics and Autonomous Systems, 94*, 43–52. DOI:10.1016/j.robot.2017.04.004

Hall D., Llinas J., Liggins II, M., Poularikas, A. (2017). Handbook of Multisensor Data Fusion. *CRC Press.* DOI:10.1201/9781420053098

Halyo, V., Hunt, A., Jindal, P. et al. (2013). GPU Enhancement of the Trigger to Extend Physics Reach at the LHC. *Journal of Instrumentation, 8*(10). DOI:10.1088/1748-0221/8/10/P10005

Hamrick, J. B., Ballard, A. J., Pascanu, R. et al. (2017). Metacontrol for Adaptive Imagination-Based Optimization. *5th International Conference on Learning Representations*

Hanea, A. M., Nane, G. F., Bedford, T., & French, S. (2021). *Expert Judgement in Risk and Decision Analysis*. Springer International Publishing. DOI:10.1007/978-3-030-46474-5

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics, 4*(2), 100–107. DOI:10.1109/TSSC.1968.300136

Havens, J., Bielby, J., & Alvarez, M. A. P. (2019). Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. *IEEE* https://www.academia.edu/38746211/ETHICALLY_ALIGNED_DESIGN_A_Vision_for_Prioritizing_Human_Wellbeing_with_Artificial_Intelligence_and_Autonomous_Systems

Hettwer, B., Gehrer, S., & Güneysu, T. (2020). Applications of machine learning techniques in side-channel attacks: a survey. *Journal of Cryptographic Engineering, 10*(2), 135–162. DOI:10.1007/s13389-019-00212-8

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780. DOI:10.1162/neco.1997.9.8.1735

Hodges, A. (2014). Alan Turing: The Enigma. *Penguin Books Australia.*

Hooker, J. N. (2019). Logic-Based Benders Decomposition for Large-Scale Optimization. *Springer Optimization and Its Applications*, 149, 1–26. DOI:10.1007/978-3-030-22788-3_1

Hu, Y., & De Giacomo, G. (2011). Generalized Planning: Synthesizing Plans that Work for Multiple Environments. *Twenty-Second International Joint Conference on Artificial Intelligence.*

Huang, Y., He, X., & Dai, H. (2015). Systematization of metrics in intrusion detection systems. *ACM International Conference Proceeding Series*. DOI:10.1145/2746194.2746222

Iglesias García, J., Diaz, J., & Arroyo, D. (2020). Hyot: Leveraging Hyperledger for Constructing an Event-Based Traceability System in IoT. *Advances in Intelligent Systems and Computing, 951*, 195–204. DOI:10.1007/978-3-030-20005-3_20

ISO (2016). ISO/TC 307 Blockchain and distributed ledger technologies. https://www.iso.org/committee/6266604.html.

Jenny Gu, & Philip E. Bourne. (2009). *Structural Bioinformatics, 4.*

Jevtic, A., Flores Valle, A., Alenya, G. et al. (2019). Personalized Robot Assistant for Support in Dressing. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(3), 363–374. DOI:10.1109/TCDS.2018.2817283

Jimenez, A. R., Seco, F., & Torres-Sospedra, J. (2019). Tools for smartphone multi-sensor data registration and GT mapping for positioning applications. *International Conference on Indoor Positioning and Indoor Navigation, 1-8*. DOI:10.1109/IPIN.2019.8911784

Jiruska, P., de Curtis, M., Jefferys, J. G. R. et al. (2013). Synchronization and demapping: Controversies and hypotheses. The *Journal of Physiology*, 591(4), 787–797. DOI:10.1113/jphysiol.2012.239590

Jiruska, P., de Curtis, M., Jefferys, J. G. R. et al. (2013). Synchronization and desynchronization in epilepsy: Controversies and hypotheses. *Journal of Physiology*, 591(4), 787–797). DOI:10.1113/jphysiol.2012.239590

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134. DOI:10.1016/s0004-3702(98)00023-x

Kennedy, P. (2019). Huawei Ascend 910 Provides a NVIDIA AI Training Alternative.

Khandpur, R. P., Ji, T., Jan, S. et al. (2017). Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media. *Proceedings of the International Conference on Information and Knowledge Management,* 1049–1057. DOI:10.1145/3132847.3132866

Kim, K., & Aminanto, M. E. (2018). Deep learning in intrusion detection perspective: Overview and further challenges. *2017 International Workshop on Big Data and Information Security*, 5–10. DOI:10.1109/IWBIS.2017.8275095

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Knyazeva, M. G., Carmeli, C., Khadivi, A. et al. (2013). Evolution of source EEG synchronization in early Alzheimer's disease. *Neurobiology of Aging*, *34*(3), 694–705. DOI:10.1016/j.neurobiolaging.2012.07.012

Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. *MIT Press*

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94. DOI:10.1007/s10462-018-9646-y

Kowalski, R. (1979). Logic for Problem Solving. *North Holland*

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. In *Nature Neuroscience,* 21(9), 1148–1160. DOI:10.1038/s41593-018-0210-5

Krutz, R. L., & Dean Vines, R. (2002). The CISSP Prep Guide, Gold Edition. *John Wiley & Sons*.

Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh: the embodied mind and its challenge to Western thought. *Basic Books*.

Lakshmanan, V., Gilleland, E., McGovern, A., Tingley, M. (2015). Machine learning and data mining approaches to climate science. *Proceedings of the 4th International Workshop on Climate Informatics*. DOI:10.1007/978-3-319-17220-0

Langley, P. (2017). Progress and Challenges in Research on Cognitive Architectures. *Proceedings of the AAAI Conference on Artificial Intelligence,* 31(1), 4870–4876

Lappas, T., Liu, K., & Terzi, E. (2009). Finding a team of experts in social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 467–475. DOI:10.1145/1557019.1557074

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature,* 521, 436–444. DOI:10.1038/nature14539

Lee, J., Kim, C., Kang, S. et al. (2019). UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision. *IEEE Journal of Solid-State Circuits*, *54*(1), 173–185. DOI:10.1109/JSSC.2018.2865489

Lee, J., Lee, J., Han, D. et al. (2019). 7.7 LNPU: A 25.3TFLOPS/W Sparse Deep-Neural-Network Learning Processor with Fine-Grained Mixed Precision of FP8-FP16. *IEEE International Solid-State Circuits Conference*, 142–144. DOI:10.1109/ISSCC.2019.8662302

Lepri, B., Oliver, N., Letouzé, E. et al. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*, *31*(4), 611–627. DOI:10.1007/s13347-017-0279-x

Li, X., Deng, S., Wang, S. et al. (2018). Review of Small Data Learning Methods. *Proceedings - International Computer Software and Applications Conference*, *2*, 106–109. DOI:10.1109/COMPSAC.2018.10212

Li, X., Neil, D., Delbruck, T., & Liu, S. C. (2019). Lip reading deep network exploiting multi-modal spiking visual and auditory sensors. *Proceedings - IEEE International Symposium on Circuits and Systems*, 1-5. DOI:10.1109/ISCAS.2019.8702565

Liakos, K. G., Busato, P., Moshou, D. et al. (2018). Machine learning in agriculture: A review. *Sensors* 18(8):2674. DOI:10.3390/s18082674

Liese, F., & Miescke, K.-J. (2007). Statistical Decision Theory, 1-52. DOI:10.1007/978-0-387-73194-0_3

Lighthill, J. (1973). Artificial intelligence: A general survey. *Artificial Intelligence: a paper symposium*.

Linares-Barranco, B., & Serrano-Gotarredona, T. (2009). Exploiting memristance in adaptive asynchronous spiking neuromorphic nanotechnology systems, 601-604.

Lison, P., & Mavroeidis, V. (2017). Automatic Detection of Malware-Generated Domains with Recurrent Neural Models. *Arxiv*.

Litman, T. (2017). Autonomous vehicle implementation predictions. *Victoria Transport Policy Institute*

London, M., & Häusser, M. (2005). Dendritic computation. *Annual Review of Neuroscience*, 28, 503–532. DOI:10.1146/annurev.neuro.28.061604.135703

López, D., Monasterio, A., Toboso, M. et al. (2020). Cartography of the Values Involved in Robotics. *Biosystems and Biorobotics*, 25, 98–104.DOI:10.1007/978-3-030-24074-5_18

Love, B. C. (2016). Cognitive Models as Bridge between Brain and Behavior. *Trends in Cognitive Sciences*, 20(4), 247–248. DOI:10.1016/j.tics.2016.02.006

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149. DOI:10.1016/j.cosrev.2009.03.005

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671. DOI:10.1016/S0893-6080(97)00011-7

Maghrebi, H. (2019). *Deep Learning based Side Channel Attacks in Practice*. *IACR Cryptol. ePrint Arch.,* 578.

Margarit-Taulé, J. M., Giménez-Gómez, P., Escudé-Pujol, R. et al. (2019). Live demonstration: A portable microsensor fusion system with real-time measurement for on-site beverage tasting. *Proceedings - IEEE International Symposium on Circuits and Systems.* DOI:10.1109/ISCAS.2019.8702184

Martínez-Rodríguez, M. C., Prada-Delgado, M. A., Brox, P., & Baturone, I. (2018). VLSI design of trusted virtual sensors. *Sensors, 18*(2). DOI:10.3390/s18020347

Martinez, D., Alenya, G., Ribeiro, T. et al. (2017). Relational Reinforcement Learning for Planning with Exogenous Effects. *Journal of Machine Learning Research*, 18(78),1–44,

Martínez, D., Alenyà, G., & Torras, C. (2017). Relational reinforcement learning with guided demonstrations. *Artificial Intelligence*, *247*, 295–312. DOI:10.1016/j.artint.2015.02.006

Mattila, J., Koivumaki, J., Caldwell, D. G., & Semini, C. (2017). A survey on control of hydraulic robotic manipulators with projection to future trends. In *IEEE/ASME Transactions on Mechatronics,* 22(2), 669–680. DOI:10.1109/TMECH.2017.2668604

McCorduck, P., & Cfe, C. (2004). Machines Who Think. *Machines Who Think*. DOI:10.1201/9780429258985

Milanés, V., Villagrá, J., Pérez, J., & González, C. (2012). Low-speed longitudinal controllers for mass-produced cars: A comparative study. *IEEE Transactions on Industrial Electronics*, *59*(1), 620–628. DOI:10.1109/TIE.2011.2148673

Minzioni, P., Lacava, C., Tanabe, T. et al. (2019). Roadmap on all-optical processing. *Journal of Optics (United Kingdom)*, *21*(6), 063001. DOI:10.1088/2040-8986/ab0e66

Mitchell, T. (1997). Machine Learning. *McGraw Hill*.

Mitchie, D. (1979). Expert Systems in the Microelectronics Age. *Edinburgh University Press*

Monasterio Astobiza, A., Toboso, M., Aparicio, M. et al. (2019). Bringing inclusivity to robotics with INBOTS. *Nature Machine Intelligence*, 1(4), 164. DOI:10.1038/s42256-019-0040-5

Müller, J. P., & Fischer, K. (2014). Application Impact of Multiagent Systems and Technologies: A Survey. *Springer*

Muñoz-Martín, I., Bianchi, S., Pedretti, G. et al. (2019). Unsupervised Learning to Overcome Catastrophic Forgetting in Neural Networks. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, *5*(1), 58–66. DOI:10.1109/JXCDC.2019.2911135

Naveiro, R., & Insua, D. R. (2019). Gradient Methods for Solving Stackelberg Games. *International Conference on Algorithmic Decision Theory*, 126–140.

Newell, A., & Simon, H. A. (2011). Computer science as empirical inquiry: symbols and search. *ACM Turing Award Lectures*, 19,113–126. DOI:10.1145/1283920.1283930

Newell, A., & Simon, H. (1972). Human Problem Solving. *Prentice-Hall*

O'Connor, P., Neil, D., Liu, S. C. et al. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in Neuroscience*, *7*, *178*. DOI:10.3389/fnins.2013.00178

Olivares-Alarcos, A., Foix, S., & Alenyà, G. (2019). On inferring intentions in shared tasks for industrial collaborative robots. *Electronics, 8*(11). DOI:10.3390/electronics8111306

Osborne, M. J., & Rubinstein, A. (1994). A Course in Game Theory.

Osman, N. (2018). Partakable technology. *IJCAI International Joint Conference on Artificial Intelligence*, 5714–5718. DOI:10.24963/ijcai.2018/815

Osterweil, L. J., Ghezzi, C., Kramer, J., & Wolf, A. L. (2008). Determining the Impact of Software Engineering Research on Practice. *Computer*, *41*(3), 39–49. DOI:10.1109/MC.2008.85

Pack Kaelbling, L., Littman, M. L., Moore, A. W., & Hall, S. (1996). Reinforcement Learning: A Survey. *Journal of Artiicial Intelligence Research*, 4, 237–285.

Parsons, S., Sierra, C., & Jennings, N. (1998). Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, *8*(3), 261–292. DOI:10.1093/logcom/8.3.261

Pasquale, F. (2015). The Black Box Society. *Harvard University Press.*

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925. DOI:10.1103/RevModPhys.87.925

Pathinarupothi, R. K., Durga, P., & Rangan, E. S. (2018). Data to diagnosis in global health: A 3P approach. *BMC Medical Informatics and Decision Making*, *18*(1). DOI:10.1186/s12911-018-0658-y

Pearl, J. (1984). Heuristics. *Addison-Wesley Publishing Company*

Pearl, J., Glymour, M. & Jewell, N. P. (2016). Causal Inference in Statistics: A Primer. *Wiley*

Peralta, A. F., Carro, A., San Miguel, M., & Toral, R. (2018). Analytical and numerical study of the non-linear noisy voter model on complex networks. *Chaos*, *28*(7), 075516. DOI:10.1063/1.5030112

Pérez-Carrasco, J. A., Zhao, B., Serrano, C. et al. (2013). Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing - Application to feedforward convnets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2706–2719. DOI:10.1109/TPAMI.2013.71

Polydoros, A. S., & Nalpantidis, L. (2017). Survey of Model-Based Reinforcement Learning: Applications on Robotics. *Journal of Intelligent and Robotic Systems: Theory and Applications*, *86*(2), 153–173. DOI:10.1007/s10846-017-0468-y

Port, R. F., & van Gelder, T. (1995). Mind as Motion: Explorations in the Dynamics of Cognition. *The MIT Press.*

Prosser, S. (2016). Experiencing Time. *Oxford University Press.*

Querejeta-Azurmendi, I., Hernandez Encinas, L., Arroyo, D., & Hernández-Ardieta, J. L. (2020). An Internet Voting Proposal Towards Improving Usability and Coercion Resistance. *Advances in Intelligent Systems and Computing*, *951*, 155–164. DOI:10.1007/978-3-030-20005-3_16

Quijada, R., Dura, R., Pallares, J., Formatje, X. et al. (2018). Large-Area Automated Layout Extraction Methodology for Full-IC Reverse Engineering. *Journal of Hardware and Systems Security*, *2*(4), 322–332. DOI:10.1007/s41635-018-0051-4

Radovic, A., Williams, M., Rousseau, D. et al. (2018). Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560(7716), 41–48. DOI:10.1038/s41586-018-0361-2

Raju, L., Sankar, S., & Milton, R. S. (2015). Distributed optimization of solar micro-grid using multi agent reinforcement learning. *Procedia Computer Science*, *46*, 231–239. DOI:10.1016/j.procs.2015.02.016

Renaudin, V., Ortiz, M., Perul, J. et al. (2019). Evaluating Indoor Positioning Systems in a Shopping Mall: The Lessons Learned from the IPIN 2018 Competition. *IEEE Access*, *7*, 148594–148628. DOI:10.1109/ACCESS.2019.2944389

Rimmer, V., Preuveneers, D., Juárez, M. et al. (2017). Automated Feature Extraction for Website Fingerprinting through Deep Learning. *ArXiv preprint*.

Rivest, R. L. (1991). Cryptography and Machine Learning. *International Conference on the Theory and Application of Cryptology*, 427–439.

Robertson, D., Giunchiglia, F., Van Harmelen, F. et al. (2008). Open knowledge coordinating knowledge sharing through peer-to-peer interaction. *Lecture Notes in Computer Science, 5118 LNAI*, 1–18. DOI:10.1007/978-3-540-85058-8_1

Rumelhart, D., McCleland, J. (1986). Parallel Distributed Processing. *MIT Press*.

Russell, S., Norving, P. (2010). Artificial Intelligence. *Prentice Hall*.

Sabater, J., & Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent System,* 475-482. DOI:10.1145/544741.544854

Sabottke, C., Suciu, O., Dumitraş, T., & Dumitras, T. (2015). Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits. *24th {USENIX} Security Symposium,* 1041-1056.

Sakkalis, V. (2011). Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, *41*(12), 1110–1117. DOI:10.1016/j.compbiomed.2011.06.020

Salathé, M., Bengtsson, L., Bodnar, T. J. et al. (2012). Digital epidemiology. *PLoS Computational Biology*, *8*(7), 1002616. DOI:10.1371/journal.pcbi.1002616

Salazar Gonzalez, J. L., Soria Morillo, L. M., Alvarez-Garcia, J. A. et al. (2019). Energy-Efficient Indoor Localization WiFi-Fingerprint System: An Experimental Study *IEEE Access*, *7*, 162664–162682. DOI:10.1109/ACCESS.2019.2952221

Santamaria-Navarro, A., Loianno, G., Solà, J. et al. (2018). Autonomous navigation of micro aerial vehicles using high-rate and low-cost sensors. *Autonomous Robots*, *42*(6), 1263–1280. DOI:10.1007/s10514-017-9690-5

Savarimuthu, T. R., Buch, A. G., Schlette, C. et al. (2018). Teaching a Robot the Semantics of Assembly Tasks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*(5), 670–692. DOI:10.1109/TSMC.2016.2635479

Schnitzler, A., & Gross, J. (2005). Normal and pathological oscillatory communication in the brain. *Nature Reviews Neuroscience*, 6(4), 285–96. DOI:10.1038/nrn1650

Segovia-Aguas, J., Jiménez, S., & Jonsson, A. (2018). Computing hierarchical finite state controllers with classical planning. *Journal of Artificial Intelligence Research*, *62*, 755–797. DOI:10.1613/jair.1.11227

Segovia-Aguas, J., Jiménez, S., & Jonsson, A. (2019). Computing programs for generalized planning using a classical planner. *Artificial Intelligence*, *272*, 52–85. DOI:10.1016/j.artint.2018.10.006

Shapiro, L. (2011). Embodied Cognition. *Routledge*

Shoshitaishvili, Y., Weissbacher, M., Dresel, L. et al. (2017). Rise of the HaCRS: Augmenting Autonomous Cyber Reasoning Systems with Human Assistance. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 347-362.

Siciliano, B., & Khatib, O. (2016). Robotics and the handbook. *Springer Handbook of Robotics*, 1-6. DOI:10.1007/978-3-319-32552-1_1

Sierra, C. & Dignum, F. (2001). Agent Mediated Electronic Commerce: The European AgentLink Perspective. *Springer-Verlag Berlin Heidelberg*, 126-147.

Simo-Serra, E., Torras, C., & Moreno-Noguer, F. (2017). 3D Human Pose Tracking Priors using Geodesic Mixture Models. *International Journal of Computer Vision*, *122*(2), 388–408. DOI:10.1007/s11263-016-0941-2

Song, J., Cho, Y., Park, J. S. et al. (2019). 7.1 An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC. *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, 130–133. DOI:10.1109/ISSCC.2019.8662476

Srivastava, S., Immerman, N., & Zilberstein, S. (2008). Learning Generalized Plans Using Abstract Counting. *AAAI*, 8, 991-997.

Stamp, M. (2017). Introduction to machine learning with applications in information security. *Chapman & Hall.*

Stark, P. B. (2016). Privacy, Big Data, and the Public Good: Frameworks for Engagement. *The American Statistician*, *70*(1), 119–119. DOI:10.1080/00031305.2015.1068625

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., & Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, *576*(7786), 266–273. DOI:10.1038/s41586-019-1787-x

Stephen Marsland. (2015). Machine Learning: An Algorithmic Perspective. *CRC Pres.*

Stroke, G.W. (1972). Optical computing. *IEEE Spectrum,* 9(12)*, 24-41.*

Su, J., Vargas, D. V., Prasad, S. et al. (2017). Lightweight Classification of IoT Malware based on Image Recognition. *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 2, 664-669.

Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental and Theoretical Artificial Intelligence*, *19*(2), 159–193. DOI:10.1080/09528130701191560

Tang, A., Sethumadhavan, S., & Stolfo, S. (2014). Unsupervised Anomaly-based Malware Detection using Hardware Features. *International Workshop on Recent Advances in Intrusion Detection,109-129.*

Taylor, M. E., & Stone, P. (2009). Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research,*10, 1633-1685.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. DOI:10.1038/381520a0

Thuruthel, G. T., Ansari, Y., Falotico, E., & Laschi, C. (2018). Control Strategies for Soft Robotic Manipulators: A Survey. *Soft Robotics,* 5(2), 149–163. DOI:10.1089/soro.2017.0007

Tzafestas, S. G. (2018). Mobile Robot Control and Navigation: A Global Overview. *Journal of Intelligent and Robotic Systems: Theory and Applications*, *91*(1), 35–58. DOI:10.1007/s10846-018-0805-9

Ucci, D., Aniello, L., & Baldoni, R. (2019). Survey of machine learning techniques for malware analysis. *Computers and Security,* 81, 123–147. DOI:10.1016/j.cose.2018.11.001

Uhlhaas, P. J., & Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, 11(2),100–113. DOI:10.1038/nrn2774

Vacca, J. (2012). Computer and Information Security Handbook.

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *Journal of Philosophy*, *92*(7), 345–381. DOI:10.2307/2941061

van Krieken, E., Acar, E., & van Harmelen, F. (2020). Analyzing Differentiable Fuzzy Logic Operators. *ArXiv*.

Varela, F. J., Thompson, E., Rosch, E., & Kabat-Zinn, J. (2016). The embodied mind: Cognitive science and human experience. *The MIT Press.* DOI:10.29173/cmplct8718

Veloso, M., Carbonell, J., Perez, A. et al. (1995). Integrating planning and learning: The prodigy architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *7*(1), 81–120. DOI:10.1080/09528139508953801

Verde, R., & Petrucci, A. (2017). SIS 2017. Statistics and Data Science: new challenges, new generations. *Firenze University Press*, 114. DOI:10.36253/978-88-6453-521-0

Vernon, D. (2014). Artificial Cognitive Systems : A Primer. *Journal of Chemical Information and Modeling*, 53(9),1–35.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939), 425–428. DOI:10.1126/science.1171990

Vidal, J. (2017). Tsunami of data'could consume one fifth of global electricity by 2025. *Climate Home News*, *11*.

Vitali, S., Glattfelder, J. B., & Battiston, S. (2011). The network of Global corporate control. *PLoS ONE*, *6*(10). DOI:10.1371/journal.pone.0025995

Vithu, P., & Moses, J. A. (2016). Machine vision system for food grain quality evaluation: A review. *Trends in Food Science and Technology*, 56, 13–20. DOI:10.1016/j.tifs.2016.07.011

Vogt, N. (2018). Machine learning in neuroscience. *Nature Methods*, 15(1), 33. DOI:10.1038/nmeth.4549

Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR) A Practical Guide. *Springer International Publishing.*

Vorobeichyk, Y., & Kantarcioglu, M. (2018). Adversarial Machine Learning. *Morgan Clayton.* DOI:10.2200/S00861ED1V01Y201806AIM039

Wallach, H. (2018). Computational social science ? computer science + social data. *Communications of the ACM*, *61*(3), 42–44. DOI:10.1145/3132698

Warman, M. (2009). Green I.T.: How many Google searches does it take to boil a kettle?

Watts, D. J. *(2011).* Everything is Obvious. How Common Sense Fails. *Crown Bussiness.*

Weisberg, S. M., & Newcombe, N. S. (2017). Embodied cognition and STEM learning: overview of a topical collection in CR:PI. *Cognitive Research: Principles and Implications*,2. DOI:10.1186/s41235-017-0071-6

Werbach, K. (2018). The Blockchain and the New Architecture of Trust. *The MIT Press.* DOI:10.7551/mitpress/11449.001.0001

Westberg, M., Zelvelder, A., & Najjar, A. (2019). A historical perspective on cognitive science and its influence on XAI research. *Lecture Notes in Computer Science, 11763,* 205–219. DOI:10.1007/978-3-030-30391-4_12

Westwood, S. J., Iyengar, S., Walgrave, S. et al. (2018). The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research*, *57*(2), 333–354. DOI:10.1111/1475-6765.12228

Yamada, Y., Uchiyama, M., Jobashi, M. et al. (2020). A 20.5 TOPS Multicore SoC with DNN Accelerator and Image Signal Processor for Automotive Applications. *IEEE Journal of Solid-State Circuits*, *55*(1), 120–132. DOI:10.1109/JSSC.2019.2951391

Yang, P., Tang, K., & Yao, X. (2019). A Parallel Divide-and-Conquer-Based Evolutionary Algorithm for Large-Scale Optimization. *IEEE Access*, *7*, 163105–163118. DOI:10.1109/ACCESS.2019.2938765

Yoo, J., Yan, L., El-Damak, D. et al. (2013). An 8-channel scalable EEG acquisition SoC with patient-specific seizure classification and recording processor. *IEEE Journal of Solid-State Circuits*, *48*(1), 214–228. DOI:10.1109/JSSC.2012.2221220

Yu, B., Gray, D. L., Pan, J. et al. (2017). Inline DGA detection with deep networks. *IEEE International Conference on Data Mining Workshops,* 683–692. DOI:10.1109/ICDMW.2017.96

Zhou, Z., Chen, X., Li, E. et al. (2019). Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing. *Proceedings of the IEEE, 107*(8).

Zimmermann, J. B., & Jackson, A. (2014). Closed-loop control of spinal cord stimulation to restore hand function after paralysis. *Frontiers in Neuroscience*, *8*. DOI:10.3389/fnins.2014.00087

CSIC white paper on Artificial Intelligence, Robotics and Data Science sketches a preliminary roadmap for addressing current R&D challenges associated with automated and autonomous machines. More than 50 research challenges investigated all over Spain by more than 150 experts within CSIC are presented in eight chapters. Chapter One introduces key concepts and tackles the issue of the integration of knowledge (representation), reasoning and learning in the design of artificial entities. Chapter Two analyses challenges associated with the development of theories –and supporting technologies– for modelling the behaviour of autonomous agents. Specifically, it pays attention to the interplay between elements at micro level (individual autonomous agent interactions) with the macro world (the properties we seek in large and complex societies). While Chapter Three discusses the variety of data science applications currently used in all fields of science, paying particular attention to Machine Learning (ML) techniques, Chapter Four presents current development in various areas of robotics. Chapter Five explores the challenges associated with computational cognitive models. Chapter Six pays attention to the ethical, legal, economic and social challenges coming alongside the development of smart systems. Chapter Seven engages with the problem of the environmental sustainability of deploying intelligent systems at large scale. Finally, Chapter Eight deals with the complexity of ensuring the security, safety, resilience and privacy-protection of smart systems against cyber threats.

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
CSIC
EDITORIAL CSIC