European Commission

JRC TECHNICAL REPORT

AI Watch
Assessing Technology Readiness Levels for Artificial Intelligence

EUR 30401 EN

Joint Research Centre

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information
Name: Fernando Martínez Plumed
Address: JRC Seville, Edificio Expo
Inca Garcilaso, 3
41092 Sevilla
Email: Fernando.MARTINEZ-PLUMED@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC122014

EUR 30401 EN

PDF ISBN 978-92-76-22987-2 ISSN 1831-9424 doi:10.2760/15025

Luxembourg: Publications Office of the European Union, 2020.

How to cite this report: Martínez-Plumed, F., Gómez, E., Hernández-Orallo, J., *AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence*, EUR 30401 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-22987-2, doi:10.2760/15025, JRC122014

Contents

Acknowledgements

The following researchers constitute the panel of experts that provided valuable comments, suggestions and useful critiques for this work (in alphabetical order): Carlos Carrascosa (Universitat Politècnica de València – Robotics and urban mobility), Blagoj Delipetrev (European Commission – Image Processing), Paul Desruelle (European Commission – Information and Communications Technologies), Salvador España (Universitat Politècnica de València – Text and speech recognition), Cèsar Ferri (Universitat Politècnica de València – Machine Learning), Ross Gruetzemacher (Auburn University – AI progress & transformative AI), Stella Heras (Universitat Politècnica de València – Multi-Agent Systems), Filipe Jones Mourao (European Commission – Artificial Intelligence and Robotics) – Alfons Juan (Universitat Politècnica de València – Language Processing), Carlos Monserrat (Universitat Politècnica de València – Machine learning and image processing), Daniel Nepelsky (European Commission – Technology innovation), Eva Onaindia (Universitat Politècnica de València – AI planning), Barry O'Sullivan (University College Cork – Artificial Intelligence and public policy), MªJosé Ramírez-Quintana (Universitat Politècnica de València – Machine Learning), Miguel Ángel Salido (Universitat Politècnica de València – Scheduling systems) and Laura Sebastià (Universitat Politècnica de València – Machine Learning).

Authors

Fernando Martínez-Plumed, Joint Research Centre, European Commission

Emilia Gómez, Joint Research Centre, European Commission

José Hernández-Orallo, Universitat Politècnica de València

Abstract

Artificial Intelligence (AI) offers the potential to transform our lives in radical ways. However, not only do we lack the tools to determine what achievements will be attained in the near future, but we even underestimate what various technologies in AI are capable of today. Certainly, the translation from scientific papers and benchmark performance to products is faster in AI than in other non-digital sectors. However, it is often the case that research breakthroughs do not directly translate to a technology that is ready to use in real-world environments. This document describes an example-based methodology to categorise and assess several AI technologies, by mapping them onto Technology Readiness Levels (TRL) (e.g., maturity and availability levels). We first interpret the nine TRLs in the context of AI and identify different categories in AI to which they can be assigned. We then introduce new bidimensional plots, called readiness-vs-generality charts, where we see that higher TRLs are achievable for low-generality technologies focusing on narrow or specific abilities, while low TRLs are still out of reach for more general capabilities. We include numerous examples of AI technologies in a variety of fields, and show their readiness-vs-generality charts, serving as a base for a broader discussion of AI technologies. Finally, we use the dynamics of several AI technology at different generality levels and moments of time to forecast some short-term and mid-term trends for AI.

Executive summary

We are still lack the capacity to predict what capabilities and products will become a reality even in the short term, a problem that is not particular for AI but any technology, and especially digital technologies. We are not always successful, even in hindsight, in understanding why some expectations are not met, and why some AI technologies have limitations or what kind of new technologies may replace them. Moreover, although many so-called breakthroughs in AI are associated with highly cited research papers or good performance in some particular benchmarks, research breakthroughs do not directly translate into a technology that is ready to use in real-world environments.

In this paper, we present a novel example-based methodology to categorise and assess several AI research and development technologies, by mapping them onto Technology Readiness Levels (TRL) (representing their maturity and availability). We first interpret the nine TRLs in the context of AI, and identify several categories in AI to which they can be assigned. The selection of technologies is representative but not exhaustive: it is based on our own experience and knowledge in the area about their relevance and "general use". Furthermore, for some specific cases, we have also considered the associated levels of research activity.

We then introduce new bidimensional plots, called readiness-vs-generality charts, in which we define the degree of generality (in terms of being able to function over many diverse specific domains and tasks) expected for a particular technology on the x-axis vs the readiness level (the TRLs) on the y-axis. Generality is a key element to be recognised, apart from the readiness levels since AI is a field that develops (cognitive) capabilities at different generality levels. Consequently, we need to assign readiness levels according to different levels of generality: a technology that is specialised for a very specific, controlled, domain may reach higher TRL than a technology that has to be more general-purpose in terms of it not-being restricted to specific tasks or scenarios. Therefore, for each technology we define the different levels of capabilities based on a comprehensive analysis of the related scientific and industrial literature. We also include examples of AI technologies in a variety of fields and show their readiness-vs-generality charts (see Table 1).

Table 1: AI categories and the sample of representative technologies evaluated for each of them.

Category	Technology
Knowledge Representation & Reasoning	Expert Systems
Learning	Recommender Systems Apprentices by Demonstration
Communication	Machine Translation Speech Recognition
Perception	Facial Recognition Text Recognition
Planning	Transport & Scheduling Systems
Physical Interaction (Robotics)	Self-Driving Cars Home Cleaning Robots
Social & Collaborative Intelligence	Negotiation Agents
Integrating Technology	Virtual Assistants

Methodologically, the examples analysed serve to illustrate the difficulties of estimating the TRLs, a problem that is not specific to AI. The use of levels on the x-axis, however, has helped us be more precise with the TRLs than would be otherwise. It should be noted that our initial assessment has undergone a profound evaluation by an independent panel of specialists, recognised in at least one of the technologies (or areas) addressed.

In the charts we see that higher TRLs are achievable for low-generality technologies focusing on narrow or specific abilities, while low TRLs are still out of reach for more general capabilities. Furthermore, the shapes of the curves seen in the charts of the previous section are informative about where the real challenges are for some technologies. Consequently, it seems that those curves that are flatter look more promising than those for which there is a steep step at some level on the x-axis. We use the dynamics of several AI technology

examples at different generality levels and moments of time to forecast some short-term and mid-term trends for AI. Finally, we illustrate that technological readiness does not mean technological success as well as the potential dangers of excessive focus on TRL when developing new AI technologies and the consequent criticisms related to the lack of generality of current AI technologies.

Valuable contributions of this work are: (1) the definition of the maturity levels for an illustrative set of AI technologies through the use of Technology Readiness Level (TRL) assessment. (2) The interpretation of the nine TRLs (introduced by NASA and adapted by the EU) in the context of AI, and then its systematic application to different categories in AI, by choosing one or two examples in each category. (3) The development of new bidimensional plots, known as readiness-vs-generality charts, as a trade-off between how general a technology is versus its readiness level. (4) The analysis of numerous examples of AI technologies in a variety of fields by means of the readiness-vs-generality charts. (5) The discussion about the future of AI as a transformative technology and how the readiness-vs-generality charts are useful for short-term and mid-term forecasting.

Foreword

This report is published in the context of AI WATCH, the European Commission knowledge service to monitor the development, uptake and impact of Artificial Intelligence (AI) for Europe, launched in December 2018.

AI has become an area of strategic importance with potential to be a key driver of economic development. AI also has a wide range of potential social implications. As part of its Digital Single Market Strategy, the European Commission put forward in April 2018 a European strategy on AI in its Communication "Artificial Intelligence for Europe" COM(2018)237. The aims of the European AI strategy announced in the communication are:

- To boost the EU's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors
- To prepare for socio-economic changes brought about by AI
- To ensure an appropriate ethical and legal framework.

Subsequently, in December 2018, the European Commission and the Member States published a "Coordinated Plan on Artificial Intelligence", COM(2018)795, on the development of AI in the EU. The Coordinated Plan mentions the role of AI Watch to monitor its implementation.

AI WATCH monitors European Union's industrial, technological and research capacity in AI; AI-related policy initiatives in the Member States; uptake and technical developments of AI; and AI impact. AI WATCH has a European focus within the global landscape. In the context of AI Watch, the Commission works in coordination with Member States. AI WATCH results and analyses are published on the AI WATCH Portal[1].

From AI Watch in-depth analyses, we will be able to better understand EU's areas of strength and areas where investment is needed. AI Watch will provide an independent assessment of the impacts and benefits of AI on growth, jobs, education, and society.

AI Watch is developed by the Joint Research Centre (JRC) of the European Commission in collaboration with the Directorate-General for Communications Networks, Content and Technology (DG CONNECT).

This report addresses the following objectives of AI WATCH: Analysis of the evolution of AI technologies. As part of this objective this report particularly aims to introduce an example-based methodology to categorise and assess several AI research and development technologies, by mapping them into Technology Readiness Levels (TRL).

1 https://ec.europa.eu/knowledge4policy/ai-watch_en

1 Introduction

Artificial Intelligence (AI) is poised to have a transformative effect on almost every aspect of our lives, from the viewpoint of individuals, groups, companies, and governments. While there are certainly many obstacles to overcome, AI has the potential to empower our daily lives in the immediate future. A great deal of this empowerment comes through the amplification of human abilities. Another important space AI systems are taking over comes from the opportunities of an increasingly more digitised and 'datafied'[2] world. Overall, AI is playing an important role in several sectors and applications, from virtual digital assistants in our smartphones to medical diagnosis systems. The impact on the labour market is already very visible, but the workplace may be totally transformed in the following years.

However, there is already a high degree of uncertainty even when it comes to determining whether a problem can be solved or an occupation can be replaced by AI *today* (Brynjolfsson et al. 2018, Martínez-Plumed et al. 2020). The readiness of AI seems to be limited to (1) areas that use and produce a sufficient amount of data and have clear objectives about what the business is trying to achieve; (2) scenarios where the suitable algorithms, approaches and software have been developed to make it fully functional into their relevant fields; and (3) situations whose costs of deployment are affordable (including data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, development time, etc., apart from monetary costs) (Martínez-Plumed et al. 2018a). To make things more complicated, AI is not one big, specific technology, but it rather consists of several different human-like and non-human-like capabilities, which currently have different levels of development (e.g., from research hypotheses and formulations to more deployed commercial applications). At a high level, AI is composed of reasoning, learning, perception, planning, communication, robotics and social intelligence. At a lower level, there are a myriad of applications that combine these abilities with many other components, not necessarily in AI, from driverless cars to chatbots.

Many products we have today were envisaged decades ago but have only come into place very recently. For instance, virtual digital assistants, such as Alexa, Siri and Google Home, are still far from some of the imagined possibilities, but they are already successfully answering a wide gamut of requests from customers, and have already become common shoulders to lean on in daily life. Similarly, computers that recognise us have been in our imagination and desiderata for decades, but it is only recently that AI-based face recognition and biometric systems populate smartphones, security cameras and other surveillance equipment for security and safety purposes. Machine learning and other AI techniques are now ubiquitous; recommender systems are used to enhance customers' experience in retailing and streaming services, fault detection and diagnosis systems are used in industry and healthcare, and planners and optimisers are used in logistics and transportation. Other applications, however, have been announced as imminent, but their deployment in the real world is taking longer than originally expected. For instance, self-driving cars are still taking off very timidly and in very particular contexts[3].

The key question is not if AI is envisaged or working in restricted situations, but whether an AI technology is sufficiently developed to be applicable in the real world, as a viable product leading to public and business value and real transformation. Only if we are able to answer this question can we really understand the impact of AI research breakthroughs and the time from different stages of their development to viable products. Policy-makers, researchers and customers need a clear technical analysis of AI capacities not only to determine what is in-scope and out-of-scope of AI (Martínez-Plumed 2018b), but also what are the current level of maturity and readiness of newly introduced technologies.

1.1 Objectives and contributions

The aim of this paper is thus to define the maturity of an illustrative set of AI technologies through the use of Technology Readiness Level (TRL) assessment. We first interpret the nine TRLs (introduced by NASA and adapted by the EU) in the context of AI, and then we apply them systematically to different categories in AI, by

2 https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#277a44c060ba

3 https://www.vox.com/future-perfect/2020/2/14/21063487/self-driving-cars-autonomous-vehicles-waymo-cruise-uber

choosing one or two examples in each category. In order to do this, we introduce new bidimensional plots, known as readiness-vs-generality charts, as a trade-off between how general a technology is versus its readiness level. We see that, in many domains, actual systems proven in operational environments are already out there, but still showing limited capabilities. For more generality in capabilities, the TRL is still at an earlier stage. We include numerous examples of AI technologies in a variety of fields and show their readiness-vs-generality charts. These are used as exemplars that work as practical guidelines for anyone interested in analysing other AI technologies using a similar methodology. The examples selected in this paper are also sufficiently representative for a discussion about the future of AI as a transformative technology and how these charts can be used for short-term and mid-term forecasting. We start this open discussion at the end of this paper.

1.2 Scope

We potentially consider all AI technologies, as defined by the areas that are usually associated with the discipline and that is one of the main reasons why we enumerate a list of AI categories that correspond to subfields in AI. In this regard we follow the AI Watch operational definition (Samoili et al., 2020) which defines a concise taxonomy that characterise the core domains of the AI research field (as well as transversal topics). This categorisation, which proceeds from the absence of a mutually agreed definition and taxonomy of AI, is used as a basis for the AI Watch monitoring activity and has been established by means of a flexible scientific methodology that allows regular revision. We do not use other characterisations of AI as comprising systems that act rationally or act like a human, which may be more restrictive. About the ingredients that make an AI technology inherently ready, we cover techniques and knowledge, but also 'compute', data and other dimensions of AI solutions. However, other factors affecting pace and adoption of a technology (e.g, financial costs of deploying solutions, labour market dynamics, economic benefits, regulatory delays, social acceptance, etc.) fall outside the scope of this report.

1.3 Intended Audience

This document is addressed to, on one side, researchers and companies writing project proposals and trying to determine which TRLs they will be able to achieve, and, on the other side, to policy-makers and evaluators assessing how far a given proposal reaches in the TRL scale. For target readers not familiar with TRLs, this document is self-contained and can also serve as an introduction to TRLs and a way of analysing progress in AI in terms of TRLs. This approach may represent a more fine-grained (in terms of AI area-specific and, more specifically, example-specific readiness analysis) and systematic scale (in terms of data collection, implementation and analysis) than using performance in benchmarks, bibliometric analysis or simply popularity.

The rest of the paper is organised as follows. Section 2 reviews the notion of technology readiness level, borrowed from NASA and adapted in the EU. Section 3 presents the key methodology: we first give the contours of what an AI technology is in particular, which is determined more precisely by those that can be assigned to one (or more) of the seven AI categories corresponding to subareas in the discipline. This section introduces the readiness-vs-generality charts, which are key for understanding the state of different technologies, by turning the conundrum between readiness and generality into a trade-off chart. Section 4 includes one or two examples of AI technologies for each of the seven categories, with a short definition, historical perspective and the grades of generality that are used in the charts. Section 5 discusses all charts together, finding different dynamics, and considers a prototypical example of AI technology, virtual assistant, covering several categories. Section 6 closes the paper with an analysis of future trends in AI according to the evolution of TRL for different levels of generality. An appendix follows after the references, including a rubric for the TRLs.

2 Technology Readiness Levels

Defined and used on-and-off for NASA space technology planning for many years, the Technology Readiness Levels (TRL) constitute a systematic measurement approach that supports consistent assessments, comparisons, and delimitations about the maturity of one or more technologies. TRL analysis was originally used for aeronautical and space projects and later generalised to any kind of project, covering the whole span from original idea to commercial deployment. The key point behind TRLs is that if we consider a specific technology and we have information about the TRL in which it is, we can get an idea of how mature it is. Therefore, the primary purpose of using TRLs is to help decision making concerning the development and transitioning of technology. TRL assessment should be viewed as one of several tools that are needed to manage the progress of research and development activity within an organisation.

The European Commission (EC) slightly adapted the TRL descriptions to be used in the Horizon 2020 Work Programmes and calls for proposals[4]. The current TRL scale used by the EC consists of 9 levels. Each level characterises the maturity of the development of a technology, from the mere idea (level 1) to its full deployment on the market (level 9)[5].

In what follows, we present these nine levels as we use them in this work (see the rubric for further details in the Appendix, and Table 1 for a summary):

- **TRL 1: Basic principles observed** (Have basic principles been observed and reported?) Lowest level of technology readiness. Research begins to be translated into applied research and development. Examples might include paper studies of a technology's basic properties.
- **TRL 2: Technology concept formulated** (Has a concept or application been formulated?) Invention begins. Once basic principles are observed, practical applications can be invented. Applications are speculative and there may be no proof or detailed analysis to support the assumptions. Examples are limited to analytic studies.
- **TRL 3: Experimental proof of concept** (Has analytical and experimental proof-of-concept been demonstrated?) Continued research and development efforts. This includes analytical studies and laboratory studies to physically validate analytical predictions of separate elements of the technology. Examples include components that are not yet integrated or representative.
- **TRL 4: Technology validated in the lab** (Has a component or layout been demonstrated in a laboratory (controlled) environment?) Basic technological components are integrated to establish that they will work together. This is relatively "low fidelity" compared to the eventual system. Examples include integration of "ad hoc" software or hardware in the laboratory.
- **TRL 5: Technology validated in a relevant environment**[6] (Has a component or layout unit been demonstrated in a relevant —typical; not necessarily stressing— environment?) Reliability is significantly increased. The basic technological components are integrated with reasonably realistic supporting elements so it can be tested in a simulated environment. Examples include "high fidelity" laboratory integration of components.
- **TRL 6: Technology demonstrated in a relevant environment** (Has a prototype been demonstrated in a relevant environment, on the target or surrogate platform?) Representative model or prototype system, which is well beyond that of TRL 5, is tested in a relevant environment. This represents a major step up in a technology's demonstrated readiness. Examples include testing a prototype in a high-fidelity laboratory environment or in a simulated operational environment.

4 https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016-2017/annexes/h2020-wp1617-annex-ga_en.pdf

5 Note that TRLs start from applied research, not covering the fundamental research that may lay the foundations of future technologies. The latter may be considered as a "TRL 0" (fundamental research), although this zero level is not contemplated in the original TRL scale, and we will not use it. The lowest level used in this paper will always be TRL 1.

6 When, in the descriptions, we talk about "relevant environment" we refer to an environment with conditions that are close enough to or simulate the conditions that exist in a real environment (production).

- **TRL 7: System prototype demonstration in operational environment** (Has the prototype unit been demonstrated in the operational environment?) Represents a major step up from TRL 6, requiring demonstration of an actual system prototype in an operational environment. Examples include testing the prototype in operational testing platforms (e.g., a real-world clinical setting, a vehicle, etc.) .
- **TRL 8: System complete and qualified** (Has a system or development unit been qualified but tools and platforms not operationally demonstrated?) Technology proved to work in its final form and under expected conditions. In most cases, this TRL represents the end of true system development. Examples include developmental test and evaluation of the system to determine if the requirements and specifications are fulfilled. By "qualified" we also understand that the system has been certified by regulators to be deployed in an operational environment (ready to be commercialised).
- **TRL 9: Actual system proven in operational environment** (Has a system or development unit been demonstrated on an operational environment?) Actual application of the technology in its final form and under mission conditions, such as those encountered in operational test and evaluation. Examples include using the system under operational conditions. This is not a necessary end point, as the technology can be improved over the months or years, especially as more and more users can give feedback. But it may also happen that general use unveils some flaws or safety issues, and the system must be retired, with one or more TRLs being reconsidered for the technology.

We may group the above nine TRLs in terms of the environment in which the project is developed. In the first four levels (TRL 1 - 4) the technology validation environment is in the laboratory, in levels TRL 5 and 6 the technology is being validated in an environment with characteristics similar to the real environment and the last three levels (TRL 7 - TRL 9) deal with the testing and validation of the technology in a real environment[7]. It can be seen graphically in Table 2 below (column "Environment").

Given the type of research, technological development and innovation being addressed, it should be noted that the first four levels would address the most basic technological research involving, mostly, laboratory results. Technological development would then be carried out from the levels TRL 5 - TRL 6 until the first prototype or demonstrator is obtained. Technological innovation projects would be between TRL 7 to TRL 9 since technological innovation requires the introduction of a new product or service on the market and for this it must have passed the tests and certifications as well as all relevant approvals. These levels would involve deployment or large-scale implementation. These concepts are shown in the column "Goal" of Table 1.

If we want to assess the life cycle of the technology to be developed in terms of outputs produced[8], TRL 1 to TRL 3 go from a first novel idea to the proof of concept. Subsequently, the technological development would be addressed (TRL 4 - TRL 7) until its validation. Finally, we would have its placing on the market and deployment (TRL 8 - TRL 9). This is shown in Table 2 below, column "Product/Evaluation".

Finally, one should also consider the results that each of the maturity levels would bring. Table 2 below shows this in the column "Outputs".

Last but not least, although TRLs have several advantages such as providing a unified and common framework for the understanding of the status of a technology, as well as helping to make decisions concerning technology funding and transition, there are some limitations. Readiness does not necessarily fit appropriateness or feasibility: a mature technology (e.g., an automated or self-driving train) may possess a greater or lesser degree of readiness to be used in a particular context (e.g., underground[9], airports[10], etc.), but the technology may not be ready to be applied to other contexts (e.g., general railways). We will deal with this issue later under the concept of generality.

Some disciplines have introduced variants or specific TRL scales, e.g., changing granularity (Charalambous et al. 2017), while others have given extra criteria for the particular discipline but keeping the original 9-level scale (Bucner et a. 2019). We will stick to the original scale here, and instead of giving a prescriptive refinement of

7 https://www.solarsteam.ca/TRL-file

8 https://www.cloudwatchhub.eu/exploitation/readiness-market-more-completing-software-development

9 https://press.siemens.com/global/en/pressrelease/europes-longest-driverless-subway-barcelona-goes-operation

10 http://www.mediacentre.gatwickairport.com/press-releases/2018/18_03_16_autonomous_vehicles.aspx

each level for AI, we will use the standard rubrics (see appendix) complemented with an exemple-based approach, as we explain in the following section.

Table 2: Summary of Technology Readiness Levels (TRLs) according to several characteristics.

Environment	Goal	Product / Evaluation	Outputs	TRL	Description
Laboratory	Research	Proof of concept	Scientific articles published on the principles of the new technology	**TRL 1**	Basic principles observed
			Publications or references highlighting the applications of the new technology.	**TRL 2**	Technology concept formulated
			Measurement of parameters in the laboratory	**TRL 3**	Experimental proof of concept
		Prototype	Results of tests carried out in the laboratory.	**TRL 4**	Technology validated in lab
Simulation	Development		Components validated in a relevant environment.	**TRL 5**	Technology validated in relevant environment
			Results of tests carried out at the prototype in a relevant environment.	**TRL 6**	Technology demonstrated in relevant environment
Operational	Implementation		Result of the prototype level tests carried out in the operating environment.	**TRL 7**	System prototype demonstration in operational environment
		Commercial product/service (certified)	Results of system tests in final configuration.	**TRL 8**	System complete and qualified
		Deployment	Final reports in working condition or actual mission.	**TRL 9**	Actual system proven in operational environment

3 Methodology

As the purpose of this paper is to determine a way to evaluate the TRLs of different AI technologies, it is key to be sufficiently general so that we could potentially consider and review any relevant and significant AI-related developments, covering both industry and academia. In this regard, we should first define what we mean by *an* AI technology, and whether this can capture new inventions and developments from all players related to innovation and production. Note that AI is not a single technology, but a research discipline in which different subareas have produced and will produce a number of different technologies. Of course, we could just enumerate a list of technologies belonging or involving AI, but it may well be imbalanced and non-representative of the full range of areas in AI. Therefore, in order to be able to cover a good representation of AI technologies that have spun off from academic or industrial research, we will identify subfields and recognise the relevant technologies they comprise.

It is also very important to recognise that apart from readiness levels, AI is a field that develops cognitive capabilities at different generality levels (e.g., voice recognition for different degrees of versatility and robustness can have different TRLs). Consequently, we need to assign readiness levels according to different levels of generality: a technology that is specialised for a very particular, controlled, domain may reach higher TRL than a technology that has to be more general-purpose (performing in a wide range of different scenarios and/or different tasks) or even open-ended (performing in uncontrolled scenarios). In order to represent the twin importance of these two concepts, in the last subsection we introduce the readiness-vs-generality charts, which will be applied over a subset of relevant AI technologies in the following sections.

3.1 What is an AI technology?

In any engineering or technological field, a particular technology is defined as the sum of techniques, skills, methods, and processes used in the resolution of concrete problems (Crabb 1823). Therefore, *technology* as such constitutes an umbrella term involving any sort of (scientific) knowledge that makes it possible to design and create goods or services that facilitate adaptation to the environment, as well as the satisfaction of individual essential needs and human aspirations. The simplest form of technology is the development and use of basic tools, either in the form of knowledge about techniques, processes, etc., or embedded into *technological* systems.

Artificial intelligence (or more precisely the technology that emerges from AI) is usually defined as a "replacing technology", or more generally as an "enabling technology" (Gadepally et al. 2019). Enabling technologies lead to important leaps in the capabilities of people or society overall. For instance, *writing* or the *computer* are such enabling technologies, as they replace or enhance human memory, information transmission or calculation. Definitely, AI introduces new capabilities, which can replace or augment human capabilities. It is important not to confound an AI system with the product of AI itself. For instance, if a generative model creates a painting, a poem or the plan of a house, the product the AI technology creates is not the painting, the poem or the plan of the house, but the generator, an AI system, which incarnates the autonomous ability. On the other hand, a tool such as a machine learning library is not an AI product, but a tool that allows *people* to create AI products; in this case, systems learning from data represent the autonomous ability.

The technologies that emerge from AI are also catalogued as "general-purpose" (Brynjolfsson et al. 2017) defined as those that can radically change society or the economy, such as electricity or automobiles. This definition, however, is not necessarily associated with how many different uses a technology has[11], so we prefer the alternative term "transformative technology". Consequently, we see AI technologies as transformative (Gruetzemacher & Whittlestone 2019). Clearly, a technology cannot be transformative if it does not reach critical elements of society or become mainstream. This is not possible if the technology does not reach TRL 9. As a result, many promising technologies in AI will only become transformative when they reach this TRL 9, and this is one reason why it is so important to assess how far we are from this final level to really determine the expected impact of AI on society.

All this is very well, but we still need a definition of AI technology. Although there are many different views on this, the overall research goal of AI is usually associated with the creation of technology that allows computers to function in an intelligent manner. However, assessing "intelligent behaviour" is still a matter of controversy and active research (Hernández-Orallo 2017). Therefore, we simply assume that an *AI technology is any sort of*

11 Actually, whether an AI technology is general-purpose or not will be considered by the term "generality" below. Some AI technologies are actually very specific.

scientific or industrial knowledge derived from the research and development in any subareas of the field. Of course, this depends on how well the contours of AI are delimited (Martínez-Plumed 2018b). Therefore, in this document, when we talk about an AI technology, we may indistinctly refer to a particular method used or introduced in an AI subdiscipline (e.g., autoencoder), a distinctive application area (e.g., machine translation), a specific product (e.g., optical character recognition system), a software tool or platform (e.g., decision support system), etc.

3.2 Categories of AI technologies

AI is not one big, specific technology. Rather, it consists of several main areas of research and development that have produced a variety of technologies. In other areas, the identification of technologies is performed through different methods, depending on the goal of the technology: craft or industrial production of goods, provision of services, organisation or performance of tasks, etc. However, the common phases in the invention and development of a new technology start with the identification of the practical problem to be solved. In the case of artificial intelligence, we can assimilate this first stage of the identification of technology with a given cognitive capability that we want to reproduce or create mechanically. These capabilities are usually grouped into areas of AI. Therefore, before starting to analyse the maturity levels of these different AI technologies, we will introduce those main fields of research in AI and what sort of relevant technologies they comprise. This categorisation is inspired by the operational definition of AI adopted in the context of AI Watch (Samoili et al., 2020), which proposes a concise taxonomy that characterises the core domains of AI research, as well as some transversal areas. In our case we focus on a list of seven categories, leaving out those more philosophical or ethical research areas related to AI. The categories selected are defined as follows:

- **Knowledge Representation and Reasoning**: This subarea of AI focuses on designing computer representations (e.g., data structures, semantic models, heuristics, etc.) with the fundamental objective to represent knowledge that facilitates inference (formal reasoning) to solve complex problems. Knowledge representation is being used, for instance, to embed the expertise and knowledge from humans in combination with a corpus of information to automate decision processes. Some specific examples are IBM Watson Health (Ahmed et al., 2017), DXplain (Hoffer et al., 2005) and CaDet (Fuchs et al., 1999).
- **Learning**: A fundamental concept of AI research since its inception is the study of computer algorithms that improve automatically through experience (Langley, 1996). While the term "learning" refers to more abstract, and generally complex, concepts in humans (such as episodic learning), today we tend to associate learning by computers with the prominent area of machine learning, in a more statistical or numeric fashion, such as implemented in neural networks or probabilistic methods (techniques that are now used in many of the other subdisciplines below). Machine learning involves a myriad of approaches, tools, techniques, and algorithms used to process, analyse and learn from data in order to create predictive models, identify descriptive patterns and ultimately extract insights (Flach 2012, Alpaydin 2020). These general algorithms can be adapted to specific problem domains, such as recommender systems (in retail or entertainment platforms), understanding human behaviour (e.g., predicting churn) or classify images or documents (e.g., filtering spam).
- **Communication**: Natural Language Processing (NLP) is the AI subfield concerned with the research of efficient mechanisms for communication between humans and machines through natural language (Clark et al. 2013, Goldberg 2017). It is mainly focused on reading comprehension and understanding of human language in oral conversations and written text. There is considerable commercial interest in the field: some applications of NLP include information retrieval, speech recognition, machine translation, question answering and language generation. Today, NLP, for instance, can be used in advertising and market intelligence to monitor social media, analyse customer reviews or process market-related news in real time to look for changes in customers' sentiment toward products and manufacturers.
- **Perception:** Machine perception is the capability of a computer system to interpret data from sensors to relate to and perceive the world around them. Sensors can be similar to the way humans perceive the world, leading to video, audio, touch, smell, movement, temperature or other kind of data humans can perceive, but machine perception can also include many other kinds of sophisticated sensors, from radars to chemical spectrograms, to massively distributed simple sensors coming from the Internet of Things (IoT). Computer *vision* (Szeliski 2010) has received most attention in the past decades and deals with computers gaining understanding from digital images or, more recently, videos. Many applications are

already in use today such as facial identification and recognition, scene reconstruction, event detection or video tracking. Computer *audition* (Gold et al. 2011) deals with the understanding of audio in terms of representation, transduction, grouping, use of musical knowledge and general sound semantics for the purpose of performing intelligent operations on audio and music signals by the computer. Applications include music genre recognition, music transcription, sound event detection, auditory scene analysis, music description and generation, emotion in audio, etc. Speech processing is covered by both perception and communication, as it requires NLP. Finally, tactile perception, dexterity, artificial olfaction, and other more physical perception problems are usually integrated into robotics (see below), but are needed in a wide range of haptic devices too and many other applications.

- **Planning and search**: This AI subject related to decision theory (Steele et. al., 2016) is concerned with the realisation of strategies or action sequences aiming at producing plans or optimising solutions for the execution by intelligent agents, autonomous robots, unmanned vehicles, control systems, etc. Note that a planning problem can be reduced to a search problem (Russell & Norvig, 2002). However, the actions to be planned or the solutions to be optimised are usually more complex than the outputs obtained in classification or regression problems, due to the multidimensional and structured space of solutions (e.g., a Markov Decision Process). In terms of applications, although planning has had real-world impact in applications from logistics (Kautz et al., 2000) to chemical synthesis (Segler et al. 2018) or health (Spyropoulos, 2000), planning algorithms have achieved remarkable popularity recently in games such as checkers, chess, Go and poker (Silver et al., 2016, 2017; Brown et al., 2019), usually in combination with reinforcement learning.
- **Physical interaction (robotics)**: This area deals with the development of autonomous mechanical devices that can perform tasks and interact with the physical world, possibly helping and assisting humans. Although robotics as such is an interdisciplinary branch of engineering and science (including remote-controlled robots with no autonomy or cognitive behaviour), AI typically focuses on robots (Murphy 2019) with a set of particular operations and capabilities: (1) autonomous locomotion and navigation, indoor or outdoor; (2) interaction, working effectively in homes or industrial environments, perceiving humans, planning their motion, communicating and being instructed to perform their physical procedures; and (3) control and autonomy, including the ability for a robot to take care of itself, exteroception, physical task performance, safety, etc. As examples of well-known applications of robots with AI we find driverless cars, robotic pets or robotic vacuum cleaners.
- **Social abilities (collective intelligence)**: The broad category covering social abilities and collective intelligence has to do with Multi-Agent Systems (MAS), Agent-Based Modelling (ABM), Swarm Intelligence as well as other related topics such as Game Theory (in auctions, networks, economics, fairness equilibria, etc.)., where collective behaviours emerge from the interaction, cooperation and coordination of decentralised self-organised agents (Shoham et al., 2008). In general terms, here we include those technologies that solve problems by distributing them to autonomous "agents" that interact with each other and reach conclusions or a (semi-)equilibrium through interaction and communication. This area overlaps with learning, reasoning, and planning. For instance, recommender engines are well-known applications where group intelligence emerges from collaboration (Chowdhury et al., 2010).

The above categorisation is sufficiently comprehensive of the areas of AI (and the capabilities that are being developed in the subject) to have a balanced first-level hierarchy where we can assign specific technologies to. Of course, there will be some technologies that may belong to two or more categories (we will include an example in the discussion), but we do not expect to have technologies that cannot be assigned to any category. Finally, note that AI technologies may be also categorised in the form of applications or programmes developed to perform specific tasks (weak AI). Actually, AI has been used to develop and advance numerous fields and industries and, therefore, we can find a wide range of examples of AI applications in areas such as healthcare (e.g., medical diagnosis), marketing (e.g., online assistants), automotive and transportation (e.g., self-parking and cruise controls), finance (e.g., electronic trading platforms), media (e.g., deep fakes), military (e.g., unmanned combat aerial vehicles), education (e.g., digital assistants/tutors), and more. These are all high-profile examples which, underneath, are using different precise AI techniques (belonging to the above list of seven categories) to successfully perform their tasks. In this sense, whatever the categorisation we use for AI technologies, any subsequent TRL analyses would draw similar conclusions as we are following an example-based approach for TRL evaluation choosing one or two examples in each of the considered categories.

3.3 TRL assessment in AI: readiness-vs-generality charts

In order to assess the readiness levels of AI technologies, we also face an important dilemma between the readiness level and the ability to act and successfully perform in real-world, open-ended (uncontrolled) scenarios. If we describe a generic technology (e.g., a robotic cleaner), we will have a very different assessment of readiness depending on whether the specification of the AI system requires more or less capabilities[12]. For instance, if the robotic cleaner is expected to clean objects, by removing them and placing them back, and also to cover vertical and horizontal surfaces, when people and pets are around, then the readiness level is expected to be lower than a vacuum cleaner roaming around on the floor, with a particularly engineered design that avoids some of the problems of a more open-ended situation. Of course, one can specify all these technologies separately, and identify different clusters of functionalities, as we see below in Figure 1 (left). These technologies are mostly independent and can reach different TRLs (shown in different darkness levels). Progress would be analysed by seeing for how many of them the TRLs increase. However, the overlaps are not systematic and high TRLs could be obtained by covering the whole space with very specific solutions.

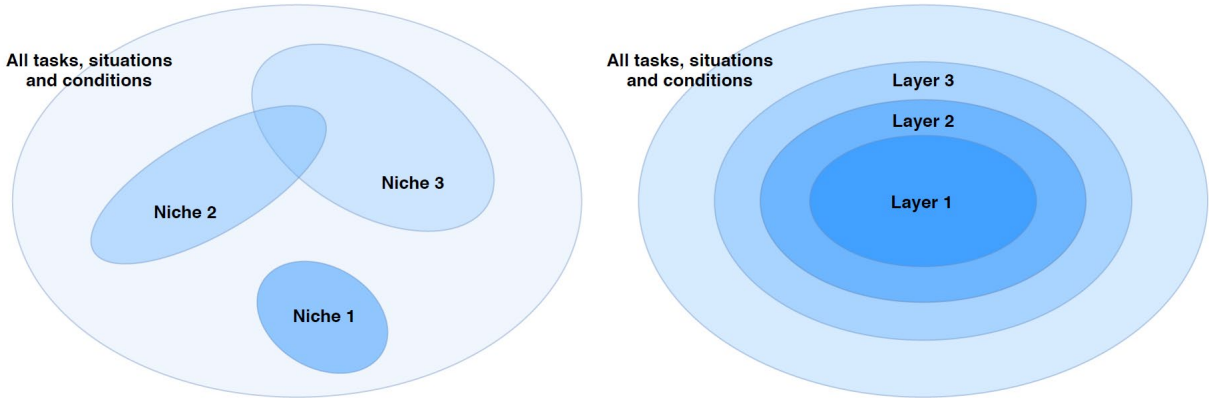

Figure 1. Left: we can consider different instances of the technology covering different niches, each of them solving a set of tasks, situations and conditions that are not hierarchically related to each other. Each cluster of functionality achieves a different TRL (shown with different darkness levels) that is mostly independent of the other niches. Right: we choose a decomposition of the space such that each instance of the technology that we analyse is a superset of the previous instances. We call these instances "levels of generality", as they are broader than the previous ones, containing them.

A different way of organising this space is a hierarchical generality model of technology, as illustrated in Figure 1 (right). In many areas, as we will see in the following sections, there is some meaningful way (many times more than one) to arrange the space of tasks, situations, and conditions in a hierarchical way. If we are able to select one hierarchy that is a total order (i.e., each pair of instances are comparable), then any instance is a subset of a more general instance and, thus, we will be able to talk about different *levels of generality* of the same technology. This ensures that no smaller task or situation is left out. Also, the idea of levels is a good representation of the fact that, very often, progress is cumulative.

Note that the higher generality is, the lower the expected readiness level becomes and vice versa. This will help understand the common situation where a technology is stuck at TRL 7, but reducing the scope of the technology, i.e., less general, or focusing on a specific functionality can lead to a product with TRL 9. Robotic vacuum cleaners are a good example of this. By limiting the scope of the technology, whether it be the task (only floor vacuuming) or the range (simple trajectories), the system is more specialised (or narrow), with the successful outcome that these devices are found in many homes today (TRL 9).

12 Note that we should not confuse capability (or functionality) with sophistication (or complexity): using a more sophisticated system does not guarantee further capabilities.

Another advantage of the hierarchical generality model of technology is that the total order allows the levels to be considered as an ordinal magnitude that can be represented in a Cartesian space along with another ordinal magnitude, the TRL. Thus, we can use two-dimensional plots[13] (readiness-vs-generality charts) with the degree of generality anticipated on the x-axis and the readiness level (the TRLs) on the y-axis. Figure 2 illustrates this idea with an example.

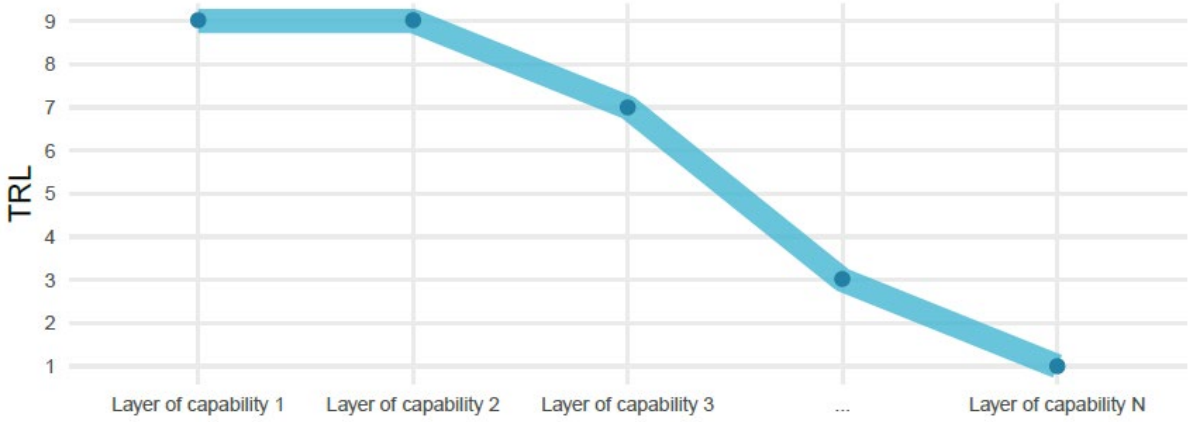

Figure 2. Readiness-vs-generality charts showing the different levels of capabilities (more specific to more general) on the x-axis and TRLs on the y-axis. Typically, the points will form a "curve" with a descending curve. Progress towards really transformative AI will be achieved by moving this curve to the top right**.**

As we move right on this plot, we have a system or application (i.e., an AI technology) that is more generally applicable. As we go up the plot, product readiness increases, in term of being used in the real world. Such a plot can be applied to any technology (e.g., a pencil is both general and ready, as a writing device), but determining the balance between generality and readiness is key in artificial intelligence, since many technologies sacrifice generality for performance in a particular niche of application to reach some narrow readiness. Only reaching the top right corner will generate really transformative technology[14]. For instance, a robotic vacuum cleaner moving around our floors has reached TRL 9 but has not transformed society. A fully-fledged robotic cleaner would do so, affecting millions of jobs and the way homes are organised for cleaning, recycling and even decoration.

The shape of these charts may reveal some important information. A steep decreasing curve that reaches high TRL levels for only low capabilities may show that important fundamental research is required to create — probably new— AI technologies that reach higher levels of generality. A flat curve that reaches only medium TRL for a wide array of capabilities may mean that reaching a commercial product or general use may depend on issues related to safety, usability or societal expectations about the technology, and not so much about rethinking the foundations of the technology. Nevertheless, a case-by-case analysis may lead to different interpretations. The next section presents the respective readiness-vs-generality chart for an illustrative set of AI technologies.

Before presenting the case-by-case analysis, we need to fix some criteria to determine the *x*-axis and the precise location of each point on the chart. Unfortunately, there is no standard scale for levels of generality that could be used for all technologies. For each technology, levels of generality are established by looking at the historical evolution of the technology, which means that some levels (e.g. word recognition for reduced vocabularies) did not get traction, while others (e.g., speech recognition for reduced vocabularies) can be

13 Both magnitudes (generality and TRL) are ordinal rather than quantitative, so technically a grid would be a more accurate representation than a Cartesian plot. Also, we connect the points with segments, but this does not mean that the intermediate points in these segments are really meaningful.

14 Here we refer to the concept of Radically Transformative AI (RTAI) from (Gruetzemacher et al., 2019) which is referred to as "AI capabilities or products which lead to societal change comparable to that precipitated by the agricultural or industrial revolutions". We may find examples of RTAI in the literature such as high-level machine intelligence (Grace et al., 2018), comprehensive AI services (Drexler, 2019) or a broadly capable system (Gruetzemacher, 2019).

identified as an early milestone in this technology. In all technologies, we can identify different dimensions that can help us define the levels. For instance, two dimensions are commonly involved in the definition of the levels of generality: how many situations the technology can cover (environments, kinds of problems), which can be associated with task generality, and the diversity of conditions for these situations (e.g., quality of the information, noise, etc.), which can be associated with robustness. The first dimension (i.e. situation covered) can unfold into two or more parameters (e.g. for speech recognition: size of vocabulary and number of languages). In our hierarchical generality model of technology, we merge all of them into one single ordinal level. There are of course cases where more challenging versions of the technology cannot easily be reduced to such a unidimensional scale, but we can still try to find a scale of levels that go from lower to higher generality. In a few cases, we will simply reuse some pre-established standard levels (usually defined at a development level rather than at a capability level) that have been used in the past for that particular technology, or even used as standards, as happens with machine translation (see the four basic types of translation (Hutchins et al. 1992)) and self-driving cars (see the US National Highway Traffic Safety Administration (NHTSA) definition of six levels of car autonomy[70]).

Once the space is defined by the generality levels and the nine readiness levels, we locate the points in the following way. First, we follow the rubric in the appendix. Second, for each level, we identify the highest TRL according to the best player (research team or company) as per 2020. The reasoning behind this choice —e.g., instead of choosing an average— is due to the fact that on AI technologies being digital, which means that they are quickly imitated by other players. Indeed, the possible slowing factors such as patents are usually compensated by open publication platforms such as arXiv[15] and open software platforms such as github[16], not to mention the common mobility of key people in AI between academia, industry and especially key tech giants, bringing the technology with them, and spreading it to other players.

Finally, even using this generality-vs-readiness space and the rubric in the appendix, there will be cases where we struggle to assess the TRL precisely. This can be caused by partial information about the technology, a definition of the TRLs that is not crisp enough, or the literature-based definitions for the levels of generality. It may also be the result of the authors of this report not being experts in each and every subarea in AI (although some detachment may also be positive). In other cases, this is caused because our assessments have been overseen by several experts (see the list in the acknowledgements at the beginning of the document) and occasionally there were some minor discrepancies. For all these cases we will use vertical error bars. We hope that some of our assessments could be replicated by other groups of experts and build these bars as proper confidence bands from the variance of results from a wider population of experts.

3.4 Methodology summary

The methodology developed in this report to define the maturity of AI technologies through the use of Technology Readiness Level (TRL) assessment covers the identification of AI technologies through to the assessment of their maturity levels:

1. **Identification of relevant AI technologies**. Based on the categorisation of AI technologies in section 3.2, we have assigned specific (illustrative) technologies to each AI area. The selection of technologies is based on our own experience and knowledge about their relevance and "general use". Furthermore, for some specific cases, we have also considered the associated levels of research activity (e.g., number of related papers, results, benchmarks, challenges, tasks, etc.) behind a particular technology. For the latter we have used the information provided in the *AIcollaboratory* (Martínez-Plumed et al. 2020a, 2020b, 2020c).

2. **Analysis of the TRLs for AI technologies.** We introduce and use two-dimensional plots called readiness-vs-generality charts in which we define the degree of generality of specific AI technologies on the x-axis vs the readiness level (the TRLs) on the y-axis. For each technology we define the different levels of capabilities based on a comprehensive analysis of the related scientific and industrial literature.

3. **Expert panel evaluation**. Our initial assessment undergoes a thorough assessment by an independent panel of specialists, recognised in at least one of the technologies (or areas) addressed. The experts are asked to follow the rubric in Appendix A to estimate the particular level in the scale for specific

15 https://arxiv.org/

16 https://github.com/

technologies. Furthermore, experts provide further information on the technology in question, such as signposting the most relevant research documents and publications which may help focus the analysis onto the most appropriate works, highlighting also any pertinent issues relating to the different technologies.

4. **Integration and evaluation**. Both our evaluations and the (qualitative) feedback and discrepancies provided by the panel of experts are then used to derive error bars in the readiness-vs-generality charts for each technology. The results are then summarised, and a briefing is produced subjected to further series of reviews and revisions. Note that a wider group of experts, using more extensive training on the TRLs and usual methods for aggregation or consensus of opinions (such as the Delphi method (Bernice 1968)) would bring more robustness to the TRL estimates, including a systematic way of deriving the error bars.

4 TRL Assessment for Representative AI Technologies

In this section, we select some illustrative AI technologies to explore how easy and insightful it is to determine the TRL for each of them. We will examine the technologies under the categories presented in section 3.2, and we use readiness-vs-generality charts for each of these technologies.

4.1 Knowledge representation and reasoning

Reasoning has always been associated with intelligence, especially when referring to humans. It is no wonder that the first efforts in AI were focused on building systems that were able to reason autonomously, going from some premises to conclusions, as in logic. We select one AI technology in this category, *expert systems*, because of its historical relevance and representativeness of *reasoning systems*.

4.1.1 Technology: Expert Systems

Expert systems, as introduced in the 1980s, is a traditional AI technology that humans can use to extend or complement their expertise. Expert systems are usually good at logical reasoning and receive inputs as facts that trigger a series of chain rules to reach some conclusions (typically as facts or statements). Expert systems are still fuelling many AI systems today, sometimes under the name "knowledge-based systems", such as some digital assistants or chatbots. In the early days of expert systems, the rules, i.e., the expertise encoded by the expert system, were usually created by experts manually, but nowadays knowledge can be extracted from document repositories or other sources such as the web or Wikipedia (Mitchell et al., 2018; Gonçalves et al., 2019). Modern systems can also revise their knowledge more easily than it was possible in the past. Such systems can deal with vast amounts of complex data in many application domains (Wagner 2017).

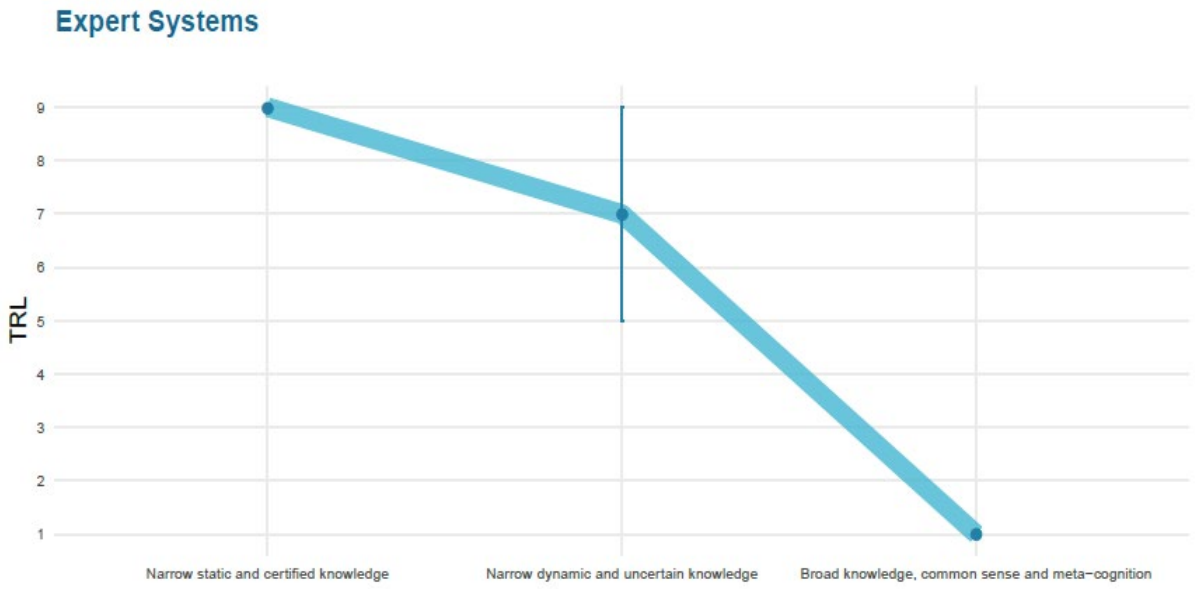

Figure 3. Readiness-vs-generality chart for expert system technology. While TRL 9 has clearly been reached for narrow systems with static and certified knowledge (early commercial systems and many expert systems still in place today), a very low TRL is estimated for expert systems dealing with general, broad knowledge and common sense. Current development is taking place at an intermediate level of expert systems, where knowledge is still narrow, but is changing, uncertain and updatable. Error bars are shown at this level because of doubts in the autonomy of some of these systems (e.g., IBM Watson).

Because of the evolution of expectations and capabilities of expert system technology, the *x*-axis of Figure 3 uses three different generality levels of expert systems:

- **Level 1 - narrow static and certified knowledge**: manually codifying narrow expertise knowledge, reason through bodies of specific knowledge, explain the reasoning, draw complex conclusions, etc.
- **Level 2 - narrow dynamic and uncertain knowledge**: automated knowledge refinement (belief revision, reason maintenance (Reinfrank 1988), etc.), reason under uncertainty, actionable insights, etc.
- **Level 3 - broad knowledge, common sense and meta-cognition**: introspective and broad knowledge, common sense, creative responses.

For the first level, early academic systems such as MYCIN (Shortliffe 2012) or CADUCEUS (Banks 1986) progressed from research papers to prototypes in relevant environments (TRL 7) in the 1970s and 1980s. Because of the excitement and expectations of expert systems in the 1980s, some commercial systems were used in business-world applications, reaching TRL 9. For instance, SID (Synthesis of Integral Design) was used for CPU design (Gibson et al, 1990). The success of former Expert Systems in TRL 9 also unveiled some limitations (e.g., narrow domains, manual knowledge acquisition, lack of common-sense knowledge, no revision, etc.). Today, many knowledge-based systems, usually coding business rules in database management systems as procedures or triggers, actually work as expert systems at this first level. Consequently, even if the term expert system is in disuse today, systems with these capabilities are still operating at TRL 9, as shown in Figure 3.

The second level represents a new level of expectations raised after the limitations of the 1980s. A new generation of expert systems was sought to overcome the knowledge acquisition bottleneck and be robust to change and uncertainty. They have been integrating automated rather than manual knowledge acquisition, and are deployed in a variety of industrial applications, such as health/diagnosis (Hoffer et al., 2005), control/management/monitoring (Jayaraman et al., 1996), stock markets (Dymova et al., 2012), space (Rasmussen 1990), etc. However, many of these systems do not meet the expectations of robustness and self-maintenance completely, and some of the features of level 2 are not fully autonomous (requiring important human maintenance). Because of this, we consider them more like market-ready research being tested and demonstrated in relevant environments, and thus covering different TRLs (from TRL 5 to TRL 9, ranging from prototypes to commercial products). This is reflected by the error bars in the figure. This can also be applied to a new generation of systems such as IBM Watson (Ahmed et al., 2017), which has already been validated and demonstrated in specific operational environments (e.g., health). Watson, in a limited sense, could be understood to be a powerful expert system, also combining a number of different techniques for natural language processing, information retrieval, hypothesis generation, etc.

At the third level of generality, we are talking about systems incorporating broad knowledge and common sense reasoning over that knowledge, including reasoning about what the system does not know (beyond assigning probabilities to their conclusions, as Watson does). While capturing and revising knowledge automatically for a wide range of domains has been illustrated in research papers and lab prototypes (Mitchell et al., 2018), nothing resembling true common sense reasoning has been shown even at a research level[17], and that is why we assign TRL 1 to this level (although it is more likely a fundamental research stage even before this level).

The schism between levels 2 and 3 (and the lack of progress on this schism in the past years) suggests there is still fundamental research to be done until AI systems exhibit more human-like common sense reasoning, being capable of predicting results and drawing conclusions that are similar to expert humans.

Of course, expert systems are not the only technology in the knowledge representation and reasoning category. Automated theorem provers, Boolean satisfiability problem (SAT) solvers, belief revision engines, truth maintenance systems, etc., as well as other different types of deductive and inference engines, are successful technologies that could also be analysed to determine their TRLs at different generality levels.

[17] Despite the recent success of NLP systems in Winograd Schema Challenge (context-based pronoun disambiguation problems) (Levesque et al, 2012), an alternative of the Turing Test (Turing, 1950), several criticisms question whether improved performance on these benchmarks represents genuine progress towards common-sense-enabled systems (see, e.g., Trichelair et al., 2019)

4.2 Learning

Learning is probably the most distinctive capability of systems that adapt to their environment. Systems that do not learn are brittle and cannot cope with any situation that was not accounted for beforehand. In the case of AI technologies, we want to consider systems that are not the result of the capability (e.g., a static classifier built with a machine learning technique that is no longer learning), but systems that continually improve with experience. We choose two technologies in this category: *recommender systems* that are constantly updating their recommendations as new data comes in, including new items, and more sophisticated *apprentices by demonstration*, which learn by observing how a (human) expert performs a task. Both are good examples of AI technologies that represent *learning systems*.

4.2.1 Technology: Recommender Systems

A recommender system (Ricci et al., 2011) is a type of information filtering system that aims to provide a way to quickly show users or users different types of topics or information items (e.g., movies, music, books, news, images, web pages, etc.) that they are looking for as well as to discover new ones that may be of their interest. A recommendation service should help cope with the flood of information by recommending a subset of objects to the user by predicting the "rating" or "weight" that the user would give to them. Recommender systems are based on the idea of similarities between items (i.e., an item is recommended based on interest of a similar item) or users (i.e., an item is recommended based on what a similar customer has consumed), or a combination between them both.

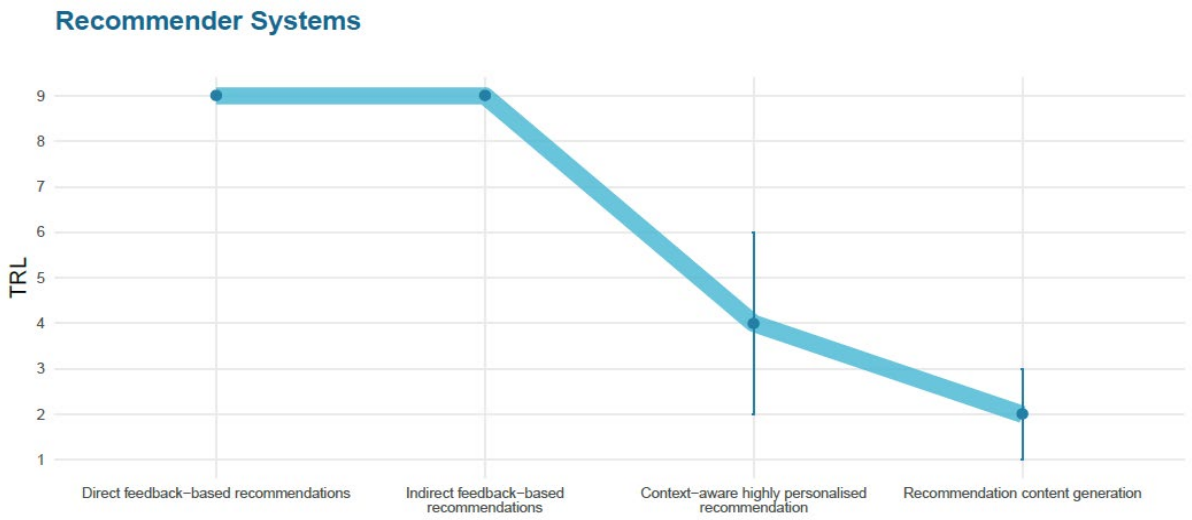

Figure 4: Readiness-vs-generality chart for recommender engines technology. TRL 9 reached for those very well-known recommender systems based on user feedback and used in a variety of areas such as playlist generators for video and music services or product recommenders. Current developments going beyond explicit feedback and using non-explicit latent attributes have already demonstrated their value in operational environments. Lower TRLs (TRL 2 to TRL 6) are estimated for more complete and flexible recommender systems being able to perform deeper personalisation using various dimensions of data. Finally, recommendation content generation would be a future direction in the field, with still little or no research nowadays.

Because of the evolution of expectations and capabilities of recommender systems technology, the *x*-axis of Figure 4 uses four different generality levels described in the following:

— **Level 1 - Direct feedback-based recommendations:** Personalised recommendation based on explicit rankings/feedback (click, buy, read, listen, watch...) over users/items and contexts.

- **Level 2 - Indirect feedback-based recommendations:** Recommendations beyond explicit feedback with latent attributes representing categories that are not obvious from the data.
- **Level 3 - Context-aware highly personalised recommendation:** User-based and context-aware personalised optimisation/recommendation balancing competing factors such as diversity, context, evidence, freshness and novelty, and using direct/indirect feedback, adding value-aware recommendations, etc.
- **Level 4 - Content generation recommendation:** Recommendations of what new items/content should be created to fill missing needs and add value.

For the first level of generality, we find those recommender systems able to find explicit similarity in users and items (making use of either or both collaborative filtering and content-based filtering (Ricci et al., 2011)) based on explicit feedback. Here we find a number of commercial systems that are or have been used in business-world applications, reaching TRL 9. For instance, we find the Pandora's Music Genome Project (Howe 2009) or the Stitch Fix's fashion box[18] as examples of content-based recommender systems. Also, the engines used in Amazon, Netflix (Gomez-Uribe, 2015), YouTube (Davidson, 2010), Spotify[19] or Linkedin (Wu et al., 2014) were (at the beginning of their development) examples of collaborative filtering-based approaches[20]. Finally, there are also popular recommender systems for specific topics like restaurants and online dating as well as to explore research articles and experts (Chen et al., 2011), collaborators (Chen et al., 2015), and financial services (Felfernig et al., 2017).

For the second level, more effective methods are currently being developed to look at similarity beyond explicit feedback as well as latent attributes (e.g., by using matrix factorization (Koren et al., 2009) or deep learning embeddings (Zhang et al., 2019)) revealing relationships that have not yet been realised. Research behind these more advanced and flexible approaches has increased exponentially in the past recent years[21] with notable examples such as those from Zillow[22], Netflix[23] and Airbnb (Grbovic, 2018) already demonstrated with success in operational and real-world environments (TRL 9).

Although successful, recommender systems still need to account for and balance multiple (competing) factors such as diversity, context, popularity/interest, evidence, freshness and novelty (Amatriain et al. 2016), to, for instance, make sure the recommendations are not biased against certain kinds of users and thus going beyond being simple proxies of accurate rating predictors. Furthermore, multi-dimensional rating would also be a step beyond (Shalom et al.,2016) for recommender systems being able to optimise and personalise the whole user experience (e.g. using a product, website, platform, etc.) via deep personalisation and using various dimensions of data. In this regard, recommendations and optimisations should be based on the understanding of a user's browsing or attention behavior. All this would correspond with the third level of generality in Figure 4, being still a matter of research and prototyping (TRL 2 to TRL 6) with some approaches and examples found in the literature (see e.g., Leonhardt et al. 2018, Ahmed et al. 2012, Kang et al. 2019).

Regarding the fourth level of generality, we are including further innovations for recommendation systems such as recommending new items/products/services/contents that do not exist and should be created to fill missing needs aiming at increasing, for instance, the value of the company or platform. Generating the content of a recommendation is still a research matter, including proof-of-concepts validated in lab (TRL 2 to TRL 4) with some ideas already published such as automatic food menu generation (Bianchini et al. 2017), music

18 https://algorithms-tour.stitchfix.com/

19 https://towardsdatascience.com/how-spotify-recommends-your-new-favorite-artist-8c1850512af0

20 Note that, currently, some of these companies use more advanced neural-based approaches (see e.g., Covington et al, 2016)

21 E.g., The leading international conference on recommendation systems (RecSys) started to organize regular workshops on deep learning in 2016.

22 https://www.zillow.com/tech/embedding-similar-home-recommendation/

23 https://help.netflix.com/en/node/100639

generation (Johansen 2018), simple fashion design[24] (Kang et al. 2017, Kumar and Gupta 2019) or even artificially generated comments (Lin et al. 2019).

As a final note, and in terms of current advances, some authors (Ekstrand et al. 2011, Konstan et al. 2013, Beel et al. 2016) have found that current research in recommender systems is stagnated because it is not providing meaningful contributions neither in terms of more advanced capabilities, nor regarding practical applications. The main reasons regarding the little impact of the research in the area are mainly the difficulty to reproduce recommender systems' research results, the lack of consistent and standard evaluations, the inexistence of comprehensive experiments and the necessity of establishing best-practice guidelines for recommender-systems research. Hence, practitioners and operators of recommender systems may find little guidance in the current research when looking for which recommendation approaches to use to address their specific tasks and problems.

4.2.2 Technology: Apprentice by Demonstration

Recommender systems are complex systems involving different types of information. However, in some way, they do not differ much from a classification problem powered by statistical correlations and patterns. In the case of humans, learning is usually associated with more complex phenomena, such as episodic learning, the creation of abstract concepts and the internalisation of new procedures. Many of these areas are still at a preliminary stage in AI (as they have always been!), but some others are beginning to show more progress in recent years. Learning by demonstration (Schaal 1997) is one of these types of learning that is more complex than a classical supervised or unsupervised machine learning problem, or even a generative model. Learning by demonstration, and the related learning by imitation (Miller et al., 1941), is the way in which culture is transmitted in apes, including humans. It is also very relevant in the workplace, as many tasks are just *taught* by an expert illustrating a procedure to an apprentice, sometimes with little verbalised instruction involved. More recently, with the popularity of short videos demonstrating simple tasks such as fixing a bike brake to cooking a fried egg, learning by demonstration is becoming the preferred way of instruction for many people. Consequently, progress in this area could have a significant impact on society.

Learning by demonstration is more technically defined in AI as learning a procedure or a task from traces, videos or examples of an operator (usually a human) performing the task. We limit our study here to tasks where the actions are discrete and relatively simple, to avoid overlapping with the robotics category. For instance, a videogame with a finite number of "action keys" is an example of this technology, or a spreadsheet automation that learns a simple programme snippet to perform an operation. A full operator in a factory is ruled out here because all the propriosensory complexity being involved. Consequently, we are referring to a technology that is usually known more specifically as *programming by demonstration* (Cypher 1993) or *programming by example* (Lieberman 2001). However, more recently, the combination of deep learning with reinforcement learning has developed new techniques, such as deep reinforcement learning, that are able to learn from the interaction with the environment. Soon, some of these techniques evolved to take advantage of traces (Sutton et al., 1998), or recorded interactions performed by a human or an artificial expert (Silver et al., 2016, Mnih et al., 2016, Harb et al., 2017).

24 https://towardsdatascience.com/the-future-of-visual-recommender-systems-four-practical-state-of-the-art-techniques-bae9f3e4c27f

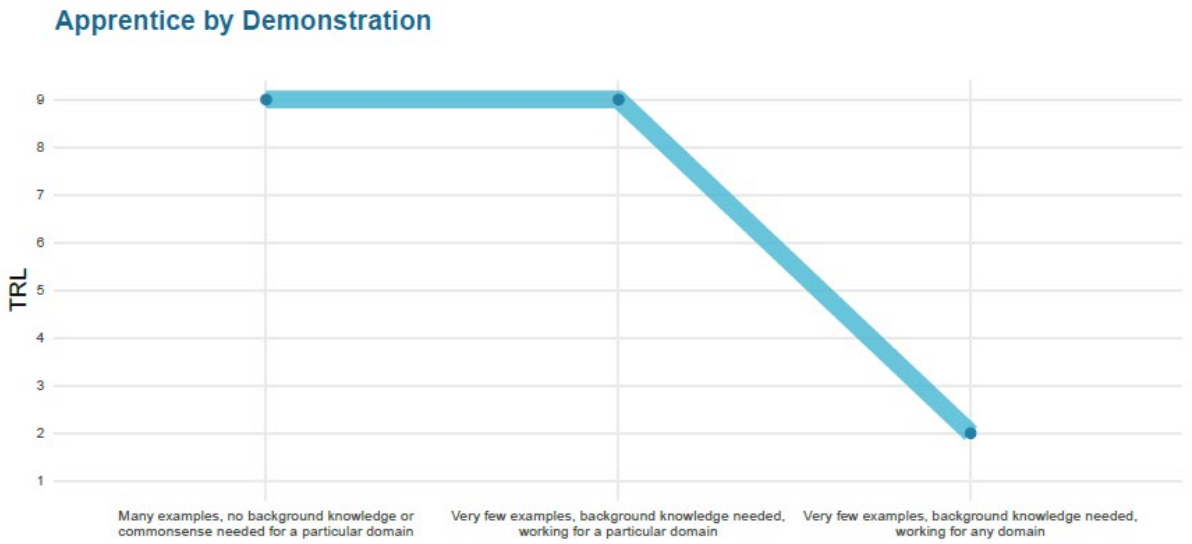

Figure 5: Readiness-vs-generality chart for learning by demonstration. We see that level 1 reaches TRL 9, especially because of the possibilities of deep reinforcement learning using human traces. level 2 also reaches TRL 9 in some domains, such as spreadsheet automation (although not in others, but we represent the maximum here, as we do in all other charts). Finally, level 3 requires learning systems that can process background knowledge in any domain, which is still at a very preliminary stage (TRL 2) with the principles and their envisaged applications.

The *x*-axis of Figure 5 uses three different generality levels, defined as follows:

- **Level 1: Many examples, no background knowledge or common-sense needed for a particular domain**: In this 'simple' case, a system can learn from a particular configuration of perceptions and actions (e.g., video games) and learn from thousands of traces of humans (or other systems) succeeding or failing at the task. Note that this is supposed to be more efficient than learning without traces, or necessary in some environments for which we lack a simulator, and a database of recorded cases is required (e.g., protocols, treatments, etc.).
- **Level 2: Very few examples, background knowledge needed, working for a particular domain**: When few examples are available, learning needs to rely on background knowledge. Here, we assume that only one domain (i.e., particular scenario or task) can be handled, by embedding sufficient background knowledge into the system or in the domain-specific language used for the representation of the policies and procedures.
- **Level 3: Very few examples, background knowledge needed, working for any domain**: In this case we want the system to handle virtually any domain. In order to reach this generality, we need the flexibility of changing the background knowledge from one domain to another, or a system that has wide knowledge about different areas, so that it can understand traces, videos, demos, etc., for different domains. For instance, the system should be able to automate a task, e.g., in a sales office or in a newspaper editorial office.

Given these levels, we can now assign the TRLs. For level 1 we can use as evidence the progress of deep reinforcement learning from traces. For instance, AlphaGo (Silver et al., 2016) was able to learn how to play go but used some hints from human traces. Similarly, many deep reinforcement learning algorithms use traces (Mnih et al., 2016, Harb et al., 2017). Because new variants of these algorithms are open source and already implemented[25], with more modest resources than in the original paper, this puts us in TRL 9, at least for the

25 See, e.g., https://github.com/openai/baselines

video game case. If we want to create an agent that can learn to play different Arcade games from observation, this can be done, and no background knowledge about the games is needed.

In level 2, the challenge comes from the limited number of examples. Humans usually need just one representative example to get the idea of a new task. This is possible because they have contextual information and background knowledge about the elements and basic actions that appear in the demonstrated task. This domain knowledge can be hardcoded into the system, either as rules or in the language itself used to express the learned procedures. We also assign a TRL 9 because of some successful systems in the domain of spreadsheet automation. In particular, Flash Fill (Gulwani et al., 2012) is based on a particular domain specific language that enables Microsoft Excel users to illustrate a simple formula with very few examples. The same idea has been brought to other domains, although each system requires a particular DSL for each domain (Polozov et al., 2015).

Finally, for level 3, we would like *the same system* to be able to learn tasks in different domains. This would mean that this apprentice could be applied for traces or videos in any domain and could replicate the task reliably. This level is still in its inception, even if there has been research for decades (Muggleton 1992; Olsson 1995; Ferri et al., 2001; Gulwani et al. 2015). While some systems have been applied to toy problems, we do not find evidence beyond having the technology concept formulated, and this is why we assign TRL 2.

Clearly, progress in this final level would have a major impact in many daily tasks that are repetitive and would not need programming or writing scripts or code snippets by hand. Such a system would have a transformative effect on the labour market and the work of programmers, among other professions. Because the challenge may depend on symbolic representations (for knowledge representation) and it has been explored for decades, we do not expect a breakthrough to high TRL 9 in the near-term future.

4.3 Communication

Computers exchange information all the time, but their format is predefined and formal. However, humans exchange information and knowledge in much more complex ways, especially through natural language. One big challenge of computers and AI has been developing tools that allow humans and machines to communicate more smoothly in natural language, and more generally about tools that can do some tasks related to language processing. We have chosen two AI technologies that are very significant in natural language processing: *machine translation* and *speech recognition*. These are two examples of AI technologies that represent *systems that (help) communicate*.

4.3.1 Technology: Machine Translation

Machine translation (MT) is the automatic translation of texts from one language into another language by a computer programme. While human translation is the subject of applied linguistics, machine translation is seen as a subarea of artificial intelligence and computer linguistics. At a basic level, although originally machine translation was based on simple substitutions of the atomic words of one natural language for those of another. Through the use of linguistic corpora, more complex translations can be attempted, allowing for more appropriate handling of differences in linguistic typology, sentence recognition, translation of idiomatic expressions and isolation of anomalies. This translation process can also be improved thanks to human intervention, for example, some systems allow the translator to choose proper names in advance, preventing them from being translated automatically. MT services have become increasingly popular in recent years, and there is an extensive range of MT software and special tools available, enabling fast processing of large volumes of text.

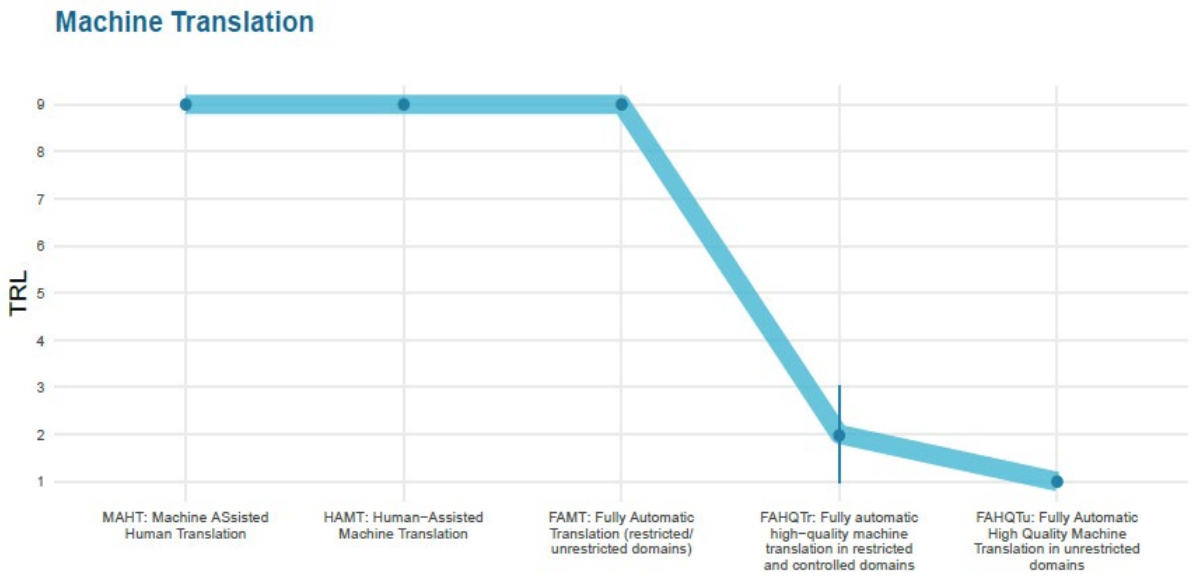

Figure 6: Readiness-vs-generality chart for Machine Translation (MT) technology. TRL 9 has been reached for the first two types of MT (MAHT and HAMT). Currently, FATM approaches have reached a crucial moment, with powerful market-ready products such as Google Translator or DeepL, and a lively research community developing and testing new systems at the expense of the improvements in neural-based approaches. The two FAHQT levels, either at controlled or uncontrolled scenarios, are estimated to have very low TRW due to the current limitations in the area of MT.

In terms of capabilities of MT, we define five levels of machine-assisted translation (see Figure 6) following the different types of translations already defined in the literature (Hutchins et al. 1992). While the level of autonomy is key in the first three of these types, and quality in levels 3 and 4, and these two factors are not necessarily aligned with levels of generality, we prefer to keep the original scale as the most interesting (challenging) levels are 4 to 5 and do correspond with varying generality:

- **Level 1 - Machine-assisted human translation (MAHT):** The translation is performed by a human translator who uses a computer as a tool to improve or speed up the translation process.
- **Level 2 - Human-assisted machine translation (HAMT):** The source and/or the target language text is modified by a human translator either before, during, or after it is translated by the computer.
- **Level 3 - Fully automatic (automated) machine translation (FAMT)**: This represents automatic production of a low-quality translation from the source language without any human intervention.
- **Level 4 - Fully automatic high-quality machine translation in restricted and controlled domains (FAHQMTr)**: This represents automatic production of a high-quality translation from the source language without any human intervention in restricted and controlled domains.
- **Level 5 - Fully automatic high-quality machine translation in unrestricted domains (FAHQMTu)**: This represents automatic production of a high-quality translation from the source language without any human intervention in unrestricted domains.

For the first two levels, it is clear we already reached TRL 9 levels, with a myriad of translation products[26] as well as dictionaries[27] and, thesaurus[28] in the market, helping to combine machine and human-based translations.

26 https://www.sdl.com/, https://www.memoq.com/ or https://www.wordfast.net/

27 https://www.wordreference.com/

In terms of current developments in FAMT (third level of capabilities), we have a number of successful MT software and applications, Google Translator being the flag bearer in **FAMT** (TRL 9). In some instances, MT services can replace human translators and dictionaries, and provide (imperfect but satisfactory) translations immediately. This is the case when getting the general meaning across is sufficient, such as with social media updates, manuals, presentations, forums, etc. In this regard, current MT software and applications[29] are best suited when we need quick, one-off translations and accuracy is not of importance. Also, MT applications are particularly effective domains where formal (structured) language is used. Finally, it should also be noted that although the technology has reached a TRL 9, MT is currently a hot area in AI in which a lot of advances are being achieved using new neural-based approaches (Sutskever et al., 2014), which have largely overcome the classical statistical approaches.

In this setting, the fourth and fifth levels correspond with the ultimate goal of MT: **FAHQMT**. As already commented, MT produces more usable outputs than when translating conversations or less standardised text. However, when aiming at professional translations of complex texts, business communication, etc., MT does not constitute, currently, a genuine or satisfactory alternative to qualified specialist translators[30]. A number of scholars questioned the feasibility of achieving fully automatic machine translation of high quality in the early decades of research in this area, first and most notably Yehoshua Bar-Hillel (Yehoshua, 1964). More recently, some research (TRL 1 to TRL 3) is being carried out for restricted scenarios (see, e.g., Muegge 2006), corresponding with the fourth level. Level 5 is still considered a utopia in MT (TRL 1) in the short or mid-terms. The most obvious scenario is the translation of literary texts: MT systems are unable to interpret text in context, understand the subtle nuances between synonyms, and fully handle metaphors, metonymy, humour, etc.

4.3.2 Technology: Speech Recognition

Speech recognition comprises those techniques and capabilities that enable a system to identify and process human speech. It involves sub-areas such as Speech Transcription (Seide et al. 2011) and Spoken Language Understanding (SLU) (Tur et al. 2011), among others, but we will focus on the former. Although speech recognition first came on the scene in the 1950s with a voice-driven machine named Audrey (by Bell Labs), which could understand the spoken numbers 0 to 9 with a 90 percent accuracy rate (Juang et al., 2005), nowadays, speech recognition programmes can recognise a virtually limitless number of spoken words, aided by cognitive and computational innovations (e.g., pure or hybrid neural models combining statistical approaches).

28 https://www.thesaurus.com/

29 https://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications

30 https://en.wikipedia.org/wiki/Machine_translation

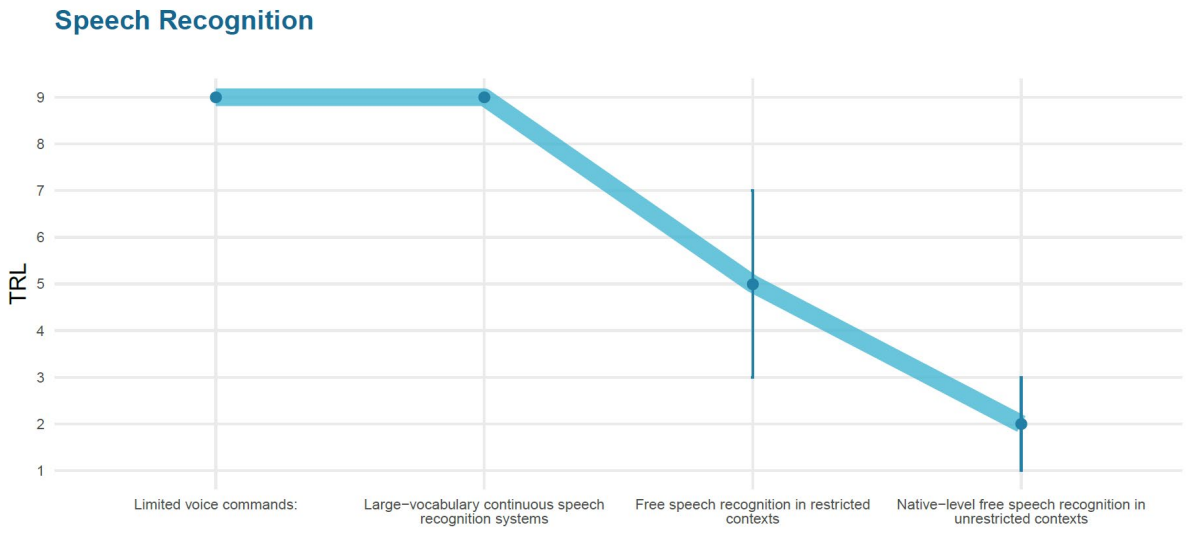

Figure 7: Readiness-vs-generality chart for speech recognition technology. TRL 9 has clearly been reached for narrow systems with limited voice commands or conversational interface such as those shown by the widespread VAs like Amazon's Alexa, Apple's Siri, etc. Current research is going towards more advanced speech capabilities including vocabulary size, speaker independence and attribution, processing speed, etc. Low TRLs are estimated for systems showing native-speaker language understanding capabilities.

Because of the evolution of expectations and capabilities of speech recognition technology, the *x*-axis of Figure 7 uses four different generality levels:

- **Level 1 - Limited voice commands**: Predefined instructions or voice commands in the recognition system with a particular (formal) syntax using, e.g., limited speech dictionaries.
- **Level 2 - Large-vocabulary continuous speech recognition systems**: Restricted-domain speech recognition systems with larger vocabularies for the spoken (formal and informal) words and phrases, some interaction with the user, high levels of robustness and accuracy of data, endpoint detection, no speech timeout, etc.
- **Level 3 - Free speech recognition in restricted contexts**: Open-ended vocabulary (formal and informal), far-field (remote) sources, speaker attribution, full transcription from any audio/video source, and able to deal with noise, echo, accents, disorganised speech, etc.
- **Level 4 - Native-level free speech recognition in unrestricted contexts**: Native-speaker multi-language recognition in adversarial environments, involving complete processing of complex language utterances, spontaneous speech, confusability, speaker independence, etc., under (possibly) adverse conditions.

For the first level, we find those voice recognition systems allowing predefined and limited system proprietary voice commands to perform specific instructions. We are able to find this technology in the market (TRL 9) since the 1980s in different products and applications, from voice-controlled operating systems (see e.g., the "Speakable Items" (Wallia 1994) in Mac OS in the 1990s) to toys (see, e.g., the Worlds of Wonder's Julie doll[31] in the 1980s) or in-car voice recognition systems (Tashev et al., 2009)

For the second level, common applications today include voice interfaces in robots, digital assistants or specific software such as voice dictation, voice dialling or call routing, domotic appliance control, preparation of structured documents, speech-to-text processing, and aircraft (e.g., direct voice input allowing the pilot to manage systems). Note that although all the above speech recognition-powered products and software are market-ready products (TRL 9) with high levels of robustness and accuracy, the capabilities achieved by these

31 http://www.robotsandcomputers.com/robots/manuals/Julie.pdf

systems are still limited to restricted domains, also having problems with noise environments, different accents, disorganised conversations, echoes, speaker distance from the microphone, etc.

Level 3 is still largely in research and evaluation phases; it is limited in that current approaches (e.g., language models and acoustic models) cannot handle the complexities of a free speech recognition application in unrestricted contexts with multiple speakers for a myriad of languages and different regional accents for the same languages. Furthermore, even in controlled contexts with a limited dictionary, there is still a lack of accuracy with common misinterpretations. Therefore, we can say that the technology achieving these capabilities is still a matter of research, prototyping and testing (TRL 3 to 7).

Finally, much more advanced capabilities in terms of a complete natural (multi-)language recognition in complex and unrestricted scenarios (as adult native speakers would do for their mother tongue) is still a long-term goal today for the research in the area (given the state being at TRL 1 to TRL 3). Working under adverse conditions (e.g., noise, different accents, complex language utterances, etc.) will be eventually solved in the short or medium term as they are problems that can be addressed with larger datasets and models. However, more complex scenarios such as language-independent speech recognition including the understanding of non-explicit information such as the use of prosody, emotions, meaningful pauses, intentional accents or even "mind reading" (e.g., speaker intention modelling) are clearly more long-term goals in the field.

4.4 Perception

Perception is a capability that we find in most animals, to a greater or lesser extent. In humans, vision is usually recognised as a predominant sense, and AI, especially in the recent years, has given this predominance to machine vision[32]. Even if we just cover vision below, we select two important technologies, *facial recognition* and *text recognition*, with very different perception targets, representing two good examples of AI technologies that incarnate *systems that perceive*.

4.4.1 Technology: Facial Recognition

A facial recognition (or identification) system is a technology capable of recognising or verifying a person from a digital image or a frame from a video source. In general, these systems work by comparing selected facial features from a given image (i.e., an "unknown" face) with faces within a database. An added difficulty is that this process may be needed in real time and, possibly, in adversarial scenarios. In recent years, facial recognition has gained a lot of attention, becoming an active research area and covering various disciplines such as image processing, pattern recognition, computer vision and neural networks. Facial recognition could also be considered within the field of object recognition, where the face is a three-dimensional object subject to variations in lighting, pose, etc., and has to be identified based on its 2D projection (except when 3D techniques are used).

Because of the evolution of expectations and capabilities of expert system technology, the *x*-axis of Figure 8 uses three different generality levels of expert systems:

Level 1 - Recognition under ideal situations: gender, age or identity recognition from high-quality stand still full-frontal faces in controlled scenarios (illumination, camera and person are controlled).

Level 2 - Recognition under partially controlled situations: gender, age or identity recognition from low-quality frontal faces (~20 degrees off) possibly in less controlled scenarios: illumination and camera are controlled, but not the person (e.g., railway stations or airports check-ins).

Level 3 - Recognition under uncontrolled situations: gender, age or identity recognition at pose variations, in low-resolution and poorly illuminated from (partial) facial photos/video in uncontrolled scenarios: neither the camera, nor the illumination, nor the person is controlled, such as in open spaces where pictures may be taken using smartphones, dashcams, etc. The recogition must be robust to person characteristics such as race, sex, etc., as well as changes in hairstyle, facial hair, body weight, and the effects of aging**.**

[32] This predominance is perhaps exaggerated, at least with a view of AI as achieving intelligent behaviour. Blind people from birth are the proof that full cognitive development is possible without sight.

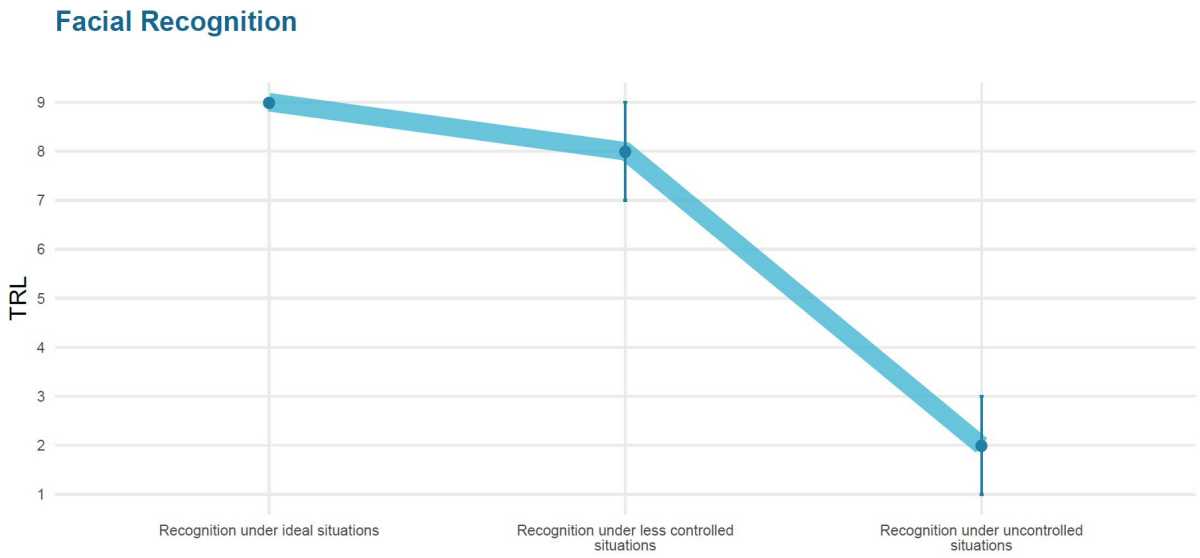

Figure 8: Readiness-vs-generality chart for facial recognition technology. TRL 9 has been clearly reached by facial recognition systems in controlled, ideal environments, with a number of systems being used for different applications (control, security, advertising, social media, etc.). Facial recognition systems under less controlled situations (such as in crowded train/metro stations or airports), and regardless of the expression, facial hair or age of the people, are also currently being tested and demonstrated in operational environments (TRL between 7 and 9). Lower TRLs are estimated when this sort of systems perform in totally uncontrolled scenarios having to deal with, for instance, pose variations, low quality/resolution, bad lighting, etc., and with people of various race, sex and other personal characteristics (e.g., facial hair, body weight, accessories, etc.).

Regarding the first level, most current facial recognition systems excel in matching one image of an isolated face with another in very controlled situations, such as when checking a driver's license or a passport. In this regard, nowadays we find lots of market-ready facial recognition applications (TRL 9) related to security, law enforcement or surveillance (helping police officers identify individuals[33], find missing people[34], etc.); retail and advertising (e.g., enabling more targeted advertising by predicting people's age and gender[35]), social media (e.g., to automatically recognise when its members appear in photos), financial services (e.g., digital payments, online account access[36], etc.), boarding controls in airports or train stations[37], among others. At present development levels, these systems are also able to detect people's gender (see, e.g., Mansanet et al., 2016), age[38] and even emotions (see, e.g., Ko 2018) with accuracy levels of over 99%[39] [40]. However, these systems still rely on full frontal face images with little or no change in illumination and orientation angle to achieve those high levels of predictive accuracy.

As for the second level, facial recognition outside of a controlled environment is no simple task. It is true that the technology is being evolved and designed to compare and predict potential matches of faces regardless of their expression (see Samadiani et al., 2019 for a review), facial hair (see, e.g., Xie et al., 2018), and age (see

33 https://www.interpol.int/How-we-work/Forensics/Facial-Recognition

34 https://www.independent.co.uk/life-style/gadgets-and-tech/news/india-police-missing-children-facial-recognition-tech-trace-find-reunite-a8320406.html

35 https://www.theguardian.com/business/2013/nov/03/privacy-tesco-scan-customers-faces

36 https://findface.pro/en/solution/finance/

37 https://www.airportveriscan.com/

38 https://labs.everypixel.com/api/demo

39 https://paperswithcode.com/task/face-recognition

40 https://neurosciencenews.com/man-machine-facial-recognition-120191/

e.g., Park et al., 2010). Also, there are currently a number of initiatives testing and demonstrating their capabilities in different operational and real-world scenarios (TRL 7 - 8) such as railway stations, airports, stadiums, etc., with different goals (e.g., security, control, etc.) (Galbally et al. 2019). However, at this level of generality, the technology is having two major drawbacks: (1) performance: facial recognition is still much more effective in "constrained situations" than for more general and uncontrolled scenarios where illumination, pose/angle/position and expression are the three major uncontrolled parameters that makes facial recognition a hard task; and (2) restrictions: current plans to install facial recognition systems in crowded public places for, e.g., surveillance reasons, are suffering from criticisms from civil society organisations as well as bans from the authorities[41] (approval is needed for TRL 8), although there are some surveillance and security systems currently operating (TRL 9) in less privacy-concerned countries such as China (see, for instance, the *YITU Dragon Eye* products used in Shanghai Metro[42]).

A further step in generality, corresponding with the third level in Figure 8, involves addressing more complex factors (in uncontrolled scenarios) such as inadequate illumination, partial or low quality image or video (e.g., only one eye is visible), multiple camera angles, poses or image variations (e.g., the subject is not looking straight into the camera), obstructions (e.g., people wearing hats, scarfs, sunglasses), etc. Furthermore, a drop in performance is obtained for facial recognition systems when trying to recognise people of different race or sex (Grother et al., 2019), this being a challenge for these systems. In terms of development, there are currently some research initiatives producing new methods for partial and unconstrained face recognition, although it is still work in progress (TRL 1 to TRL 3) and recognition accuracy can be as low as 30% - 50% in some cases (Elmahmudi, at al., 2019).

4.4.2 Technology: Text Recognition

Text Recognition is the process of digitising text by automatically identifying symbols or characters from an image belonging to a certain alphabet, making them accessible in a computer-friendly form for text processing programmes or similar. Text recognition involves both offline recognition (e.g., input scanned from images, documents, etc.) and online recognition (i.e., input is provided in real time from devices such as tablets, smartphones, digitisers, etc.). Here we will focus on the former. Large amounts of written, typographical, or handwritten information exist and is continuously generated in all types of media. In this context, being able to automate the conversion (or reconversion) into a symbolic format implies a significant saving in human resources and an increase in productivity, while maintaining or even improving the quality of many services. Optical character recognition (OCR) has been in regular use since the 1990s, developed significantly with the widespread use of the fax by the end of the 20th century. Today, they are already in wide use, but the possibilities and requirements have evolved with a more digital society.

Figure 9 tries to model the evolution of expectations in terms of the different capabilities of text recognition technology through the following levels of generality:

- **Level 1 - Template-based typewritten and handwritten character recognition:** recognition of typewritten and handwritten character in structured documents (e.g., postal systems, bank-check processing, passports, invoices, etc.).
- **Level 2 - Free-form handwritten character recognition**: recognition of (non-)separable/segmentable handwritten characters with automatic layout analysis in unstructured documents.
- **Level 3 - Free-form unconstrained handwritten word recognition:** recognition of unconstrained (non-)separable/segmentable handwritten words in unstructured documents.
- **Level 4 - Complex non-pundit-readable text recognition**: recognition, interpretation and deciphering of non-pundit-readable media (e.g., ancient or badly damaged) unconstrained texts in any format.

41 Some examples include: https://www.euractiv.com/section/data-protection/news/german-ministers-plan-to-expand-automatic-facial-recognition-meets-fierce-criticism/, https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html or https://sciencebusiness.net/news/eu-makes-move-ban-use-facial-recognition-systems

42 https://www.yitutech.com/en

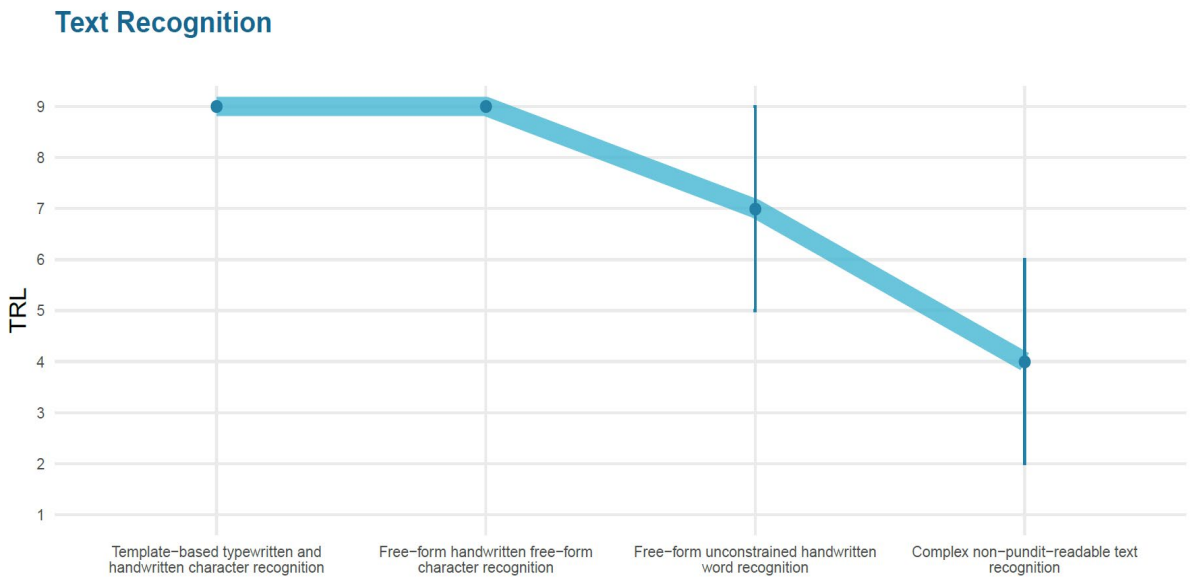

Figure 9. Readiness-vs-generality chart for text recognition technology. TRL 9 has been clearly reached by OCR systems. For free-form character recognition, current developments in machine learning and computer vision are improving the performance of these systems, where we may find prototypes for testing and demonstrating new capabilities as well as market-ready products (TRL 5 to TRL 9). More advanced capabilities in terms of unconstrained, free-form recognition of handwritten text is still a matter of research and development (TRL 2 to TRL 6). Very low TRLs are estimated for text recognition systems addressing the interpretation and deciphering of non-human-readable media.

For the first level we find the simplest (and common) form of character recognition: template-based optical character recognition (OCR). OCR as a technology has been instrumental in automating the processing of managing physical typewritten documents. For instance, enterprises using OCR software can create digital copies of structured documents such as invoices, receipts, bank statements and any type of accounting documents that needs to be managed. Passports, , and other forms of structured documentation that need to be managed are also the target of OCR software. The accuracy of these systems is dependent on the quality of the original document, but levels are usually around 98% or 99% for printed text (Holley 2009), which is good enough for most applications, or 95% when addressing, for instance, very specific handwritten recognition task such as postal address interpretation (see, e.g., (Srihari et al 1997)). Most commercial products and software are of this type (TRL 9)[43].

Currently, OCR technology has been improved by using a combination of machine learning and computer vision algorithms to analyse document layout during pre-processing to pinpoint what information has to be extracted. This technology is usually called “Intelligent Character Recognition” (ICR) and targets both unconstrained typewritten and handwritten text, imposing new challenges to the technology. This represents thus the second level of capabilities in Figure 9. Because this process is involved in recognising handwriting text, accuracy levels may, in some circumstances, not be very good but can achieve 97-99% accuracy rates in structured forms when handling capital letters and numbers (Ptucha et al., 2019) which are easily separable/segmentable, but it fails when addressing more complex scenarios such as unconstrained texts or non-separable (e.g., cursive) handwriting. However, these error rates do not preclude these systems from massive use, with plenty of ICR products and software currently in the market[44] (TRL 9). It is also an active area of research (see, e.g., Bai et al., 2014; Oyedotun et al., 2015; Yang et al.,2016; Ptucha et al., 2019;) where new alternatives (e.g., neural approaches) are being developed and assessed.

The third level of capabilities represents further advancements in this sort of technology involving recognition of unconstrained (i.e., non-easily separable/segmentable) and free-form handwritten word (instead of

43 https://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software

44 See http://www.cvisiontech.com/library/ocr/text-ocr/intelligent-character-recognition-software.html, https://abbyy.technology/en:features:ocr:icr or https://www.scanstore.com/Forms_Processing_Software/ICR_Software/

"character") recognition[45]. "Intelligent word recognition" (IWR) technologies[46] may fall into this level. IWR is optimised for processing real-world documents that contain mostly free-form, hard-to-recognise data fields that are inherently unsuitable for ICR. While ICR recognizes on the character-level, IWR works with unstructured information (e.g., full words or phrases) from documents. Although IWR is said to be more evolved than hand print ICR, it is still an emerging technology (TRL 5 to TRL 9) with some products performing capabilities to decode (scanned) printed or handwritten text (see, e.g., Google Vision API[47] used in Google Docs[48] and Google Lens app[49]), as well as number of prototypes being tested and validated in relevant environments (Yuan et al., 2012; Acharyya et al., 2013).

Finally, much more advanced uses of text recognition systems would be, for instance, to interpret ancient or badly damaged texts that can only be deciphered by pundits or even not deciphered by humans. In this line there we nowadays find some efforts in terms of research and projects (see, e.g., Lavrenko et al. 2004, Sánchez at al. 2013, Granell et al. 2018, Toselli et al. 2019), but without going beyond successful validations and demonstrations from laboratory to relevant scenarios (TRL-2 to TRL-6).

4.5 Planning

In this AI category, planning usually deals with choosing the best sequence of actions according to some utility function, and scheduling is about arranging a set of actions (or a plan) in a timeline subject to some constraints. Not surprisingly, this is one of the areas in AI that had early successful applications in different domains. We choose *transport scheduling systems* as a well-delineated example of an AI technology that represents systems that plan.

4.5.1 Technology: Transport Scheduling Systems

Transport scheduling refers to those tactical decisions associated with the creation of vehicle service schedules (also called "timetabling") aiming at minimising the net operating costs (Boyle 2009). In order to determine an appropriate vehicle schedule, there are also other factors having a direct effect on the operating costs: the number of vehicles required, the total mileage and hours for the vehicle fleet as well as the crew schedule. These activities are usually assisted by software systems with or without direct interaction with the planner in charge. This sort of systems take as input several parameters, including the frequency of service in different routes, the expected travel times, etc., as well as different operating conditions and constraints (e.g., "clockface" values, vehicle reutilisation/repositions, layovers, coordination of passenger transfers, number of vehicles, etc.), to generate high-quality solutions (e.g., departure times).

Because of the evolution of the expectations and capabilities of transport scheduling technology, the *x*-axis of Figure 10 uses three different generality levels described as follows:

- **Level 1 - Specific-purpose offline scheduling:** all the information is available beforehand with no uncertainty, which can be used as an input and an optimised schedule is output. The particularities of the domain are embedded into the system and only the data is given as an input.
- **Level 2 - Specific-purpose online scheduling/rescheduling:** all or part of the input information comes in real time, with uncertainty in measurements or in the information (e.g., a train that should arrive at 3:30 but arrives at 3:40). Still, the particularities of the domain are embedded into the system.

45 Note that the transcription at further levels (e.g., line or paragraph) goes beyond this technology as it involves other technologies such as (joint) line segmentation (Bluche 2016)

46 https://www.efilecabinet.com/what-is-iwr-intelligent-word-recognition-how-is-it-related-to-document-management/, https://content.infrrd.ai/case-studies/global-investment-firm-uses-infrrds-intelligent-data-processing

47 https://cloud.google.com/vision/docs/handwriting

48 https://docs.google.com/

49 https://lens.google.com/

— **Level 3 - General-purpose online scheduling/rescheduling:** the information also comes in real time and with uncertainty, but the system is now designed to be extended with new subsystems that have different specific behaviours. For instance, a train station scheduling system can include the behaviour, utilities and constraints of bus and metro subsystems connecting with the station, as well as events in the city, and optimise globally.

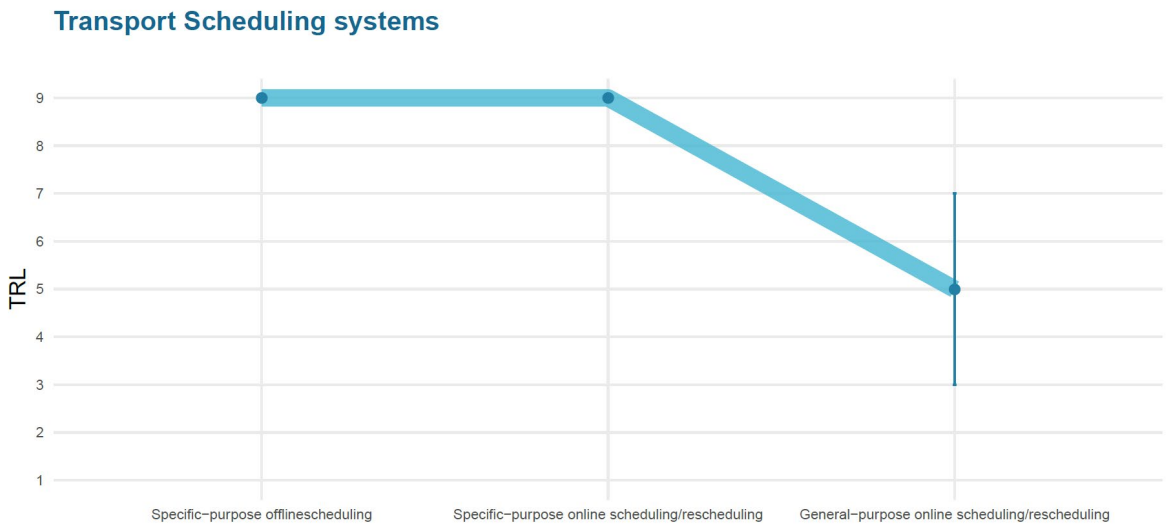

Figure 10: Readiness-vs-generality chart for transport scheduling system technology. The range of software systems that are able to perform offline and online scheduling for particular domains implies a TRL 9 for the first two levels. More general-purpose scheduling systems have a lower TRL, between 3 and 7.

Although, traditionally, transport timetables have been manually generated (e.g., using time-distance diagrams (Chakroborty et al., 2017) where schedules are manually adjusted to meet all the constraints), this process can take a long time and it is unfeasible when dealing with high-loaded transport networks. At the first level of generality, computer-based scheduling and planner systems have appeared over the last decades to provide automated and optimised transport scheduling for vehicles and drivers. These systems have been launched, after years of research, for different areas of application (TRL 9) including, among others, (a) trains (Ghoseiri et al., 2004, Ingolotti et al., 2004, Abril et al. 2006) with a huge number of commercial products such as RAILSYS[50], OTT[51] or MULTIRAIL[52]; (b) flights (Feo et al, 2009), also with a myriad of commercial applications such as FLIGHTMANAGER[53], OASIS[54] or TAKEFLIGHT[55]; (c) buses and shuttles (Gavish et al., 1978), with software platforms as GOALBUS[56], TRIPSPARK[57] or REVEAL[58]; (d) maritime transport (Meng et al., 2014) with commercial software such as MJC2[59], or MES[60]: or (e) road transport (Törnquist 2006), with software products such as

50 https://www.rmcon-int.de/railsys-en/

51 https://www.via-con.de/en/development/opentimetable/

52 https://www.oliverwyman.com/our-expertise/insights/2013/jan/multirail-pax-_-integrated-passenger-rail-planning-.html

53 https://www.topsystem.de/en/flight-scheduling-1033.html

54 http://www.osched.com/

55 https://tflite.com/airline-software/Passenger-Service-System/flight-schedule/

56 https://www.goalsystems.com/en/goalbus/

57 https://www.tripspark.com/fixed-route-software/scheduling-and-routing

58 http://reveal-solutions.net/bus-routing-scheduling-software/bus-scheduling-software-101/

59 https://www.mjc2.com/transport-logistics-management.htm

PARAGON[61] or PARADOX[62]. Note that all these systems are specialised (or adapted) for performing in very particular scenarios, and there is no general-purpose tool.

For the second degree, we consider that the input information can be provided online so an automated scheduling system needs to process it in real time. The systems should have then two parts: off-line scheduling (for known information) and on-line re-scheduling. While the former oversees scheduling vehicles and crews from known information, the latter has to be applied in response to the new specific needs and/or incidents that may appear. The schedules have to be dynamically updated balancing the resources (vehicles, timeslots, crew, etc.) available. Examples of real-time requirements or incidents may be, for instance, to meet specific travel demands or requests of passengers (e.g., new stops), to adapt to perturbations or problems regarding resources/demand (e.g., failures in vehicles), or manage new schedule intervals between new events (e.g., as volcano eruptions or heavy weather-related issues), etc. Dealing with real-time needs also entails that scheduling systems have to be able to confront different levels of uncertainty in terms of measurements or in the information they are provided (e.g., a train will arrive at 3:30 but it arrives at 3:40). Like in the first level, we are able to find plenty of research in this regard (see, e.g., Eberlein et al., 1998; Fu et al., 2002; D'Ariano et al., 2008, Verderame et al., 2010; Wegele et al., 2010; Reiners et al., 2012) as well as market-ready applications (e.g., MJC2[63] for road traffic, TPS[64] for trains, OPTIBUS[65] for bus/shuttles) applied to different transport scenarios, this implying a TRL 9 for this sort of more capable scheduling systems.

Finally, for the third level, we introduce a further level of generality in terms of these systems being able to be extended to any sort of transport scheduling problem with a combination of other transportation systems and other constraints and utility functions (e.g., a coach service combined with a train service). However, having general-purpose scheduling software systems is more difficult due to the varietal intrinsic characteristics of each scenario (it is not the same scheduling a fleet of trucks based on road-traffic characteristics as scheduling flights based on airflows, hub bankings and other flight characteristics). However, although the previously introduced products and software platforms are all domain-specific systems, the task of automating scheduling or timetabling (as a multi-objective constrained optimisation problem) is a general problem creating a feasible/optimised schedule for any kind of service or a combination of them. In (Hassold et al., 2014; Liu et al., 2016) we can see some general-purpose solutions (at the research level), but they are still being tested and demonstrated in particular domains. That is why we give a TRL value between 3 and 7.

4.6 Physical interaction (robotics)

Many people have a paradigmatic view of intelligent systems as robots that physically interact with the world. While a great part of AI applications are digital, it is those tasks that require physical interaction with the world and with humans in particular that usually shape people's imagination about AI. When asking people about AI systems, navigation (e.g., going from one place to another safely) is an important subgoal of many of these systems. We have selected two very relevant and different technologies in this category, *self-driving cars* and *home cleaning robots*. Again, when robotics is combined with AI we expect these physical systems not to be controlled by humans (locally or remotely) but be given instructions (e.g., where to go and what to clean) and follow them autonomously. The following two examples are good examples of AI technologies that represent *systems that interact physicall*y.

4.6.1 Technology: Self-Driving Cars

AI is changing the act of driving itself: automated technologies already assist drivers and help prevent accidents. As vehicle automation is progressively reaching new levels, these technologies are becoming one of

60 https://cirruslogistics.com/products/marine-enterprise-suite/

61 https://www.paragonrouting.com/en-gb/our-products/routing-and-scheduling/integrated-fleets/

62 https://www.paradoxsci.com/transportation-logistics-software-rst

63 https://www.mjc2.com/transport-logistics-management.htm

64 https://www.hacon.de/en/solutions/train-capacity-planning/

65 https://www.optibus.com/

the greatest forces transforming modern transportation systems. However, and despite extraordinary efforts from many of the leading names in tech and in automaking, fully-autonomous[66] cars are still out of reach except in special trial programmes[67], and their potential impact with respect to timing, uptake, and penetration remains uncertain[68].

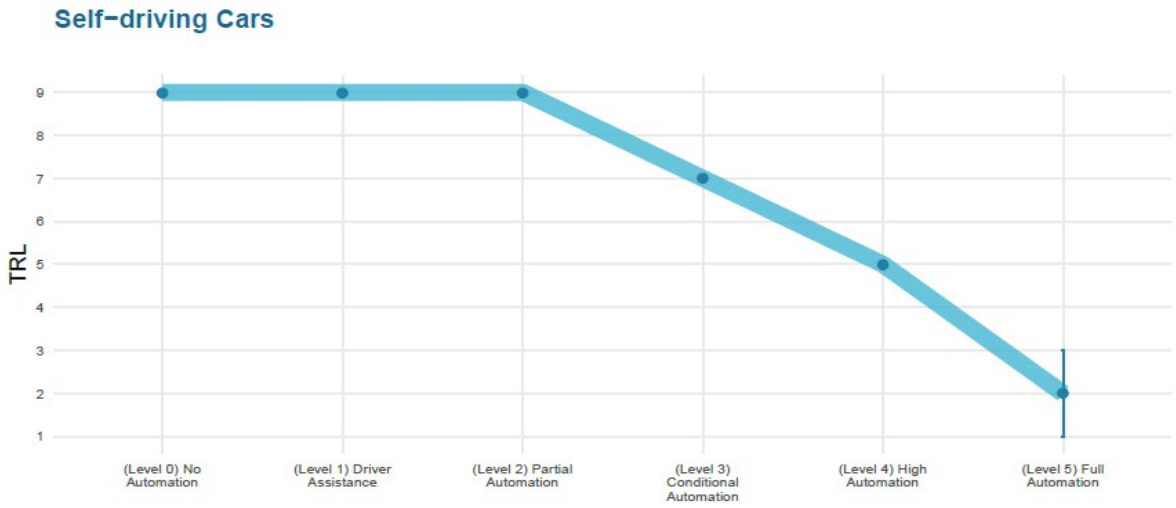

Figure 11. Readiness-vs-generality chart for self-driving cars technology. TRL 9 has been clearly reached by many cars in our roads in the levels between NHTSA levels 0 and 2 of automation. For NHTSA levels 3 and 4, current developments from automobile companies are currently performing research, prototyping and testing with self-driving cars (so TRLs are between 5 and 7). However, very low TRLs are still estimated for fully self-driving cars requiring no human attention at all.

While a generality scale could be developed in terms of the scenarios a fully-automated car could manage (e.g., from simple trips to complex situations), the discussion is usually set at identifying several levels of driving automation based on the amount of driver intervention and attentiveness required. In particular we use the US National Highway Traffic Safety Administration (NHTSA) definition of six levels of car autonomy to evaluate the self-driving capabilities of cars[69]. They released this guidance to both push forward and standardise autonomous vehicle testing. The 'NHTSA levels' (which we use here as levels of generality) are the following:

- **Level 1 (NHTSA Level 0) - No Automation**: A Level 0 car has no self-driving capabilities at all.
- **Level 2 (NHTSA Level 1) - Driver Assistance**: A Level 1 vehicle can assist with either steering or braking, but not both at the same time.
- **Level 3 (NHTSA Level 2) - Partial Automation**: A Level 2 vehicle can assist with both steering and braking at the same time.
- **Level 4 (NHTSA Level 3) - Conditional Automation**: The vehicle itself controls all monitoring of the environment (using sensors like LiDAR).
- **Level 5 (NHTSA Level 4) - High Automation**: At Levels 4 and 5, the vehicle is capable of steering, braking, accelerating, monitoring the vehicle and roadway as well as responding to events, determining when to change lanes, turn, and use signals.

66 We do not want completely-autonomous vehicles choosing where to go. By autonomous, we usually mean a vehicle that is capable of sensing its environment and moving safely with little or no human input, apart from the destination and preferences commands.

67 https://www.vox.com/future-perfect/2020/2/14/21063487/self-driving-cars-autonomous-vehicles-waymo-cruise-uber

68 https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world

69 https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

- **Level 6 (NHTSA Level 5) - Full Automation**: This level of autonomous driving requires absolutely no human attention.

For the first level (level 0), there is no automation at all as the human does all the driving at all times. Realistically, until recently, most vehicles were Level 0 (TRL 9). In turn, in the second level (level 1) we find some assistance systems for driving and maintenance such as the Adaptive Cruise Control (ACC) that is in charge of handling the braking systems to, for instance, keep a specified distance from the car in front of you, but it has no other control over the car (e.g., steering). In this regard, most, if not all, brands and automobile groups (e.g., PSA, VAG, General Motors, Daimler, etc.) incorporate ACC to their models nowadays (TRL 9).

Moving to the third level (level 2), the vehicle may assist with both steering and braking at the same time but it still requires full driver attention, and the driver must be ready to take over at any time. Combining adaptive cruise control (from Level 1) with lane centering (or auto steer, a mechanism that keeps a car centered in the lane) capabilities met the definition of Level 2. Tesla's Auto-Pilot[70] feature, as seen on the Model S, X, and 3, currently falls into the Level 2 category (TRL 9).

As for the fourth level (level 3), the driver's attention is still critical but they can leave the handling of some (critical) functions such as braking, and delegate them to the autonomous system in the vehicle when conditions are safe. Many current Level 3 vehicles require no human attention to the road at speeds under 37 miles per hour. Audi and others have already announced Level 3 autonomous cars to launch in 2018, but it has not actually happened due to the restrictive regulatory, technical, safety, behavioral, legal and business-related complications (TRL 8).

At the fifth level (level 4), although the vehicle is capable of steering, braking, accelerating, it would first notify the driver when there are safe conditions to take over the driving task, and only then does the driver may decide to switch the vehicle into autonomous mode. However, vehicles reaching this level of autonomy cannot determine between more complex and dynamic driving scenarios (e.g., traffic jams). In terms of developments, Honda has announced it is working towards a Level 4 vehicle by 2026[71]. Uber and Google's Waymo have also announced they have been working on Level 4 vehicles, but the reality is all their cars require safety drivers and they are currently testing their vehicles at Level 2 and 3 standards. Waymo is the exception as they are testing their prototypes at Level 4 conditions in the Early Access programme[72], but they are limiting the conditions in which the vehicles are allowed to drive (e.g., in dry weather areas).

Finally, for the sixth level (level 5), human attention should not be required at all and, therefore, there would be no need for pedals, brakes, or a steering wheel. The autonomous vehicle system would control all critical tasks, monitoring of the environment and identification of unique driving conditions like traffic jams. In this regard, although no commercial production of a level 5 vehicle exists, some of the aforementioned companies such as Waymo, Tesla or Uber are currently working towards this goal. As successful proof-of-concept we find Nuro[73] has been partnering with Krogers to test small cars that handle deliveries (within a short distance in a small, controlled area). Also, Waymo cars are navigating the streets of Arizona with no one behind the wheel[74], but fully self-driving cars are not here yet (TRL 1 to 3).

In general terms, and even if the technology is ready, most cars still sit between levels 1 and 3, typically with few or limited automated functions. There are some exceptions, such as certain Tesla models and Google's Waymo featuring a limited set of self-driving capabilities (e.g., enabling the car to steer, accelerate and brake on behalf of the driver), but still these are research projects in initial or testing/trial programmes[75]. Indeed, note that almost every major car manufacturer is currently performing research and testing with self-driving cars. This is yet another indication that manufactures have not even met the expectations (Narla et al. 2013) or

70 https://www.tesla.com/autopilot

71 https://hondanews.com/releases/honda-targeting-introduction-of-level-4-automated-driving-capability-by-2025

72 https://waymo.com/apply/

73 https://www.reviewgeek.com/1717/two-ex-googlers-want-nuro-a-new-self-driving-car-to-handle-your-deliveries/

74 https://www.theverge.com/2019/12/9/21000085/waymo-fully-driverless-car-self-driving-ride-hail-service-phoenix-arizona

75 https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/

the media announcements made just a few years ago[76] [77] [78] [79] claiming that by 2020 we all would be permanent backseat drivers[80]. This provides very gloomy evidence of the complexity and difficulty of the driving tasks (with the high level of reliability required (König et al. 2017)), and that even those simplest subtasks (e.g., tracking other vehicles and objects around a car on the road) are actually much trickier than they were thought to be.

4.6.2 Technology: Home cleaning robots

Home cleaning robots were one of the expectations of early AI and robotics. Home chores are at the same time considered to require low qualification and seen as a nuisance for which automation would represent a liberation. Many partial (non-AI) solutions have gone in this direction during the 20th century such as washing machines, non-robotic vacuum cleaners, and dish washers. However, robotic cleaners started to flourish as late as the 1990s[81]. They are currently used for helping humans with many kinds of simple domestic chores such as vacuum cleaning, floor cleaning, lawn mowing, pool cleaning or window cleaning.

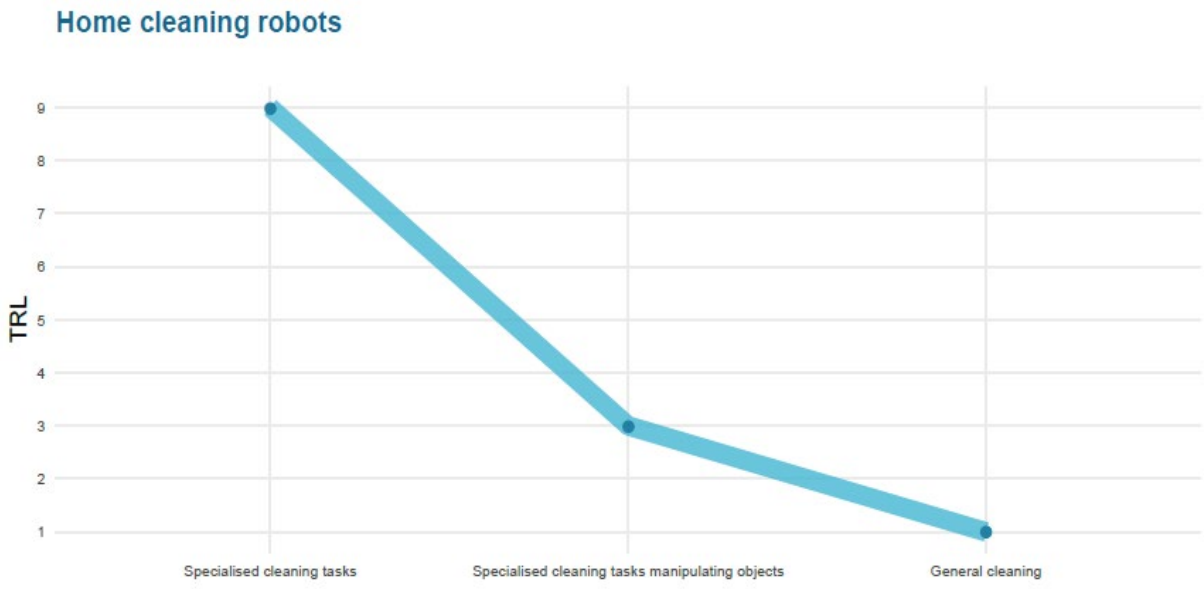

Figure 12. Readiness-vs-generality chart for home cleaning robot technology. While TRL 9 has been clearly reached by those specialised robots for dusting, vacuuming, mopping etc., lower TRLs are estimated when considering more complex house-cleaning tasks involving manipulation, flexibility, interaction, or coordination at any level.

However, despite their popularity, we can analyse how far we are from reaching the original goals if we analyse these technologies by their level of generality. We identify the following three levels:

- **Level 1 - Specialised cleaning tasks:** In this level we consider a robot that is able to do a particular cleaning task in a very specific way, such as dusting, vacuuming, mopping, doing the windows or cleaning the swimming pool, and other tasks that do not require manipulating new and diverse sets of objects (just

76 https://www.wired.com/story/gms-cruise-rolls-back-target-self-driving-cars/

77 https://www.theatlantic.com/technology/archive/2018/03/the-most-important-self-driving-car-announcement-yet/556712/

78 https://www.wsj.com/articles/toyota-aims-to-make-self-driving-cars-by-2020-1444136396

79 https://www.autotrader.com/car-shopping/self-driving-cars-honda-sets-2020-as-target-for-highly-automated-freeway-driving-266836

80 https://www.theguardian.com/technology/2015/sep/13/self-driving-cars-bmw-google-2020-driving

81 An early example of this is the 2001 Electrolux robot vacuum cleaner https://www.electroluxgroup.com/en/trilobite-advert-elubok115-2/

well-defined objects such as walls and windows, and avoiding obstacles). In this case, the robot has a physical configuration that exploits the particularities of the task, either as a roundish device moving on the floor or a small autonomous drone doing the windows.

- **Level 2 - Specialised cleaning tasks manipulating objects**: Here we consider that the task is still specific but involves the manipulation of a variability of household objects, including removing and putting back decorations and many other items, fold laundry, iron clothes, etc. (beds do not have a predefined size or location, clothes are very different, etc.). The robots, still purposed for a single task, may have different shapes and sizes for each application.
- **Level 3 - General cleaning**: in this final level of generality we expect the same robot to do many home chores. This means that the robot may have a more flexible physical configuration (probably with some type of robotic limbs) and a sophisticated interplay between perceptors and actuators. At this level we are not implying that these robots must be humanoid, clean exactly as humans do or use the same instruments (broom, vacuum cleaners, etc.).

Looking at Figure 12, for level 1, a clear evidence of TRL 9 can be found in the roundish robotic cleaners that roam around our houses vacuuming and sometimes mopping the floor. Many models exist, with some simple perception and navigation capabilities. Most of the innovations in the last decade have been towards better identifying walls and avoiding stairs using built-in sensors for autonomous navigation, mapping, decision making and planning. For instance, they are able to scan the room size, identify obstacles and perform the most efficient routes and methods. Some of them include capabilities from other categories (such as, speech recognition for voice commands or even basic conversation capabilities). However, they are still at this level, as they are not able to manipulate objects. A similar situation happens with other specific tasks such as windows cleaning[82], pool cleaning, lawn mowing or car washing.

The second level involves the manipulation of objects, which requires more advanced recognition of the environment and dexterity. There are current prototypes[83] to fold laundry (Bersch et al, 2011; Miller et al., 2012) or iron clothes[84] (Estevez et al., 2020). More complex tasks such as making the bed or clean the bathroom[85] are still a bit below working prototypes. Nevertheless, considering the best situation of all these specific cases, we have evidence of a TRL 3.

Finally, the third level is still in very early stages, and we do not have evidence to assign a value beyond TRL 1. About the near future, innovations are required at level 2, before moving to significant progress at level 3, with general-purpose service robots (Walker et al., 2019), which would become the real transformation drivers. Nevertheless, technology companies working on home robots (e.g., iRobot, Amazon, Samsung, Xiaomi, etc.) are still fighting for some other competitive advantages at level 1. For example, they add video conferencing and voice assistants to their devices rather than the ability to actually manipulate objects or diversify the physical tasks they can do. While some specialisation may be positive in the long term for cleaning (as any other activity), and there are some marketing and economic interests for going in this direction, having dozens of different gadgets at home has some limitations in terms of maintenance, sustainability and adoption. In the end, we could even envision the possibility that a robot at level 3 could replace dishwashing machines, vacuum cleaners and other specialised devices towards a more general home cleaner, especially in small apartments.

4.7 Social and collaborative intelligence

One of the key characteristics for the success of some species and human collectives is that they act as swarms, herds or social communities. Being able to interact successfully and collaborate with a diversity of other agents is an important capability that AI has focused on quite intensively, particularly in the area of multi-agent

82 https://www.digitaltrends.com/home/best-window-cleaning-robots/

83 https://www.calcalistech.com/ctech/articles/0,7340,L-3768535,00.html

84 https://helloeffie.com/

85 Some simple products (https://www.digitaltrends.com/home/giddel-toilet-cleaning-robot/) and incipient prototypes already exist (https://techcrunch.com/2020/03/04/this-bathroom-cleaning-robot-is-trained-in-vr-to-clean-up-after-you/).

systems (Wooldridge 2009). In the last category of this section, we again look for a technology that has limited overlap with some other categories (e.g., a robotic swarm would belong to this category and the previous one). Accordingly, we choose a paradigmatic case of this kind of social and collaborative agents, the technology about *negotiation agents*. This AI technology is representative of *systems that collaborate socially*.

4.7.1 Technology: Negotiation Agents

Negotiation is a complex decision-making between two or more peers to reach an agreement, such as an exchange of goods or services (Jonker et al. 2012). Even if decision theory (Steele et al. 2016), game theory (Myerson 2013) and multi-agent theories (Janssen 2002) are consolidated disciplines, many promises for the technology of negotiation agents are usually expressed as partial automation, i.e., as assistants for a negotiation. Here, we do not want to consider a third dimension about the level of automation, so we will cover the levels of generality and the levels of readiness assuming full autonomy: agents that negotiate autonomously (Jennings et al. 2001). Of course, guidelines and supervision may be given by humans (apart from the objective functions), but these agents should operate autonomously —the typical example is a stock market agent doing transactions in the night. For instance, this was the assumption of the automated negotiating agents competition (Baarlag et al. 2015), although the latter has incorporated new challenges over the years[86] (e.g., preference elicitation, human-agent negotiation; supply chain management, etc.).

By negotiation we also consider trading agents (Rodríguez-Aguilar et al. 1998, Wellman 2011) and we are transparent on the techniques that are used (argumentation techniques[87] or others), but we are a bit more specific than some umbrella terms such as "agreement technologies" (Ossowski 2012, Heras et al. 2012). In the end, the history of this area dates back to decision theory and game theory, which can find optimal policies when the protocol is known as well as the behaviour of other agents (Parsons et al. 2012). Things become more complicated in situations where agents can reach local optima instead of more desirable equilibria, or the rules of the game change during operation. In more general multi-agent systems, especially heterogeneous multi-agent systems (Perez et al. 2014), things become even more complicated as one has to consider that other agents may have different functions (proactiveness, involving different goal-directed behaviours) or they may even change. Finally, a more open-ended situation happens when there is bounded rationality, usually given by resources or by constraints imposed by real-time scenarios (Rosenfeld and Kraus 2009) and cases where theory of mind is needed for negotiation or coalitions (Von Der Osten et al. 2017).

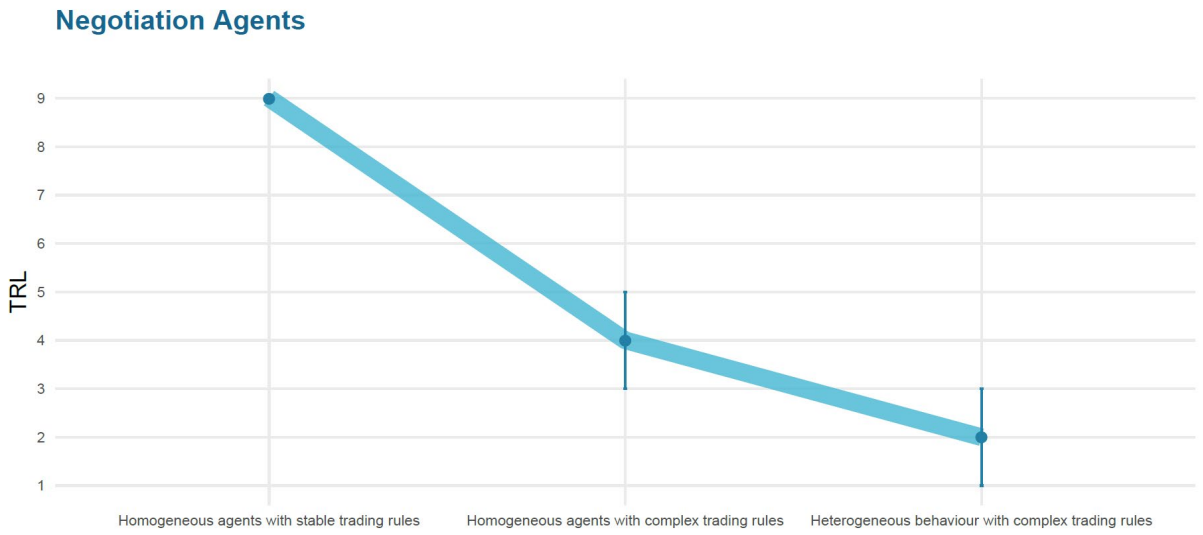

Figure 13: Readiness-vs-generality chart for negotiation agents technology. Level 1 reaches TRL 9, with some negotiation bots running in simple scenarios. Level 2 is more challenging, and TRL ranges between 3 and 5. Finally, level 3 is still far ahead in the future, with an estimated TRL between 1 and 3.

86 https://web.tuat.ac.jp/~katfuji/ANAC2020/

87 Note that considering "argumentation" as a negotiation technique is debatable; different views can be found from the area of computational argumentation, where negotiation is considered one of the multiple types of argumentative dialogues (see McBurney et al. 2002)

The levels we will use for negotiation agents are as follows:

- **Level 1 - Homogeneous agents with stable trading rules:** agents can get good trading and negotiation results if they maximise their utility and choose the immediate best action, independently of the other agents, or assuming all agents behave equally (homogeneous multi-agent system, with similar utility functions but possibly different parameters). The market rules are fixed and specific (e.g., a stock market trader), with no (frequent) local maxima and deadlocks.
- **Level 2 - Homogeneous agents with complex trading rules**: the trading rules become more complex and the global regulations can change. Local maxima and deadlocks are frequent and agents should act or coordinate to avoid them. Here we still assume all agents behave equally (i.e., homogeneous multi-agent system, with similar utility functions but possibly different parameters). There is no need to model different capacities as all the other agents are assumed to work under perfect rationality wrt. their functions.
- **Level 3 - Heterogeneous behaviour with complex trading rules**: Here we can now have agents with bounded rationality, changing goals and erratic behaviour, adversarial or malicious agents, including humans with very different motivations. At this level, we expect agents could benefit from a diversity of social strategies for negotiation such as persuasion, alliance-building, decoys, lying, manipulation, etc. This requires modelling the capabilities, goals and mind states of other agents, possibly in terms of BDI (beliefs, desires and intentions).

Note that the generality increases mostly because of the complexity and diversity of the trading rules and the other agents.

Early negotiation agents can be found at level 1 using the basics of decision theory (Parsons et al. 2012), and at this level many negotiating agents do not even need AI (Lin and Kraus 2012) but are coded manually with a few rules. Many of these systems populate restricted scenarios, such as the electricity grid, where participants must follow some strict regulations (which try to avoid deadlocks and shortages), but still leave enough flexibility for trading and rewarding those agents that behave more intelligently in the "smart grid" (Ramchurn et a. 2012). Still today, some systems exist at the macro-level, i.e., companies in electricity markets (Pereira et al. 2014), illustrated with real-data simulations, but the generalised use of smart agents at homes is still very incipient. Clearly, the area where trading agents are a developed product is in the stock and the currency markets, and more recently in cryptocurrencies. While they reach high TRLs at this level, there is the question of whether they really help their users (or owners) make profits88. Another common case both in research and with commercial applications is auction sniping as happens with online platforms such as ebay (Hu and Bolivar 2008). According to all this, we can assign TRL 9 to this level.

Level 2 expects the global regulations to change and the utility functions to have different values. These two aspects are sometimes referred together as "domain knowledge and preference elicitation" and, per 2017, is considered a "challenge" (Baarslag et al 2017), with some research in terms of on-line or incremental preference extraction (Baarslag et al. 2015b, Baarslag et al. 2017b), as well as in domain modeling (Hindriks et al. 2008, Sanders et al. 2008, Simonsen et al 2012). However, in some scenarios such as e-commerce between companies, there have been some patents being filed89 (Krasadakis 2016). Furthermore, in (Fatima et al. 2014) [chapter 12] a number of applications (e.g., grid computing, load balancing, resource allocation, etc.) can be found regarding trading agents with bounded rationality and limited knowledge about the domain. Given all the above, we consider a range between TRL 3 and TRL 5 for this level as all the activity is still in the research and prototyping phases.

When it comes to level 3, we have seen much activity at the research levels, with methods with bounded rationality and heterogeneous utility functions, working for simulations, with specific contexts (Rosenfeld and Kraus 2009) or theoretically (Sofy and Sarne 2014), or considering volatility of information or partial knowledge (Adam et al. 2014). Only a few are trying to use mind modelling in a general way (Von Der Osten et al. 2017),

88 https://medium.com/@victorhogrefe/how-effective-are-trading-bots-really-1684acc1f496, https://3commas.io/blog/best-crypto-trading-bot

89 https://medium.com/innovation-machine/a-buyer-bot-negotiating-with-a-seller-bot-7026f79ac51e

but still in restricted scenarios (games). Because of the lack of working evidence in general settings we assume a value of TRL between 1 and 3 for this level.

Level 3 captures a wide spectrum of possibilities and could be refined in the future as agents start to have better mind modelling capabilities. However, if we take the high edge of this level, such as understanding and performing well in complex machine-human environments, even if only restricted to trading, these are clearly challenging scenarios even for human scientists (Rahwan et al. 2019), so we expect a long time to reach high levels at this level.

5 Discussion: Rearranging the Generality

After the series of examples of AI technologies seen in the previous section, organised into one of the seven AI categories, we can extract general insights from what we observe in the readiness-vs-generality plots more globally.

Methodologically, the examples serve to illustrate the challenges of estimating the TRLs, a problem that is not specific to AI. The use of levels of generality on the *x*-axis, however, has helped us be more precise with the TRLs than would be otherwise. In fact, there is no such a thing as TRL 3 or TRL 7 for machine translation, unless we also specify the level of generality (scope of functionality) for the technology. This is the first take-away of this methodology. Of course, the levels in these examples could be refined and made even more granular, possibly reducing the error bars in some cases. In those cases where there is no standardised scale for the generality axis (as for self-driving cars or machine translation), an open discussion in the particular community to find a consensus would be very welcome.

The shapes of the curves seen in the charts of the previous section are informative about where the real challenges are for some technologies. Going from 70% to 80% in a benchmark is usually a matter of time and can be circumvented without a radical new innovation, but in many cases going from TRL 1 to TRL 7, for instance, needs something more profound than incremental research and development. Consequently, it seems that those curves that are flatter (see Figures 4 - recommendation engines, 8 - facial recognition, 10 - transport scheduling systems and 11 - self-driving cars) look more promising than those for which there is a steep step at some level on the x-axis (see Figures 3 - expert systems, 5 - apprentice by demonstration, 9 - text recognition, 12 - home cleaning robots and 13 - negotiation agents). Importantly, the shape of the curves depends on the definition of levels in the *x*-axis (all charts are summarised in the following subsection, see Figure 15).

Refining one level into two or three more granular levels may produce a flatter curve (e.g., smoothing the step curves). This is also a good indication of a way in which an insurmountable level of generality can be disaggregated into more gradual steps, which may lead to new research and development tracks taking AI to high TRLs. This is also what happened in the past with some technologies. For instance, robotic vacuum cleaners added a small, yet relevant, intermediate step that took the technology to TRL 9, created an ecosystem of companies and users, which in the end paves the way for more research effort and investment in the following steps or refinements on the *x*-axis.

In the opposite direction to disaggregation, there is also a unifying trend to consider technologies that, by definition, are expected to integrate many capabilities. A very good example of these integrating AI technologies is represented by virtual assistants, because they are expected to cover a wide range of tasks that integrate capabilities that are associated with many categories in AI, including knowledge representation and reasoning, learning, perception, communication, etc. Let us explore this technology in particular and derive its readiness-vs-generality charts.

5.1 An integrative AI technology: virtual assistants

Virtual Assistants (VA), also known as intelligent personal assistants or digital assistants, are applications or devices meant to interact with an end user in a natural way, to answer questions, follow a conversation or accomplish other tasks. VAs have expanded rapidly over the last decade with many new products and capabilities (EC Report 2018, Hoy 2018). Alexa, Siri, Cortana or Google Assistant are very well-known examples of this technology. The idea of a computer humans could *meaningfully and purposely* dialogue with is also one of the early visions of AI (Turing 1950), but as with many early visions, it is taking decades to materialise. Having a meaningful conversation is not always easy with humans of different backgrounds, culture, and knowledge, and making it purposeful (so that the speaker gets things done) is also a challenge in *human* communication. It is no surprise then that these are two important hurdles to overcome when trying to get something similar with *machines*.

As said above, domain generality is very important, because we want these systems to do a wide range of tasks. However, this is more a desideratum than a reality, or even a necessity for some applications. This is similar to the cases with other AI technologies analysed in the previous section, such as expert systems. In

particular, making an assistant for a narrow domain (a telecommunication company assistant or a ticket-purchasing service avatar) is easier than a more general assistant (an executive assistant in the workplace).

Given these considerations, we introduce a tentative three-level scale for generality of virtual assistants as shown in Figure 14, which may of course be refined in the future.

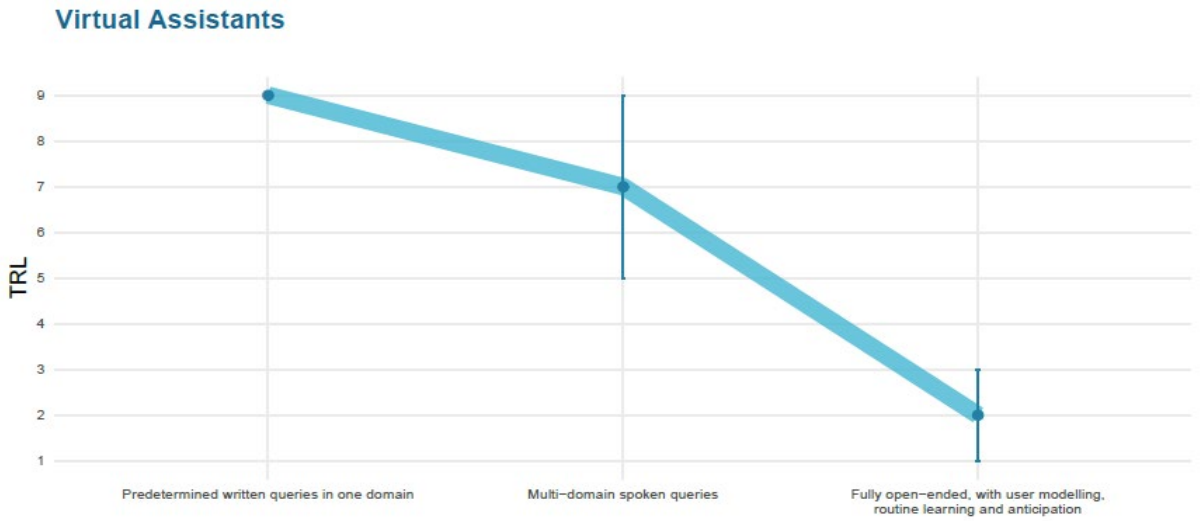

Figure 14. Readiness-vs-generality chart for virtual assistant technology. TRL 9 has been reached for systems that work with predetermined written queries (generality level 1), high TRL are more diverse with open-ended spoken queries (generality level 2). Finally, the most advanced level requires generality in terms of domains, types of interactions and queries from the user (generality level 3). Error bars show some uncertainty in the assessment.

The *x*-axis of Figure 14 reflects three generality levels of virtual assistants:

- **Level 1 - predetermined written queries in one domain**: queries are restricted or should contain some keywords the system recognises to find the topic and some related information. Answers are either template-based or pre-recorded as text (possibly read by synthesisers).
- **Level 2 - multi-domain spoken queries**: text and voice commands can be received with an unrestricted language, the answers are constructed and not stored. Questions and queries may cover a diverse range of domains.
- **Level 3 - fully open-ended, with user modelling, routine learning and anticipation**: the most advanced level requires generality in terms of domains, types of interactions and queries from the user. The system may be proactive rather than just reactive.

In terms of capabilities, and as shown in Figure 14, the simplest VAs (generality level 1) are conceived as straightforward software agents able to perform simple tasks or give straight answers based on templates or predefined commands or questions. We can find examples of this type of VAs in commercial products, in the form of simple chatbots in customer-service applications on websites and other apps for restricted (simple) domains (e.g., ticket purchase assistants, VAs for QA of Coronavirus-related content[90] [91], etc.). Consequently, we could assign TRL 9 to these assistants.

Focusing on level 2, these VAs should be able to interpret human speech and respond via constructed complex answers using synthesised voices, sometimes emulating simple dialogues and conversations. Users may be able

90 https://avaamo.ai/projectcovid/

91 https://www.hyro.ai/covid-19

to ask their assistants (open) questions (with limited proactivity), control home automation devices and media playback via voice, and manage other basic tasks such as email, to-do lists, and calendars with verbal commands. It seems that TRLs are high in this case too. However, although VAs are seen (and marketed) as intelligent assistants capable not just of understanding but of taking decisions and fully supporting humans, this vision has not fully materialised yet. Currently, there are a number of VAs in the market, with Google Home, Amazon Echo, Apple Siri and Microsoft Cortana (Hoy 2018) being the main examples. These companies are constantly developing, testing, and demonstrating new features and capabilities for their VAs, and we can see this evolution and improvements as new versions are launched on the market. Because of this, we plot a range of values between TRL 7 and 9, as shown in the figure.

Finally, VAs with level 3 of generality are envisaged to have more advanced capabilities, including background knowledge so humans will be able to have (professional) conversations and discussions on any topic, more advanced dialogue management, or improved reasoning about the world, among other things[92]. In this level, VAs are assumed to understand context-based language complexities such as irony, prosody, emotions, meaningful pauses, etc. We think this is at a research stage today (TRLs 1 to 3). Note that, even in level 3, VAs are not expected to perform complex rationales or make sophisticated decisions. This is covered by technologies such as expert systems or planning. Of course, once high TRLs are obtained in these technologies they may end up being incorporated in VAs, as they are usually shipped as integrators of AI services.

5.2 Delineating technologies more precisely

From the previous discussion we see how important it is to refine the levels of generality such that levels are sufficiently crisp for a more accurate assessment of TRLs. This becomes more difficult as the technology is broader, especially those that are defined by integrating capabilities from different categories of AI, such as the VA in the previous section. Precisely because of this difficulty, we have to be wary of the bias and misconceptions our explicit or implicit assumptions of generality can create.

For instance, many funding calls, especially after the H2020 programme, ask for a particular TRL. While this is relevant in calls that are oriented towards products that can be distributed in the market as the project is completed, it is important to look at the dynamics of readiness-vs-generality charts and the pressure for avoiding generality. For the purpose of high TRLs, some research projects may be tempted to solve simplified versions of the problems or solutions for very narrow domains, with many ad-hoc tweaks, rather than solving the general problem. These calls even encourage that the technology is illustrated in one domain, which is carefully chosen by the researchers as one in which a very specific set of techniques can really work. But, in the end, the technology may not extrapolate to other domains, and its transformative effect may be very limited. This is particularly important in calls such as FET (Future and Emerging Technologies[93]). Of course, some bottom-up approaches that work in a particular domain end up being generalisable to other domains, but this should be explicit for evaluation purposes.

This generality issue is also critical in early stages of research. Research papers and benchmarks should consider a wide range of domains, especially when new principles and techniques are introduced. Otherwise, their purported performance should be scrutinised very carefully. Media and the scientific community itself are usually more amazed by the first time something is achieved (e.g., beating a human master in Go) than how it is achieved. For instance, the first publications about AlphaGo (Silver et al, 2016) had more public repercussions than other research papers that followed generalising the techniques for any board game, and without precoded human knowledge (Silver et al., 2017; Silver et al., 2017b; Schrittwieser et al., 2019).

Figure 15 includes a summarised view of all readiness-vs-generality charts. Note that, there are many other ways in which we can look at these plots to compare the twelve AI technologies we have analysed. We can see, for instance, the highest level with TRL-9 as a simple way of comparing the technologies. Alternatively, we can just focus on the narrowest (leftmost) level. Although, this would give a much promising view of the state of the art of AI technologies, it has sometimes been the case that specialised early stages have paved the way for more general versions of the technology.

92 https://www.cnet.com/news/facebook-ai-chief-we-want-to-make-smart-assistants-that-have-common-sense/

93 https://ec.europa.eu/programmes/horizon2020/en/h2020-section/future-and-emerging-technologies

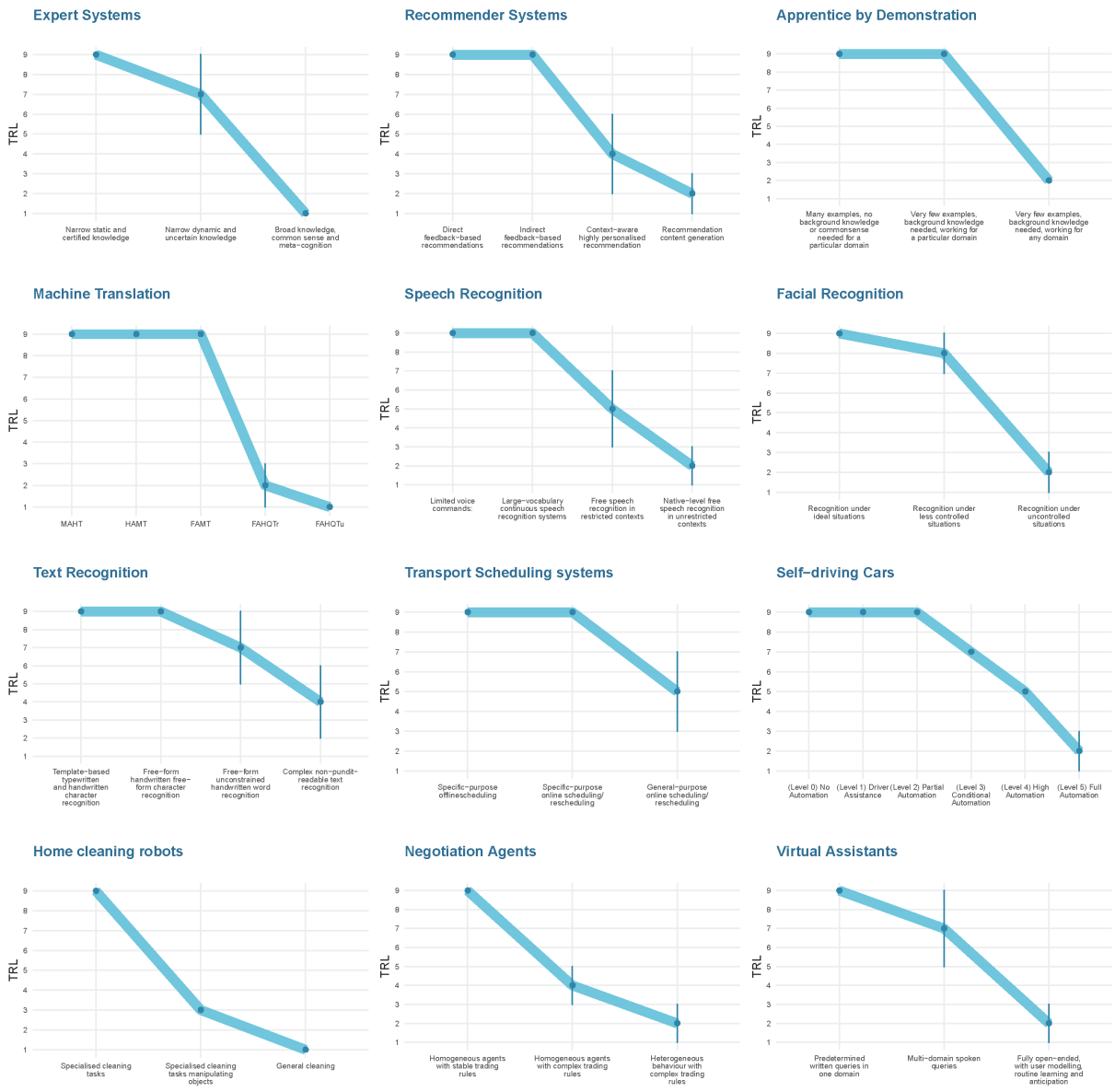

Figure 15. A composition of all readiness-vs-generality charts from Figures 3 to 14.

On the other hand, asking for too much generality has the risk of entering an area that is not well understood yet (Bhatnagar et al, 2017, Martínez-Plumed et al. 2020a, 2020b), and a project or a paper may end up aiming at some vague understanding of "artificial general intelligence" or slip into dubious terms such as "human-level machine intelligence", which cannot be properly evaluated (Hernández-Orallo 2020). In contrast, we think that the use of TRLs, while at the same time being precise and ambitious on how to certify the position on these readiness-vs-generality charts, may be of utmost importance to track the impact (Makridakis 2017) of AI and anticipate the key transformations of the future. We explore this in more detail in the following subsection.

5.3 Assessing TRLs more precisely: the *AIcollaboratory*

In this report, we have assessed the TRL of each technology (at a particular level) by asking experts (including ourselves) to follow the guideline in the Appendix to estimate the particular readiness level in the scale. A

wider group of experts, using more extensive training on the TRLs and usual methods for aggregation or consensus of opinions (such as Delphi) would bring more robustness to these estimates, including a systematic way of deriving the error bars. However, the estimates would still be based on expert evidence but not quantitative evidence.

There are some sources of information that allow us to assess TRLs such as the number of patents or the sales of particular AI-related products. However, we do not think that this information would be sufficient on its own to understand or quantify the TRL for many AI technologies, especially considering such data is historical (i.e. analysing the past), and therefore not ideal to address the future-oriented concerns of this report. Coverage on the media could also be a relevant source, and we could use relevant sources such as AI topics[94] (Martínez-Plumed et al. 2018b, Hernández-Orallo 2020). However, there is an important source of quantitative information on the progress in AI: benchmarks and competitions (Hernández-Orallo et al. 2017).

The relation between benchmarks and TRLs is more complex than it may seem. Some AI benchmarks (e.g., Atari games) would qualify as “simulated environments'' mentioned in the description of TRL 5 or TRL 6, depending on whether only some components or a complete autonomous system are being assessed through them. Other benchmarks, such as those used for self-driving cars would qualify as "operational testing platforms" for TRL 7. Other benchmarks, e.g., some Kaggle competitions, are about real cases and their models could be applied directly, showing evidence for TRL 8. Benchmarks sometimes contain standardized information regarding elements that map onto different TRLs, and therefore can be useful in a TRL assessment. We have used these connections in some of the assessments in the previous sections. Doing a more systematic analysis of all benchmarks in AI, its corresponding technology and what kind of technology readiness level they could be associated with, would enable a more quantitative approach to estimating TRLs.

In this regard, we could use the AIcollaboratory[95] (Martínez-Plumed et al. 2020a, 2020b, 2020c) to collect intrinsic information characterising benchmarks and map out the relationships between them and TRLs. This initiative was conceived for the analysis, evaluation, comparison, and classification of AI systems, creating a unifying setting that incorporates data, knowledge and measurements to characterise them. The *AIcollaboratory* is designed to enable this kind of mapping. For the moment, we leave such mapping and quantitative analysis for future work and out of the scope of this report. It is not just the sheer volume of the endeavour but also because there are some issues to discuss and solve first in order to do this meaningfully and reliably. For instance, most benchmarks are not just pass or fail but are accompanied by one or more metrics, such as the performance level, which depend on the application domain and may even be opposed to each other for particular tasks. We should determine the minimum level of accuracy in a given benchmark that would be considered sufficient evidence for the associated TRL to be met.

Defining benchmarks to map onto TRLs could generate tension with assessments of the technology's generality. For instance, 70% performance on a face recognition benchmark could be considered useful for some applications and a proof of TRL7 but it may well happen that most of the remaining 30% errors would focus on a particular niche of the technology (e.g., noisy pictures). Would this be evidence of TRL7 at that particular level of generality or, rather, would it indicate the technology belongs to a lower level? We believe that performance thresholds to assign a TRL should be much higher (e.g., 99%) to avoid this kind of specialisation problem. Nevertheless, there are other issues, such as systems being specialised to the benchmark but not to the real problem (so that a TRL7 would never translate into a TRL9). Despite these challenges, we do think that clarifying and utilising the relationship between benchmark results and TRLs is a promising avenue of research, which we hope to develop in future work.

94 https://aitopics.org/

95 http://www.aicollaboratory.org/

6 AI progress through TRLs: the future

The analysis of a readiness-vs-generality chart may constitute a useful tool to understand the state of the art of a particular technology. However, can it be useful for anticipating the future?

In the first place, as we already mentioned, a static picture can give us hints about what is expected in the near future. A very steep curve (such as in Figure 4 - apprentice by demonstration) suggests that there may be a long way to go from one generality level to the next in the technology to the next one. The gap may include significant discoveries, results, or inventions at some low TRLs, which may involve fundamental research, usually linked to slower progress. A flatter curve (such as in Figure 7 - facial recognition) may correspond to situations where the fundamental ideas are already there, and progress could be smoother. But this has another reading, a flatter curve with no level reaching TRL 9 means that the technology has not reached the market successfully and the industry ecosystem is non-existent, which would otherwise invest money and research teams on the problem. Yet, at the same time, this is only partially true, as some sectors already exist before automation. For self-driving cars, there is an ecosystem of very powerful automobile multinationals, with no self-driving car technology until very recently. These companies have invested huge amounts of money in this technology. Also, some tech giants can go from low TRLs emerging from new techniques to working products in less than a year, as happened, for instance, with the language model BERT (Devlin et al., 2018) being applied to Google's search engine[96].

To better understand the speed of progress, we also need to consider the notion of technology "hyper adoption", which is related to Gartner's Hype Cycle from Gartner (Linden et al. 2003). This theory states that people adapt to and adopt new technologies much faster than they used to in the past. This may be partially caused by the so-called "democratisation" of new technology innovations, as they become available to large parts of the population as soon as they enter the market. For instance, electricity took 70 years for mass adoption, but the Internet took just 20 years. The same is happening with AI technologies. A clear example is the current hyper-adoption of voice-related technology[97], with all the tech giants such as Amazon, Google and Microsoft launching new products every few months. It may be the case that developments in this sort of technology has enhanced the adoption rates of voice assistants, and vice versa. The trend may even stop because of ageing populations in many countries, which are more reluctant towards technological innovations.

In order to have more ground for extrapolations we would need a less static picture of the evolution of AI technologies. Having information about the charts in past years would give us data about how curves evolve, and how some TRL transitions are faster than others. Of course there may be no clear trends or trends that cease to hold because of some changes in the AI playground or in society (e.g., a financial crisis, a pandemic or I the lack of market enthusiasm and/or low investment). We can do a simple exercise with the VA technology seen in the previous section. Can we compare the "picture" (i.e., the readiness-vs-generality charts) with a historical perspective?

6.1 Readiness trends

Figure 14 shows that, in the case of virtual assistants, there has been important progress at level 2 of generality in recent years, and level 3 may be changing rapidly to higher TRLs because of high investment and the ubiquity of VAs of level 2 of generality. We see this evolution from the 1990s, where digital speech recognition technology became a feature of personal computers of brands such as Microsoft, IBM or Philips, but without conversational or Question and Answering (QA) capabilities. In 1994, IBM launched the very first smartphone (IBM Simon) with some assistant-like capabilities: sending emails, setting up calendars and agenda, taking notes (with a primitive predictive text system installed) or even downloading programmes! However, it was a menu-based interaction, very different from the assistants we know today. In this regard we may estimate that some research on this was being performed (TRL 1 to TRL 3), mostly focused on the field of speech recognition. This went in parallel with advances during the 1970s and 1980s in computational linguistics leading to the development of text comprehension and question answering projects for restricted scenarios such as the Unix Consultant (Wilensky 1987) for answering questions about Unix OS or LILOG (Rollinger 1991) in the domain of tourist information. These projects never went past the stage of successful demonstrations in relevant scenarios (TRL 7).

96 https://www.blog.google/products/search/search-language-understanding-bert/

97 https://www.forbes.com/sites/forbestechcouncil/2018/06/08/the-hyper-adoption-of-voice-technology

By the decade of the 2000s, not only were there relevant advances in speech recognition technology, but also in QA (with market-ready products such as Wolfram Alpha (Wolfram 2009)), Information Retrieval and knowledge-Based Systems that paved the way for future VA systems. One important milestone in this decade was the launch of Google Voice Search in 2002 (Franz et al., 2002). The system allowed users to request information by speaking to a phone or computer rather than typing in the search box. This can be considered as the first step in launching Google's VA. This is a significant milestone not only due to the change in the power-efficient computing paradigm (they offload the processing power to its data centres), but because Google was able to collect gigantic amounts of data from billions of searches, which could help the company improve its prediction models of what a person is actually saying. At the same time, IBM also pushed its research in QA and information retrieval during this decade (from 2005 onwards) with a specific goal in mind: to be able to compete successfully on *Jeopardy!* The first prototypes and demonstrations of their system, *Watson* (Ferrucci 2012), were developed and tested between 2007 and 2010, prior to their success in 2011. From all the above, we may extract that much research, testing and development was being performed in those areas related to the VA (TRL 3 to 7) even without market-ready products being launched.

Finally, VAs have witnessed a quick growth in terms of development, products and adoption by consumers during the last decade. The very first modern digital virtual assistant with voice-based communication capabilities installed on a smartphone was Siri[98], specifically on the iPhone 4S in 2011. Apple hit the market first but was soon followed by some big players' developments and products including Google Now (2012) Microsoft Cortana (2013) or Amazon Echo (2014) (Hoy 2018). As already explained, all these VAs have been further developed and improved during the last few years, where manufacturers are constantly testing and including new and more powerful capabilities (TRL 7 to TRL 9) in terms of interpreting human speech (via open questions), answering via constructed complex outputs, simple dialogue and conversational capabilities, and further advanced control over basic tasks (email, calendar, etc.) as well as home automation devices and media playback via verbal commands.

For level 3 of generality, VAs are foreseen to have much more advanced capabilities (e.g., background knowledge, open-domain conversations, common-sense reasoning, etc.) that were not found in the research agenda (TRL 1 to TRL 3) of natural language processing, planning, learning or reasoning until high TRLs have been obtained for the second level of generality. Note that the high TRLs of the latter were largely due to the huge advancements in hardware (e.g., computing infrastructure), software (e.g., powerful neural-based approaches) and data (e.g., people's behaviour, language corpus, etc.).

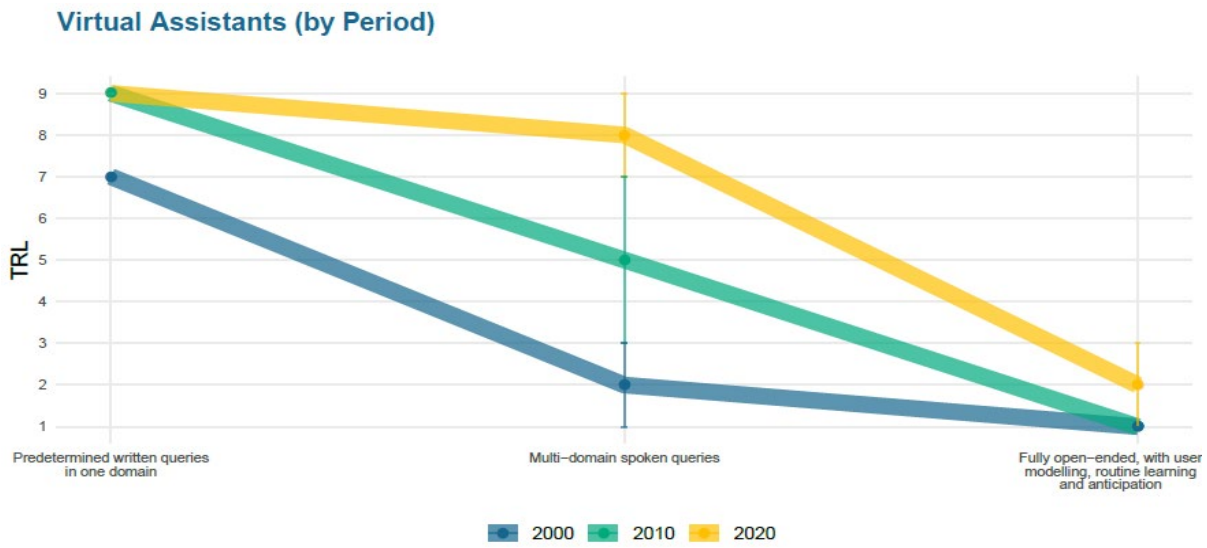

Figure 14. Readiness-vs-generality chart for virtual assistant technology at different moments in time (yellow: 2020, green: 2010, blue: 2000). We see how the "curve" has evolved from a steep one in the year 2000 located on the first level to another, also steep, from the second in 2020.

98 https://www.apple.com/siri/

Even if there are many uncertainties when assessing and inspecting these curves, with time we think that the juxtaposed historical view of TRL evolution for a given AI technology is more robust than the evolution of a single point (the technology at the same level). And it is much better than the analysis of the evolution of the technology mixing levels on the *x*-axis, because each period has a potential horizon for the technology. With this usual mistake we could have said that there has been no progress in smart phones in the past ten years once the penetration of devices reached near 100%. The percentage of time we use them has increased, because they have increased the generality of tasks and activities they can do, so their transformation goes on.

6.2 AI futures

There are many ways in which AI futures can be extrapolated, from expert panels (Müller and Bostrom 2016, Betz et al. 2018) to role-play scenarios (Avin 2019). There are also many visions about what will be possible in the future, with mixed success (Kurzweil 2005), poor specification[99] or not meeting any AI forecasting desiderata[100] (Dafoe 2018, Ap. A). Be relying on measurable indicators, it is possible to connect the progress in AI with some economic indicators (such as the PREDICT dataset[101]). In this paper, however, we have adopted an approach based on TRLs, to describe the state of the art of a discipline (which may be of use in applications such as project assessment or product development). For this reason, we have outlined some ideas on how to use this methodology for forecasting purposes.

The truth is that we are still terribly bad at predicting what capabilities and products will become a reality even in the short term, a problem that is not specific to AI but all technology, and particularly digital technologies. We are not always successful, even with hindsight (Martínez-Plumed et al. 2018b), in understanding why some potentials are not fulfilled, and why some technologies have limitations, and what kind of new technologies may replace them (Marcus 2020). While some criticisms in the early days of AI were related to scalability (the ideas worked for toy problems but were intractable in general), more recently most criticisms of AI are related to the lack of generality of current AI technologies. This is one reason for expressing generality as a dimension in our representations and measurements and is key to determine the maturity of a technology and forecast its transformative power.

Generality is also a key element when related to mass production and hence society's digital transformation. If a system is specialised for one particular domain, the return on investment —R&D investment— would be smaller than if the technology is applicable to a wide range of areas. Even a minor gain that takes place in many devices usually represents more money than a major gain in a few devices. Of course, many of these devices or apps can still be very specific (e.g., a watch), so this does not necessarily go in the direction of full generality but can still achieve massive penetration. When a widespread system becomes more general (e.g., a mobile phone, useful for calls and SMSs, turns into a smart phone, with apps), the transformation becomes huge. It is no wonder that virtual assistants, which can be distributed on every device (from phones to smart homes), if combined with a highly-general of tasks, may represent a major transformation in the years to come. Hence the interest by tech giants in investing in this technology.

If the dimensions are right, **high TRLs for high-level (i.e., broad) generalities should indicate potential short-term or mid-term massive transformative power** (see, for instance, Figures 6 - speech recognition, 7 - facial recognition, 8 - text recognition or 10 - self-driving cars). However, generality requires effort, and has associated costs. There are some internalities and externalities about a technology (e.g., environmental footprints, user privacy, skill atrophy, etc.) that should be considered refining these predictions. For instance, a given technology can be ready but the costs of deployment may not be affordable for the consumers (these costs can include data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, development time, etc., apart from monetary costs for research and development) (Martínez-Plumed et al. 2018a, Spelda et al., 2020). For instance, self-driving car technology can be based on

99 https://www.lesswrong.com/posts/yy3FCmdAbgSLePD7H/how-to-write-good-ai-forecasting-questions-question-database

100 Indicators for relevant AI-related achievements (e.g., a new capability that would pose a substantial employment threat to a large group of people)

101 https://ec.europa.eu/jrc/en/publication/2018-predict-dataset

radar or cheap cameras. While mass production can reduce the cost of radars, having self-driving capabilities for cheap cars (those most people have) may give advantage to technologies that rely on computer vision rather than radar tracking[102]. Even if a device is flooding the market, that does not mean it will be used extensively: if the novelty just wears off, it will be forgotten shortly after (as happens with many gadgets). Sometimes products are sold before they are effectively ready, just to make a positioning in the market, or because of some other commercial reasons such as meeting customers' expectations. The success of a technology is therefore an even more difficult variable to estimate, as many social and economic factors may interplay (Schilling 1998). For instance, if a technology is deployed too early, it may rebound with a backlash from consumers (e.g., Microsoft Clippy created aversion against assistants (Veletsianos 2007, PCMagazine 2001)), or human labour costs may fluctuate, accelerating or slowing the adoption of certain technology (e.g., mechanisation and automation have facilitated an increase in the speed of production (Miozzo et al. 2005, Borghans et al. 2006, Suri 2011))). In other words, **technological readiness does not mean technological success**. Analysing all the factors contributing to such success is out of the scope of this paper, and in the case of AI may require a particular analysis in the same way we have done here for the TRLs, but in terms of technology success rather than maturity.

What we have covered in this paper is an example-based methodology where (1) we identify the technology, its category and its scope, (2) we recognise and define the levels of generality that are most meaningful for the technology and appropriate to estimate the TRLs accurately, (3) we find evidence in the scientific literature and industry to identify the points on the readiness-vs-generality chart, and (4) we use the chart to understand the state of the art of the technology and extrapolate its future trends. The examples selected in this paper are also sufficiently representative for a discussion about the future of AI and how these charts can be used for short-term and mid-term forecasting.

As future work, there are many avenues we would like to see explored. First, the reliability of the assessments could be increased by using external experts for each chart. There is an opportunity for a consultation with the AI community asking for their views, suggestions, and evidence of the TRL levels. With a larger and wider group of experts methods such as Delphi could be used. Furthermore, we could develop new scales based on generality, autonomy, intelligence, etc., better understanding the different AI technologies and their evolution. We could also derive the TRLs from the results of the related benchmarks for each technology, as discussed at the end of the previous section. Second, covering many more AI technologies and their evolution would give a more complete picture than what we portray here, with a choice of representative AI technologies. Third, for many technologies there is an important discussion about the "right" levels of generality. In some cases there may be different scales or even multidimensional (e.g., hierarchical) scales to explore. Finally, there is also an opportunity to use the proposed methodology and results to generate an agenda of challenges for AI, particularly for those higher levels of generality which are currently acting as constraints to higher TRLs.

There is an enormous interest in the futures of AI and its impact. But massive impact can only be reached when the technology is really transformative. This only happens when new ideas, expertise and innovation reach maturity and they are widely applicable. Using the technology readiness levels and combining them with levels of generality, as we have done in this paper, can allow for the exploration of fresh perspectives on the state of the art of artificial intelligence, and how it may come to affect our society in the near future.

102 It has been argued that episodes of acceleration in technological progress were driven by particular General Purpose Technologies (GPTs) as this sort of technologies have the power to change the pace and direction of economic progress. Illustratively, in (Petralia 2017) the case of electrical and electronic technologies is discussed.

References

Acharyya, A., Rakshit, S., Sarkar, R., Basu, S., & Nasipuri, M. (2013). Handwritten word recognition using MLP based classifier: a holistic approach. International Journal of Computer Science Issues (IJCSI), 10(2 Part 2), 422.

Abril, M., Barber, F., Tormos, P., Lova, A., Ingolotti, L. and Salido, M.A., 2006. A Decision Support System for railway timetabling (MOM): the Spanish case. Computers in Railways X: Computer System Design and Operation in the Railway and Other Transit Systems, 10, p.235.

Adam, E., Grislin, E., & Mandiau, R. (2014). Autonomous agents in dynamic environment: a necessary volatility of the knowledge. In Trends in Practical Applications of Heterogeneous Multi-agent Systems. The PAAMS Collection (pp. 103-110). Springer.

Ahmed, A., Teo, C.H., Vishwanathan, S.V.N. and Smola, A., 2012, February. Fair and balanced: Learning to present news stories. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 333-342).

Ahmed, M.N., Toor, A.S., O'Neil, K. and Friedland, D., 2017. Cognitive computing and the future of health care cognitive computing and the future of healthcare: the cognitive power of IBM Watson has the potential to transform global personalized medicine. IEEE pulse, 8(3), pp.4-9.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Amatriain, X., & Basilico, J. (2016, September). Past, present, and future of recommender systems: An industry perspective. In Proceedings of the 10th ACM Conference on Recommender Systems (pp. 211-214).

Avin, S., 2019. Exploring artificial intelligence futures. Journal of AI Humanities. Available at https://doi. org/10.17863/CAM, 35812.

Bai, J., Chen, Z., Feng, B., & Xu, B. (2014, October). Image character recognition using deep convolutional neural network learned from different languages. In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 2560-2564). IEEE.

Baarslag, T., Aydoğan, R., Hindriks, K. V., Fujita, K., Ito, T., & Jonker, C. M. (2015). The automated negotiating agents competition, 2010–2015. AI Magazine, 36(4), 115-118.

Baarslag, T., & Gerding, E. H. (2015b). Optimal incremental preference elicitation during negotiation. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Baarslag, T., Kaisers, M., Gerding, E., Jonker, C. M., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. International Joint Conferences on Artificial Intelligence.

Baarslag, T., & Kaisers, M. (2017b). The value of information in automated negotiation: A decision model for eliciting user preferences. In Proceedings of the 16th conference on autonomous agents and multiagent systems (pp. 391-400).

Banks, G (1986). "Artificial intelligence in medical diagnosis: the INTERNIST/CADUCEUS approach". Critical Reviews in Medical Informatics. 1 (1): 23–54. PMID 3331578.

Beel, J., Breitinger, C., Langer, S., Lommatzsch, A., & Gipp, B. (2016). Towards reproducibility in recommender-systems research. User modeling and user-adapted interaction, 26(1), 69-101.

Bernice B. Brown (1968). "Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts.": An earlier paper published by RAND (Document No: P-3925, 1968, 15 pages)

Bersch, C., Pitzer, B., & Kammel, S. (2011, September). Bimanual robotic cloth manipulation for laundry folding. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1413-1419). IEEE.

Betz, U. A., Betz, F., Kim, R., Monks, B., & Phillips, F. (2019). Surveying the future of science, technology and business–A 35 year perspective. Technological Forecasting and Social Change, 144, 137-147.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., et al., (2017,). Mapping intelligence: Requirements and possibilities. In 3rd Conference on" Philosophy and Theory of Artificial Intelligence (pp. 117-135). Springer.

Bianchini, D., De Antonellis, V., De Franceschi, N., & Melchiori, M. (2017). PREFer: A prescription-based food recommender system. Computer Standards & Interfaces, 54, 64-75.

Bluche, T. (2016). Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In Advances in Neural Information Processing Systems (pp. 838-846).

Borghans, L., & Ter Weel, B. (2006). The division of labour, worker organisation, and technological change. The Economic Journal, 116(509), F45-F72.

Boyle, D.K., 2009. Controlling System Costs: Basic and Advanced Scheduling Manuals and Contemporary Issues in Transit Scheduling (Vol. 135). Transportation Research Board.

Brown, N. and Sandholm, T., 2019. Superhuman AI for multiplayer poker. Science, 365(6456), pp.885-890.

Brynjolfsson, E., Rock, D. and Syverson, C., 2017. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics (No. w24001). National Bureau of Economic Research.

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy?. In AEA Papers and Proceedings (Vol. 108, pp. 43-47). https://www.aeaweb.org/articles?id=10.1257/pandp.20181019.

Buchner, G. A., Stepputat, K. J., Zimmermann, A. W., & Schomäcker, R. (2019). Specifying technology readiness levels for the chemical industry. Industrial & Engineering Chemistry Research, 58(17), 6957-6969.

Chakroborty, P. and Das, A., 2017. Principles of transportation engineering. PHI Learning Pvt. Ltd.

Charalambous, G., Fletcher, S. R., & Webb, P. (2017). The development of a Human Factors Readiness Level tool for implementing industrial human-robot collaboration. The International Journal of Advanced Manufacturing Technology, 91(5-8), 2465-2475.

Chen, H.H., Gou, L., Zhang, X. and Giles, C.L., 2011, June. Collabseer: a search engine for collaboration discovery. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (pp. 231-240).

Chen, H.H., Ororbia, I.I., Alexander, G. and Giles, C.L., 2015. ExpertSeer: A keyphrase based expert recommender for digital libraries. arXiv preprint arXiv:1511.02058.

Chowdhury, S.R., Rodríguez, C., Daniel, F. and Casati, F., 2010, December. Wisdom-aware computing: on the interactive recommendation of composition knowledge. In International Conference on Service-Oriented Computing (pp. 144-155). Springer, Berlin, Heidelberg.

Clark, A., Fox, C., & Lappin, S. (Eds.). (2013). The handbook of computational linguistics and natural language processing. John Wiley & Sons.

Covington, P., Adams, J. and Sargin, E., 2016, September. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems (pp. 191-198).

Crabb, G. (1823). Universal Technological Dictionary: Or, Familiar Explanations of the Terms Used in All Arts and Sciences (Vol. 1). Baldwin, Cradock, and Joy.

Cypher, A. (1993), Watch What I Do: Programming by Demonstration, Daniel C. Halbert, MIT Press.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B. and Sampath, D., 2010, September. The YouTube video recommendation system. In Proceedings of the fourth ACM conference on Recommender systems (pp. 293-296).

Drexler, K. Eric. (2019). Reframing Superintelligence, Technical Report, Future of Humanity Institute, University of Oxford, Oxford, UK.

Dymova, L., Sevastianov, P. and Kaczmarek, K., 2012. A stock trading expert system based on the rule-base evidential reasoning using Level 2 Quotes. Expert Systems with Applications, 39(8), pp.7150-7157.

Elmahmudi, A. and Ugail, H., 2019. Deep face recognition using imperfect facial data. Future Generation Computer Systems, 99, pp.213-225.

Estevez, D., Victores, J. G., Fernandez-Fernandez, R., & Balaguer, C. (2020). Enabling garment-agnostic laundry tasks for a Robot Household Companion. Robotics and Autonomous Systems, 123, 103330.

European Commission Report (2018), Digital Transformation Monitor: The rise of Virtual Personal Assistants, https://ec.europa.eu/growth/tools-databases/dem/monitor/content/rise-virtual-personal-assistants

Ekstrand, M. D., Ludwig, M., Konstan, J. A., & Riedl, J. T. (2011, October). Rethinking the recommender research ecosystem: reproducibility, openness, and LensKit. In Proceedings of the fifth ACM conference on Recommender systems (pp. 133-140).

Felfernig, A., Isak, K., Szabo, K. and Zachar, P., 2007, July. The VITA financial services sales support environment. In Proceedings of the national conference on artificial intelligence (Vol. 22, No. 2, p. 1692). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Feo, T.A. and Bard, J.F., 1989. Flight scheduling and maintenance base planning. Management Science, 35(12), pp.1415-1432.

Ferri-Ramírez, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2001, March). Incremental learning of functional logic programs. In International Symposium on Functional and Logic Programming (pp. 233-247). Springer, Berlin, Heidelberg.

Ferrucci, D. A. (2012). Introduction to "this is watson". IBM Journal of Research and Development, 56(3.4), 1-1.

Flach, P. (2012). Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press.

Franz, A., & Milch, B. (2002). Searching the web by voice. In Proceedings of the 19th international conference on Computational linguistics-Volume 2 (pp. 1-5). Association for Computational Linguistics.

Fuchs, J., Heller, I., Tolpilsky, M. and Inbar, M., 1999. CaDet, a computer-based clinical decision support system for early cancer detection. Cancer detection and prevention, 23(1), p.78.

Gadepally, V., Goodwin, J., Kepner, J., Reuther, A., Reynolds, H., Samsi, S., Su, J. and Martinez, D., 2019. AI Enabling Technologies: A Survey. arXiv preprint arXiv:1905.03592.

Galbally, J., Ferrara, P., Haraksim, R., Psyllos, A., & Beslay, L. (2019). Study on Face Identification Technology for its Implementation in the Schengen Information System. EUR 29808 EN, Publication Office of the European Union, Luxemburg, 2019, ISBN 978-92-76-08843-1, doi:10.2760/661464, JRC116530.

Gavish, B., Schweitzer, P. and Shlifer, E., 1978. Assigning buses to schedules in a metropolitan area. Computers & Operations Research, 5(2), pp.129-138.

Ghoseiri, K., Szidarovszky, F. and Asgharpour, M.J., 2004. A multi-objective train scheduling model and solution. Transportation research part B: Methodological, 38(10), pp.927-952.

Gibson, C. et al, VAX 9000 SERIES, Digital Technical Journal of Digital Equipment Corporation, Volume 2, Number 4, Fall 1990, pp. 118-129.

Gold, B., Morgan, N., & Ellis, D. (2011). Speech and audio signal processing: processing and perception of speech and music. John Wiley & Sons.

Goldberg, Y. (2017). Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies, 10(1), 1-309.

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS), 6(4), 1-19.

Gonçalves, R. and Dorneles, C.F., 2019. Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues. ACM Computing Surveys (CSUR), 52(5), pp.1-30.

Grace, Katja, et al. (2018). When will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research 62: 729-754.

Granell, E., Romero, V., & Martínez-Hinarejos, C. D. (2019). Image–speech combination for interactive computer assisted transcription of handwritten documents. Computer Vision and Image Understanding, 180, 74-83.

Grbovic, M. and Cheng, H., 2018, July. Real-time personalization using embeddings for search ranking at airbnb. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 311-320).

Grother, P., Ngan, M. and Hanaoka, K., 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology.

Gruetzemacher, R., & Whittlestone, J. (2019). Defining and Unpacking Transformative AI. arXiv preprint arXiv:1912.00747.

Gruetzemacher, Ross. (2019). A Holistic Framework for Forecasting Transformative AI. Big Data and Cognitive Computing 3(3): 35.

Gulwani, S., Harris, W. R., & Singh, R. (2012). Spreadsheet data manipulation using examples. Communications of the ACM, 55(8), 97-105.

Gulwani, S., Hernández-Orallo, J., Kitzelmann, E., Muggleton, S. H., Schmid, U., & Zorn, B. (2015). Inductive programming meets the real world. Communications of the ACM, 58(11), 90-99.

Heras, S., De la Prieta, F., Julian, V., Rodríguez, S., Botti, V., Bajo, J., & Corchado, J. M. (2012). Agreement technologies and their use in cloud computing environments. Progress in Artificial Intelligence, 1(4), 277-290.

Harb, J., & Precup, D. (2017). Investigating recurrence and eligibility traces in deep Q-networks. arXiv preprint arXiv:1704.05495.

Hernández-Orallo, J. (2017). The measure of all minds: evaluating natural and artificial intelligence. Cambridge University Press.

Hernández-Orallo, J., Baroni, M., Bieger, J., Chmait, N., Dowe, D.L., Hofmann, K., Martínez-Plumed, F., Strannegård, C. and Thórisson, K.R., (2017). A new AI evaluation cosmos: Ready to play the game?. AI Magazine, 38(3), pp.66-69.

Hernández-Orallo, J.; Martínez-Plumed, F.; Avin, S.; Whittlestone, J. and O h'Eigeartaigh, S. (2020) "AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues" European Conference on Artificial Intelligence.

Hindriks, K., Jonker, C., & Tykhonov, D. (2008). Avoiding approximation errors in multi-issue negotiation with issue dependencies. In Proc. of The 1st International Workshop on Agent-based Complex Automated Negotiations (ACAN 2008) (pp. 1347-1352).

Hoffer, E.P., Feldman, M.J., Kim, R.J., Famiglietti, K.T. and Barnett, G.O., 2005. DXplain: patterns of use of a mature expert system. In AMIA Annual Symposium Proceedings (Vol. 2005, p. 321). American Medical Informatics Association.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4).

Howe, M., 2009. Pandora's Music Recommender. A Case Study, I, pp.1-6.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.

Hu, W., & Bolivar, A. (2008). Online auctions efficiency: a survey of ebay auctions. In *Proceedings of the 17th international conference on World Wide Web* (pp. 925-934).

Hutchins, W.J. and Somers, H.L., 1992. An introduction to machine translation (Vol. 362). London: Academic Press.

Ingolotti, L., Barber, F., Tormos, P., Lova, A., Salido, M.A. and Abril, M., 2004, November. An efficient method to schedule new trains on a heavily loaded railway network. In Ibero-American Conference on Artificial Intelligence (pp. 164-173). Springer, Berlin, Heidelberg.

Janssen, M. (Ed.). (2002). Complexity and ecosystem management: the theory and practice of multi-agent systems. Edward Elgar Publishing.

Jayaraman, V. and Srivastava, R., 1996. Expert systems in production and operations management. International Journal of Operations & Production Management.

Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Sierra, C., & Wooldridge, M. (2001). Automated negotiation: prospects, methods and challenges. International Journal of Group Decision and Negotiation, 10(2), 199-215.

Johansen, E.S., 2018. Personalized Content Creation using Recommendation Systems (Master's thesis, The University of Bergen).

Jonker, C. M., Hindriks, K. V., Wiggers, P., & Broekens, J. (2012). Negotiating agents. AI Magazine, 33(3), 79-79.

Juang, B.H. and Rabiner, L.R., 2005. Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, p.67.

Kang, W. C., Fang, C., Wang, Z., & McAuley, J. (2017). Visually-aware fashion recommendation and design with generative image models. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 207-216). IEEE.

Kang, W. C., Kim, E., Leskovec, J., Rosenberg, C., & McAuley, J. (2019). Complete the Look: Scene-based Complementary Product Recommendation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10532-10541).

Kautz, H. and Walser, J.P., 2000. Integer optimization models of AI planning problems. The Knowledge Engineering Review, 15(1), pp.101-117.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. sensors, 18(2), 401.

König, M., & Neumayr, L. (2017). Users' resistance towards radical innovations: The case of the self-driving car. Transportation research part F: traffic psychology and behaviour, 44, 42-52.

Konstan, J. A., & Adomavicius, G. (2013, October). Toward identification and adoption of best practices in algorithmic recommender systems research. In Proceedings of the international workshop on Reproducibility and replication in recommender systems evaluation (pp. 23-28).

Krasadakis, G. (2016) Artificial intelligence negotiation agent, File by Microsoft, US20170287038A1, United States Patent.

Kubrick, S., Clarke, A.C. (1968) Screenplay for "2001: A Space Odyssey", Stanley Kubrick Productions

Kumar, S., & Gupta, M. D. (2019). c+GAN: Complementary Fashion Item Recommendation. arXiv preprint arXiv:1906.05596.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. Penguin.

Langley, P., 1996. Elements of machine learning. Morgan Kaufmann.

Lavrenko, V., Rath, T. M., & Manmatha, R. (2004, January). Holistic word recognition for handwritten historical documents. In First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings. (pp. 278-287). IEEE.

Leonhardt, J., Anand, A. and Khosla, M., 2018, April. User Fairness in Recommender Systems. In Companion Proceedings of the The Web Conference 2018 (pp. 101-102).

Levesque, H., Davis, E., & Morgenstern, L. (2012, May). The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.

Li, X., & Da, F. (2012). Efficient 3D face recognition handling facial expression and hair occlusion. Image and Vision Computing, 30(9), 668-679.

Lieberman, H. (2001), Your Wish is My Command: Programming By Example, Ben Shneiderman, Morgan Kaufmann.

Lin, R., Oshrat, Y., & Kraus, S. (2012). Automated agents that proficiently negotiate with people: Can we keep people out of the evaluation loop. In New Trends in Agent-Based Complex Automated Negotiations (pp. 57-80). Springer, Berlin, Heidelberg.

Lin, Y., Ren, P., Chen, Z., Ren, Z., Ma, J., & De Rijke, M. (2019). Explainable outfit recommendation with joint outfit matching and comment generation. IEEE Transactions on Knowledge and Data Engineering.

Linden, A., & Fenn, J. (2003). Understanding Gartner's hype cycles. Strategic Analysis Report Nº R-20-1971. Gartner, Inc, 88.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46-60.

Mansanet, J., Albiol, A., & Paredes, R. (2016). Local deep neural networks for gender recognition. Pattern Recognition Letters, 70, 80-86.

Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

Martínez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., hÉigeartaigh, S. Ó., & Hernández-Orallo, J. (2018a). Accounting for the neglected dimensions of ai progress. arXiv preprint arXiv:1806.00610.

Martínez-Plumed, F., Loe, B. S., Flach, P., O hEigeartaigh, S., Vold, K., & Hernández-Orallo, J. (2018b). The facets of artificial intelligence: a framework to track the evolution of AI. In International Joint Conference on Artificial Intelligence, IJCAI (pp. 5180-5187).

Martínez-Plumed, F., Tolan, S., Pesole, A., Hernández-Orallo, J., Fernández-Macías, E., & Gómez, E. (2020). Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), February 7-8, 2020, New York, NY, USA. https://dl.acm.org/doi/abs/10.1145/3375627.3375831

Martínez-Plumed, F., Hernández-Orallo, J., Gómez, E., (2020a). Tracking AI: The Capability is (Not) Near: In 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain. IOS Press.

Martínez-Plumed, F., Hernández-Orallo, J., Gómez, E., (2020b). Tracking the Evolution of AI: The AIcollaboratory. In 1st International Workshop: Evaluating Progress in Artificial Intelligence (EPAI 2020), Spain.

Martínez Plumed, F., Hernández-Orallo, J., Gómez, E., (2020c) AI Watch: Methodology to Monitor the Evolution of AI Technologies. Publications Office of the European Union, Seville, 2020, ISBN 978-92-76-17153-9 (online), doi:10.2760/643950 (online), JRC120090, 2020.

McBurney, P., & Parsons, S. (2002). Games that agents play: A formal framework for dialogues between autonomous agents. Journal of logic, language and information, 11(3), 315-334.

Meng, Q., Wang, S., Andersson, H. & Thun, K., 2014. Containership Routing and Scheduling in Liner Shipping: Overview and Future Research Directions. Transportation Science, 48(2), pp. 265-280.

Miller, N. E., & Dollard, J. (1941). Social learning and imitation.

Miller, S., Van Den Berg, J., Fritz, M., Darrell, T., Goldberg, K., & Abbeel, P. (2012). A geometric approach to robotic laundry folding. The International Journal of Robotics Research, 31(2), 249-267.

Miozzo, M., Dewick, P., & Green, K. (2005). Globalisation and the environment: the long-term effects of technology on the international division of labour and energy demand. Futures, 37(6), 521-546.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B. and Krishnamurthy, J., 2018. Never-ending learning. Communications of the ACM, 61(5), pp.103-115.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937).

Muegge, U., 2006. Fully Automatic High Quality Machine Translation of Restricted Text-A Case Study. Translating and the computer, 28, p.15.

Muggleton, S. (Ed.). (1992). Inductive logic programming (No. 38). Morgan Kaufmann.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence (pp. 555-572). Springer.

Murphy, R. R. (2019). Introduction to AI robotics. 2nd Edition, MIT press.

Myerson, Roger B. (2013). Game theory. Harvard university press.

Narla, S. R. (2013). The evolution of connected vehicle technology: From smart drivers to smart cars to... self-driving cars. Ite Journal, 83(7), 22-26.

Olsson, R. (1995). Inductive functional programming using incremental program transformation. Artificial intelligence, 74(1), 55-81.

Ossowski, S. (Ed.). (2012). Agreement technologies (Vol. 8). Springer Science & Business Media.

Oyedotun, O. K., Olaniyi, E. O., & Khashman, A. (2015). Deep learning in character recognition considering pattern invariance constraints. International Journal of Intelligent Systems and Applications, 7(7), 1.

Park, U., Tong, Y., & Jain, A. K. (2010). Age-invariant face recognition. IEEE transactions on pattern analysis and machine intelligence, 32(5), 947-954.

Parsons, S. D., Gymtrasiewicz, P., & Wooldridge, M. (Eds.). (2012). Game theory and decision theory in agent-based systems. Springer.

PCMagazine, "20th Anniversary of the PC Survey Results." vol. 2004, 2001.

Pereira, R., Sousa, T. M., Pinto, T., Praça, I., Vale, Z., & Morais, H. (2014). Strategic Bidding for Electricity Markets Negotiation Using Support Vector Machines. In Trends in Practical Applications of Heterogeneous Multi-agent Systems. The PAAMS Collection (pp. 9-17). Springer.

Perez, J.B., Rodríguez, J.M.C., Mathieu, P., Campbell, A., Ortega, A., Adam, E., Navarro, E.M., Ahrndt, S., Moreno, M.N. and Julián, V. eds., (2014). Trends in Practical Applications of Heterogeneous Multi-agent Systems. The PAAMS Collection. Springer.

Petralia, S. (2017). Unravelling the Trail of a GPT: The Case of Electrical & Electronic Technologies from 1860 to 1930. Mimeo.

Polozov, O., & Gulwani, S. (2015, October). FlashMeta: a framework for inductive program synthesis. In Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (pp. 107-126).

Ptucha, R., Such, F. P., Pillai, S., Brockler, F., Singh, V., & Hutkowski, P. (2019). Intelligent character recognition using fully convolutional neural networks. Pattern Recognition, 88, 604-613.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O. and Jennings, N.R., (2019). Machine behaviour. Nature, 568(7753), pp.477-486.

Ramchurn, S. D., Vytelingum, P., Rogers, A., & Jennings, N. R. (2012). Putting the 'smarts' into the smart grid: a grand challenge for artificial intelligence. Communications of the ACM, 55(4), 86-97.

Rasmussen, A.N., 1990. The INCO Expert System Project: CLIPS in Shuttle Mission Control. First CLIPSConference, p.305.

Reinfrank, M., 1988. Reason maintenance systems. In Begründungsverwaltung (pp. 1-26). Springer, Berlin, Heidelberg.

Ricci, F., Rokach, L. and Shapira, B., 2011. Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35). Springer, Boston, MA.

Rodríguez-Aguilar, J. A., Martín, F. J., Noriega, P., Garcia, P., & Sierra, C. (1998). Towards a test-bed for trading agents in electronic auction markets. Aí Communications, 11(1), 5-19.

Rollinger, C.R. ed. (1991). Text understanding in LILOG: integrating computational linguistics and artificial intelligence: final report on the IBM Germany LILOG-Project. Berlin: Springer.

Rosenfeld, A., & Kraus, S. (2009). Modeling agents through bounded rationality theories. In Twenty-First International Joint Conference on Artificial Intelligence.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.H., Xiang, Y. and He, J., 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors, 19(8), p.1863.

Samoili, S., Cobo, M. L., Gomez, E., De Prato, G., Martinez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence (No. JRC118163). Joint Research Centre.

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., ... & de Does, J. (2013, September). tranScriptorium: a european project on handwritten text recognition. In Proceedings of the 2013 ACM symposium on Document engineering (pp. 227-228).

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. Co-design, 4(1), 5-18.

Sar Shalom, O., Koenigstein, N., Paquet, U. and Vanchinathan, H.P., 2016, April. Beyond collaborative filtering: The list recommendation problem. In Proceedings of the 25th international conference on world wide web (pp. 63-72).

Segler, M.H.S.; Preuss, M. and Waller, M.P. (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 555(7698):604.

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In Twelfth annual conference of the international speech communication association.

Schaal, S. (1997). Learning from demonstration. In Advances in neural information processing systems (pp. 1040-1046).

Schilling, M. A. (1998). Technological lockout: An integrative model of the economic and strategic factors driving technology success and failure. Academy of management review, 23(2), 267-284.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al., (2019). Mastering atari, go, chess and shogi by planning with a learned model. arXiv preprint arXiv:1911.08265.

Srihari, S. N., & Kuebert, E. J. (1997, August). Integration of hand-written address interpretation technology into the united states postal service remote computer reader system. In Proceedings of the Fourth International Conference on Document Analysis and Recognition (Vol. 2, pp. 892-896). IEEE.

Shoham, Y. and Leyton-Brown, K., 2008. Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press.

Shortliffe, E. ed., 2012. Computer-based medical consultations: MYCIN (Vol. 2). Elsevier.

Silberg, G., Wallace, R., Matuszak, G., Plessers, J., Brower, C., & Subramanian, D. (2012). Self-driving cars: The next revolution. White paper, KPMG LLP & Center of Automotive Research, 9(2), 132-146.

Silver, David; Huang, Aja; Maddison, Chris J.; Guez, Arthur; Sifre, Laurent; Driessche, George van den; Schrittwieser, Julian; Antonoglou, Ioannis; Panneershelvam, Veda; Lanctot, Marc; Dieleman, Sander; Grewe, Dominik; Nham, John; Kalchbrenner, Nal; Sutskever, Ilya; Lillicrap, Timothy; Leach, Madeleine; Kavukcuoglu, Koray; Graepel, Thore; Hassabis, Demis (28 January 2016). "Mastering the game of Go with deep neural networks and tree search". Nature. 529 (7587): 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al., (2017). Mastering the game of go without human knowledge. Nature, 550(7676), 354-359.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. and Lillicrap, T., 2017b. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815.

Simonsen, J., & Robertson, T. (Eds.). (2012). Routledge international handbook of participatory design. Routledge.

Szeliski, R. (2010). Computer vision: algorithms and applications. Springer.

Sofy, N., & Sarne, D. (2014). Effective deadlock resolution with self-interested partially-rational agents. Annals of Mathematics and Artificial Intelligence, 72(3-4), 225-266.

Spelda, P., & Stritecky, V. (2020). The future of human-artificial intelligence nexus and its environmental costs. Futures, 117, 102531.

Spyropoulos, C.D. (2000).AI planning and scheduling in the medical hospital environment. Artif. Intell. Med. 20, 2 (October 2000), 101–111.

Steele, Katie and Stefánsson, H. Orri (2016)., Decision Theory, The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.).

Steinert, M., & Leifer, L. (2010). Scrutinizing Gartner's hype cycle approach. In Picmet 2010 Technology Management for Global Economic Growth (pp. 1-13). IEEE.

Suri, T. (2011). Selection and comparative advantage in technology adoption. Econometrica, 79(1), 159-209.

Sutton, R. S., & Barto, A. G. (1998). Introduction to reinforcement learning (Vol. 135). Cambridge: MIT press.

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

Tashev, I., Seltzer, M., Ju, Y.C., Wang, Y.Y. and Acero, A., 2009. Commute UX: Voice enabled in-car infotainment system.

Toselli, A. H., Vidal, E., Puigcerver, J., & Noya-García, E. (2019). Probabilistic multi-word spotting in handwritten text images. Pattern Analysis and Applications, 22(1), 23-32.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). How reasonable are common-sense reasoning tasks: A case-study on the Winograd Schema Challenge and SWAG. arXiv preprint arXiv:1811.01778.

Tur, G., & De Mori, R. (2011). Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.

Veletsianos G (2007) Cognitive and affective benefits of an animated pedagogical agent: considering contextual relevance and aesthetics. J Educ Comput Res 36(4):373–377

Verderame, P.M., Elia, J.A., Li, J. and Floudas, C.A., 2010. Planning and scheduling under uncertainty: a review across multiple sectors. Industrial & engineering chemistry research, 49(9), pp.3993-4017.

Von Der Osten, F. B., Kirley, M., & Miller, T. (2017). The Minds of Many: Opponent Modeling in a Stochastic Game. In IJCAI (pp. 3845-3851).

Wagner, W.P., 2017. Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. Expert systems with applications, 76, pp.85-96.

Walker, N., Peng, Y. T., & Cakmak, M. (2019, July). Neural Semantic Parsing with Anonymization for Command Understanding in General-Purpose Service Robots. In Robot World Cup (pp. 337-350). Springer.

Wallia, C.J.S., 1994. Talking and listening to a Mac Quadra 840 AV. Technical Communication, 41(1), pp.130-131.

Wellman, M. P. (2011). Trading agents. Synthesis Lectures on Artificial Intelligence and Machine Learning, 5(3), 1-107.

Wilensky, R. (1987). The Berkeley UNIX consultant project. In Wissensbasierte Systeme (pp. 286-296). Springer, Berlin, Heidelberg.

Wooldridge, M. (2009). An introduction to multiagent systems. John Wiley & Sons.

Wolfram, S. (2009). Wolfram|Alpha. On the WWW. URL http://www.wolframalpha.com

Wu, L., Shah, S., Choi, S., Tiwari, M. and Posse, C., 2014, October. The Browsemaps: Collaborative Filtering at LinkedIn. In RSWeb@ RecSys.

Yang, W., Jin, L., Tao, D., Xie, Z., & Feng, Z. (2016). DropSample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition. Pattern Recognition, 58, 190-203.

Yehoshua Bar-Hillel (1964). Language and Information: Selected Essays on Their Theory and Application. Reading, MA: Addison-Wesley. pp. 174–179.

Yuan, A., Bai, G., Yang, P., Guo, Y., & Zhao, X. (2012, September). Handwritten English word recognition based on convolutional neural networks. In 2012 International Conference on Frontiers in Handwriting Recognition (pp. 207-212). IEEE.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), 1-38.

Appendix A: Technology Readiness Levels Rubric

In this appendix, we include more detail about each TRL in the form of a rubric, as has been used to assign the TRLs in this document. These extended descriptions have been adapted from some "TRL calculators"[103] [104], developed by the US Air Force Research Laboratory developed for assisting in the process of evaluating the TRL of project or product. Each entry below includes level, title, rubric question, description, and main characteristics.

TRL - 1 Basic principles observed - *Have basic principles been observed and reported?*

Lowest level of technology readiness. Research begins to be translated into applied research and development. Examples might include paper studies with the basic properties of a technology.

- "Back of envelope" environment
- Basic scientific principles observed
- Research hypothesis formulated
- Mathematical formulations of concepts that might be realisable in software
- Initial scientific observations reported in scientific journals, conference proceedings and technical reports

TRL - 2 Technology concept formulated - *Has a concept or application been formulated?*

Invention begins. Once basic principles are observed, practical applications can be invented. Applications are speculative and there may be no proof or detailed analysis to support the assumptions. Examples are limited to analytic studies.

- Desktop environment
- Paper studies show that application is feasible
- An apparent theoretical or empirical design solution identified
- Basic elements of technology have been identified
- Experiments performed with synthetic data
- Individual parts of the technology work (no real attempt at integration)
- Know what experiments you need to do (research approach)
- Analytical studies reported in scientific journals, conference proceedings and technical reports

TRL - 3 Experimental proof of concept - *Have analytical and experimental proof of concepts been demonstrated?*

Continued research and development efforts. This includes analytical studies and laboratory studies to physically validate analytical predictions of separate elements of the technology. Examples include components that are not yet integrated or representative.

- Academic environment

103 https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2003/systems/nolte2.pdf, https://faaco.faa.gov/index.cfm/attachment/download/100020.

104 US Air Force Research Laboratory "TRL Calculator" (for Excel): (http://aries.ucsd.edu/ARIES/MEETINGS/0712/Waganer/TRL%20Calc%20Ver%202_2.xls)

- Preliminary system performance characteristics and measures have been identified and estimate
- Outline of software algorithms available
- Laboratory experiments verify feasibility of application
- Metrics established
- Experiments carried out with small representative data sets
- Algorithms run on surrogate processor in a laboratory environment
- Existing software examined for possible reuse
- Limitations of presently available software assessed (analysis of current software completed)
- Scientific feasibility fully demonstrated
- Analysis of present state of the art shows that technology fills a need

TRL - 4 Technology validated in the laboratory - *Has a breadboard unit been demonstrated in a laboratory (controlled) environment?*

Basic technological components are integrated to establish that they will work together. This is relatively "low fidelity" compared to the eventual system. Examples include integration of “ad hoc” software and/or hardware in the laboratory.

- Controlled laboratory environment
- Individual components tested in laboratory or by supplier
- Formal system architecture development begins
- Overall system requirements for end user's application are known
- Analysis provides detailed knowledge of specific functions software needs to perform
- Technology demonstrates basic functionality in simplified environment
- Analysis of data requirements and formats completed
- Experiments with full scale problems and representative data sets
- Individual functions or modules demonstrated in a laboratory environment
- Some ad hoc integration of functions or modules demonstrates that they will work together
- Low fidelity technology “system” integration and engineering completed in a lab environment
- Functional work breakdown structure developed

TRL - 5 Technology validated in a relevant environment - *Has a breadboard unit been demonstrated in a relevant (typical; not necessarily stressing) environment?*

Fidelity and reliability is significantly increased. The basic technological components are integrated with reasonably realistic supporting elements so it can be tested in a simulated environment. Examples include “high fidelity” laboratory integration of components.

- Laboratory environment modified to approximate operational environment
- System interface requirements known
- System software architecture established
- Coding of individual functions/modules completed

- High fidelity lab integration of system completed, ready for test in realistic or simulated environment
- Individual functions tested to verify that they work
- Individual modules and functions tested for bugs
- Integration of modules/functions demonstrated in a laboratory environment

TRL - 6 Technology demonstrated in a relevant environment - *Has a prototype been demonstrated in a relevant environment, on the target or surrogate platform?*

Representative model or prototype system, which is well beyond that of TRL 5, is tested in a relevant environment. This represents a major step up in the demonstrated readiness of a technology. Examples include testing a prototype in a high fidelity laboratory environment or in a simulated operational environment.

- Operating environment for eventual system known
- Representative model / prototype tested in high-fidelity lab / simulated operational environment
- Realistic environment outside the lab, but not the eventual operating environment
- Prototype implementation includes functionality to handle large scale realistic problems
- Algorithms partially integrated with existing hardware / software systems
- Individual modules tested to verify that the module components (functions) work together
- Representative software system or prototype demonstrated in a laboratory environment
- Laboratory system is high-fidelity functional prototype of operational system
- Limited software documentation available
- Engineering feasibility fully demonstrated

TRL - 7 System prototype demonstration in operational environment - *Has a prototype unit been demonstrated in the operational environment?*

Represents a major step up from TRL 6, requiring demonstration of an actual system prototype in an operational environment. Examples include testing the prototype in operational testing platforms (e.g., a real-world clinical setting, a vehicle, etc.) .

- Each system/software interface tested individually under stressed and anomalous conditions
- Algorithms run on processor(s) in operating environment
- Operational environment, but not the eventual platform
- Most functionality available for demonstration in simulated operational environment
- Operational/flight testing of laboratory system in representational environment
- Fully integrated prototype demonstrated in actual or simulated operational environment
- System prototype successfully tested in a field environment

TRL - 8 System complete and qualified - *Has the system/development unit been qualified but not operationally demonstrated?*

Technology proved to work in its final form and under expected conditions. In most cases, this TRL represents the end of true system development. Examples include developmental test and evaluation of the system to determine if the requirements and specifications are fulfilled.

- Final architecture diagrams have been submitted
- Software thoroughly debugged
- All functionality demonstrated in simulated operational environment
- Certifications and licenses given by regulators

TRL - 9 Actual system proven in operational environment - *Has the system/development unit been demonstrated on an operational environment?*

Actual application of the technology in its final form and under mission conditions, such as those encountered in operational test and evaluation. Examples include using the system under operational conditions.

- Operational concept has been implemented successfully
- System has been installed and deployed.
- Actual system fully demonstrated

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30401-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/15025

ISBN 978-92-76-22987-2