Telematics and Informatics 58 (2021) 101525

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Futures of artificial intelligence through technology readiness levels

Fernando Martínez-Plumed [a,b,*], Emilia Gómez [a], José Hernández-Orallo [b]

[a] Joint Research Centre, European Commission, Spain
[b] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Spain

ARTICLE INFO

Keywords:
AI technologies
Generality
Capabilities
Technology readiness
TRLs

ABSTRACT

Artificial Intelligence (AI) offers the potential to transform our lives in radical ways. However, the main unanswered questions about this foreseen transformation are its *depth*, *breadth* and *timelines*. To answer them, not only do we lack the tools to determine what achievements will be attained in the near future, but we even ignore what various technologies in present-day AI are capable of. Many so-called breakthroughs in AI are associated with highly-cited research papers or good performance in some particular benchmarks. However, 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 exemplar-based methodology to categorise and assess several AI technologies, by mapping them onto Technology Readiness Levels (TRL) (representing their *depth* in 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. We then introduce a generality dimension, which represents increasing layers of *breadth* of the technology. These two dimensions lead to the new *readiness-vs-generality charts*, which show that higher TRLs are achievable for low-generality technologies, focusing on narrow or specific abilities, while high 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 exemplars. Finally, we show how the *timelines* of several AI technology exemplars at different generality layers can help forecast some short-term and mid-term trends for AI.

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. An important space AI systems are also taking over comes from the opportunities of an increasingly more digitised and 'datafied' (Hintz et al., 2018) 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.

* Corresponding author.
E-mail addresses: Fernando.MARTINEZ-PLUMED@ec.europa.eu (F. Martínez-Plumed), Emilia.GOMEZ-GUTIERREZ@ec.europa.eu (E. Gómez), jorallo@dsic.upv.es (J. Hernández-Orallo).

https://doi.org/10.1016/j.tele.2020.101525
Received 16 June 2020; Received in revised form 8 September 2020; Accepted 4 November 2020
Available online 23 November 2020
0736-5853/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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., 2017; Garcia-Murillo et al., 2018; Martinez-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. The latter includes some usually neglected dimensions, in addition to performance, such as data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, development time, etc., apart from monetary costs (Martinez-Plumed et al., 2018). 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 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.

The key question is not whether AI is envisaged to work or already 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 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 needed 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 (Martinez-Plumed et al., 2018), but also what are the current level of maturity and readiness of newly introduced technologies.

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 (Mankins et al., 1995). We first interpret the nine TRLs (introduced by NASA Mankins et al., 1995) in the context of AI, and then we apply them systematically to different categories in AI, by choosing illustrative exemplars in each category. In this regard, we potentially consider all AI technologies, as defined by the areas that are usually associated with the discipline; this is one of the main reasons why we enumerate a list of AI categories that correspond to subfields in AI. We do not use other characterisations of AI systems, such as those that act rationally or act like a human (Russell and Norvig, 2020), which may be more restrictive.

For the readiness assessment, we introduce new bidimensional plots, which we call readiness-vs-generality charts, as a trade-off between how broad a technology, its generality, 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. When it comes to 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 work. 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.

The rest of the paper is organised as follows. Section 2 reviews the notion of technology readiness level, borrowed from NASA. Section 3 presents the key methodology: we first give the contours of what an AI technology is in particular, by assigning it 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 exemplar 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 exemplar of AI technology, the virtual assistants, covering several categories. Section 6 closes the paper with an analysis of future trends in AI according to the evolution of TRL for different layers of generality. An appendix follows after the references, including a rubric for the TRLs and the description of the AI categories introduced.

2. Technology readiness levels

Defined and used on-and-off for NASA space technology planning for many years, the Technology Readiness Levels (TRL) (Mankins et al., 1995) conform a systematic measurement system 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 current TRL scale 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).[1] In Table 1 we summarise these nine levels as we use them in this work (see the rubric for further details in A).

We may group the nine TRLs in terms of the environment in which the project is developed, as shown in column "Environment" in Table 1. 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.

Given the type of research, technological development and innovation being addressed, it should be noted that the first three levels would address the most basic technological research involving, mosty, 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. In these levels, the deployment or implementation on a large scale would come. We show these concepts in the column "Goal" of Table 1.

In case we want to assess the life cycle of the technology to be developed in terms of artefacts produced, 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–RL 7) until its validation. Finally, we would have its placing in the market and deployment (TRL 8–RL 9). This is shown in Table 1, column "Artefacts".

Finally, one should also consider the results that each of the maturity levels would bring. We show this in Table 1 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 technology status, 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,[2] airports,[3] 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 (Buchner et al., 2019). Regarding AI, some authors have also assessed innovation projects in manufacturing and logistics in terms of specification and use of AI technologies (Eljasik-Swoboda et al., 2019; Ellefsen et al., 2019), being limited to particular industrial domains. Some other have also proposed modifications on the original TRL scale for the development of machine learning projects (Lavin et al., 2020). Here, we stick to the original scale and will remain as general as possible. Instead of giving a prescriptive refinement of each level for AI, we use the standard rubrics (see A) complemented with an exemplar-based approach, as we explain in the following section.

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 comprehensive 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 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 breadth layers (e.g., voice recognition with different degrees of versatility and robustness can be mapped to different TRLs). Consequently, we need to assign readiness levels according to different layers 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 and open-ended. In order to represent this, 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

[1] 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.

[2] See, e.g., https://press.siemens.com/global/en/pressrelease/europes-longest-driverless-subway-barcelona-goes-operation.

[3] See, e.g., http://www.mediacentre.gatwickairport.com/press-releases/2018/18_03_16_autonomous_vehicles.aspxn.

Table 1
Summary of Technology Readiness Levels (TRLs) according to several characteristics.

Environment	Goal	Product/ Evaluation	Outputs	TRL	Description
			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
		Proof of concept	Measurement of parameters in the laboratory	TRL 3	Experimental proof of concept
Laboratory	Research		Results of tests carried out in the laboratory.	TRL 4	Technology validated in lab
			Components validated in a relevant environment.	TRL 5	Technology validated in relevant environment
Simulation	Development		Results of tests carried out at the prototype in a relevant environment.	TRL 6	Technology demonstrated in relevant environment
		Prototype	Result of the prototype level tests carried out in the operating environment.	TRL 7	System prototype demonstration in operational environment
		Commercial (certified) product	Results of system tests in final configuration.	TRL 8	System complete and qualified
Operational	Implementation	Deployed product	Final reports in working condition or actual mission.	TRL 9	Actual system proven in operational environment

used in the resolution of concrete problems. 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 the 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 the AI system itself. For instance, if a generative model (Salakhutdinov, 2015) 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,[4] so we prefer the alternative term "transformative technology". Consequently, we refer to AI technologies as potentially transformative (Gruetzemacher et al., 2019). Clearly, a technology is transformative as much as it reaches critical elements of society or becomes mainstream. Critical elements could cover specialised domains, such as manufacturing robotics, and mainstream could also be used for a given domain (e.g., transforming technology for manufacturing). 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 technology derived from the research and development in any subareas of AI. Of course, this depends on how well the contours of AI are delimited (Martinez-Plumed et al., 2018). 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. TRL assessment in AI: readiness-vs-generality charts

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

[4] 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.

problem to be solved. In the case of AI, 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. In this regard, we introduce a categorisation of those main fields of research in AI and what sort of relevant technologies they comprise. AI as a field is thus categorised into seven main AI capabilities: knowledge representation and reasoning, learning, communication, planning, perception, physical interaction and social abilities. The detailed organisation of the AI categories and technologies evaluated, including details of the subsections covering them, can be found in Table 2 (full descriptions of the categories in B). This categorisation is inspired by the operational definition of AI adopted in the context of the AI Watch initiative (Samoili et al., 2020) from the European Commission (EC).

The above categorisation is sufficiently comprehensive of the areas of AI (and the cognitive capabilities that are being developed by the discipline) 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 include an example in the discussion section), but we do not expect to have technologies that we cannot be assigned to any category. In the following section, we identify relevant technologies and assign them to some of these categories, as summarised in Table 2. Before that, however, we need to better understand the scope of each AI technology.

In order to assess the readiness levels of AI technologies, we face an important dilemma between the readiness level and the autonomous performance for open-ended 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. 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 situation. Of course, one can specify all these specific technologies separately, and identify different niches, as we see in Fig. 1 (left). These instances of the technology 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. Fig. 2.

A different way of organising this space is an onion model, as we illustrate in Fig. 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. Some of these hierarchies we can think of might well be partial orders. But if we are able to select one hierarchy that is a total order, where each instance is a subset of a more general instance, we will be able to talk about different layers of generality of the same technology. This ensures that in order to reach layer n, no task or situation of layer $n-1$ is left out. In other words, progress, when following layers, is cumulative. This is also more meaningful since the higher generality is, the lower the expected readiness level becomes. And the other way around. This will help understand the common situation where a technology is stuck at TRL 7, but a reduction of the expectation in generality (finding a successful specific niche) can lead to a product with TRL 9. Robotic vacuum cleaners are a good example of this. By limiting the task (only floor vacuuming) and the range (simple trajectories), the system is specialised with the successful outcome that these devices are found in many homes today (TRL 9).

Another advantage of the onion model is that the total order allows the layers to be considered as an ordinal magnitude that can be represented in a Cartesian space along with another ordinal magnitude, the TRL. Therefore, we use two-dimensional plots[5] (readiness-vs-generality charts) with the degree of generality expected on the x-axis versus the readiness level (the TRLs) on the y-axis. Figure 2 shows one figurative plot.

As we move right in this plot, we have a system, procedure or component, a technology, that is more generally applicable. As we go up in this plot, we reach products that are more ready to be used in the real world. This plot can be applied to any technology (e.g., a pencil is both general and ready, as a writing device), but determining the precise location in the plot 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 make the technology become really transformative. For instance, a robotic vacuum cleaner moving around our floors has reached TRL 9, but has not transformed society. A fully-fledged robotic cleaner which much more advanced capabilities would do, 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 to be done to create—probably new—AI technologies that reach higher layers of capabilities. A flat curve that reaches medium TRL for a wide layer of capabilities may mean that reaching the commercial product or general use may depend on issues related to safety, usability or expectations about the technology, and not that much about rethinking the foundations of the technology. Nevertheless, looking case by case may lead to different interpretations. This is precisely what we do next; we present these charts for an illustrative set of AI technologies.

Just before this analysis, we need to fix some criteria to determine the x-axis and the precise location of each point in the chart. Unfortunately, there is no standard scale for these layers that could be used for all technologies. For each technology, generality layers are established by looking at the historical evolution of the technology, and this dictates that some layers (e.g., word recognition for reduced vocabularies) did not get traction, while others (e.g., speech recognition for reduced vocabularies) can be identified as an early milestone in this technology. In all technologies we can identify different dimensions that can help us shape the layers. For instance,

[5] 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.

Table 2
AI categories and the sample of representative technologies evaluated for each of them.

Category	Technology
(§4.1) Knowledge Representation & Reasoning	(§4.1.1) Knowledge Inference Engines
(§4.2) Learning	(§4.2.1) Recommender Systems
	(§4.2.2) Apprentices by Demonstration
(§4.3) Communication	(§4.3.1) Machine Translation
	(§4.3.2) Speech Recognition
(§4.4) Perception	(§4.4.1) Facial Recognition
	(§4.4.2) Text Recognition
(§4.5) Planning	(§4.5.1) Transport Scheduling Systems
(§4.6) Physical Interaction (Robotics)	(§4.6.1) Self-Driving Cars
	(§4.6.2) Home Cleaning Robots
(§4.7) Social & Collaborative Intelligence	(§4.7.1) Negotiation Agents
(§5.1) Integrating Technology	(§5.1) Virtual Assistants

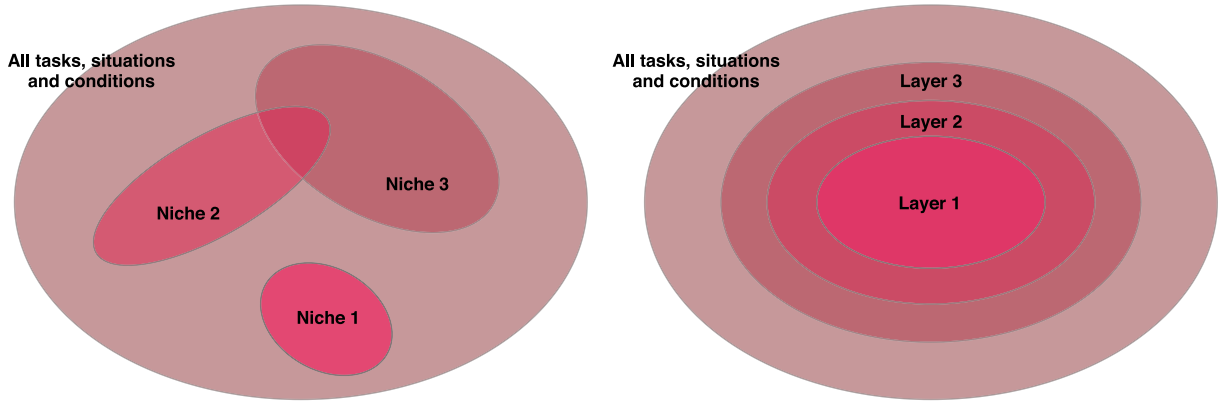


Fig. 1. Left: we can consider different instances of the technology covering different niches, each of them solving a set of tasks, situations and conditions that are not hierarchically related to each other. Each "niche" achieves a different TRL (shown with different darkness levels), which is mostly 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 "layers of generality", as they are broader than the previous ones, containing them.

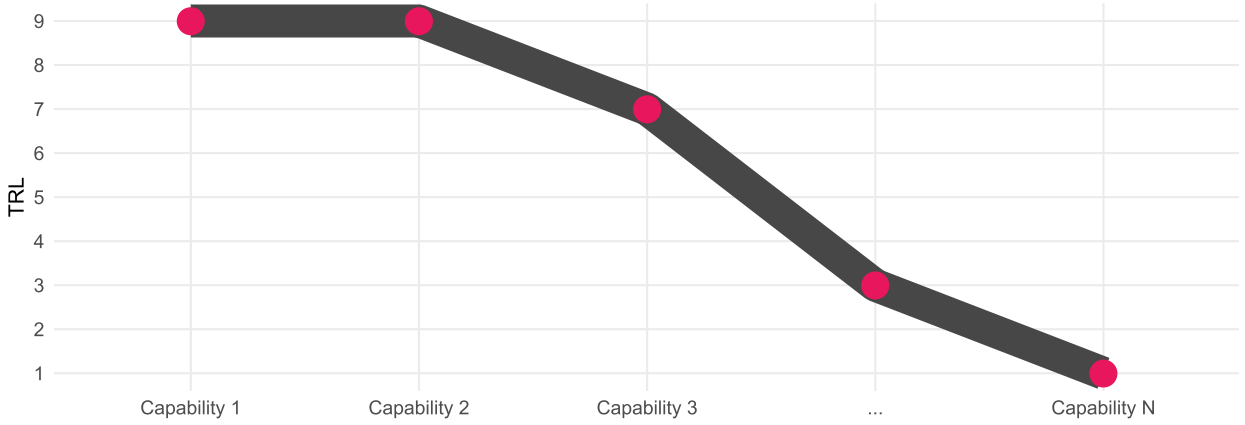


Fig. 2. Readiness-vs-generality charts showing the different layers of capabilities (more specific to more general) on the *x*-axis and TRL levels on the *y*-axis. Typically, the points will form a "curve" with decreasing shape. Progress towards really transformative AI will be achieved by moving this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

there are two dimensions that are commonly involved in the definition of the generality layers: 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 generality. Sometimes, the first dimension unfolds into two (e.g., speech recognition: size of vocabulary and number of languages) or more. In our onion model, we

merge all of them into one single ordinal dimension. There are of course cases where more challenging versions of the technology cannot easily be reduced to this unidimensional scale, but we will still try to find a scale of layers that go from lower to higher generality. In a few cases, we will simply reuse some scales that have been used in the past of that particular technology, or even used as standards, as happens with machine translation and self-driving cars.

Once the space is defined by the generality layers and the nine readiness levels, we locate the points in the following way. First, we follow the rubric in A. Second, for each layer, we identify the highest TRL according to the best player (research team or company) as per 2020. The reasoning behind this choice—e.g., in front of choosing an average—relies on AI technologies being digital, which means that they are quickly imitated by other players. Indeed, possible slowing factors such as patents are usually compensated by open publication platforms such as arXiv[6] and open software platforms such as github,[7] not to mention the common transfer of key people in AI between academia, industry and especially key tech giants (Kwok, 2019), bringing the technology with them, and spreading it to other players.

Finally, even using this generality-vs-readiness space and the rubric in A, 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 our own definitions for the layer of generality. It may also be the result of ours not being experts in each and every subarea in AI (although some detachment may also be positive). In other cases, this may be originated by the revision done by a panel of experts (see the list in the acknowledgements), which occasionally had some minor discrepancies. For all these cases we use vertical error bars in the charts. We hope that some of our assessments could be replicated by other groups of experts, and these bars built as proper confidence bands from the variance of results from a wider population of experts.

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 examine the technologies under the seven categories presented above (and further explained in B), 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, knowledge inference engines, because of its historical relevance and representativeness of *systems that reason*.

4.1.1. Technology: knowledge inference engines

Knowledge inference engines, as introduced in the 1980s in the form of Expert Systems, is a traditional AI technology that humans can use to extend or complement their expertise. Knowledge inference engines 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). Knowledge inference engines systems are still fuelling many 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 and retrieved from document repositories or other sources such as the web or Wikipedia (Mitchell et al., 2018; Gonçalves and Dorneles, 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).

Because of the evolution of expectations and capabilities of this technology, the *x*-axis of Fig. 3 uses four different generality layers of knowledge inference engines, as described in the box on the right of the figure, from very narrow systems to broader ones, even using meta-cognition.

For the first layer, early academic expert 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 this sort of knowledge inference engines 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, 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 layer. Consequently, even if the term expert system is in disuse today, systems with these capabilities are still operating at TRL 9, as shown in Fig. 3.

The second layer represents a new level of expectations raised after the limitations of the 1980s. A new generation of knowledge inference engines 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; Nilashi et al., 2017), control/management/monitoring (Jayaraman and Srivastava, 1996), stock markets (Dymova et al., 2012), space (Rasmussen, 1990), etc. However, many of these systems do not meet the expectations of

[6] https://arxiv.org/.
[7] https://github.com/.

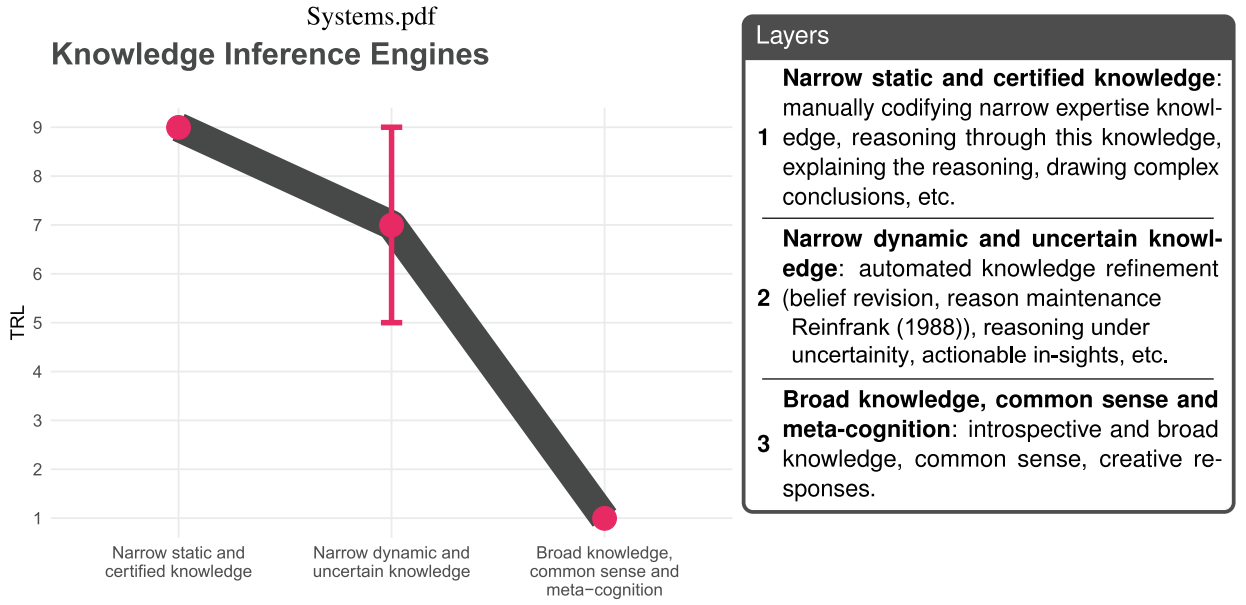


Fig. 3. Readiness-vs-generality chart for knowledge inference engines technology. While TRL 9 has clearly been reached for narrow systems with static and certified knowledge (early commercial systems and many expert systems still in place today), a very low TRL is estimated for knowledge inference engines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

robustness and self-maintenance completely, and some of the features of layer 2 are not fully autonomous (requiring important human maintenance). Because of this, we consider them more like market-ready research results 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 knowledge inference engines, also combining a number of different techniques for natural language processing, information retrieval, hypotheses generation, etc.

At the third layer, 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 commonsense reasoning has been shown even at a research level, and that is why we assign TRL 1 to this layer (if not at a fundamental research stage even before this level).

The schism between layers 2 and 3 (and the lack of progress in this schism in the past years) suggests there is still a long way to go until AI systems exhibit more human-like commonsense reasoning, being capable of acquiring knowledge and drawing conclusions, of a real domain expert.

Of course, expert systems is not the only technology in the knowledge representation and reasoning category. Automated theorem provers, SAT systems for solving boolean satisfiability problems, 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 layers.

4.2. Learning

Learning is probably the most distinctive capability of systems that adapt to their environment. Many AI systems are nowadays based on machine learning, creating models from training data that are finally deployed. However, they can be deployed without further learning (e.g., a classification model), or they can learn continually, as more evidence is given or as they interact with an environment. While the former type of learning (batch learning) is able to generalise to some unseen situations, it is usually brittle to concept shift or other variations of the problem or its distribution. It represents a *model* taken from a system that learned. The latter type of learning (continual learning) is more properly associated with what we would call a *system*that learns. Consequently, we want to consider AI 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 exemplars of AI technologies that represent *systems that learn*.

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 several 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 in 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.

Because of the evolution of expectations and capabilities of recommender systems technology, the *x*-axis of Fig. 4 uses four different generality layers described in the right box. For the first layer, 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 real-world applications, reaching TRL 9. For instance, we find the Pandora's Music Genome Project (Howe, 2009) or the Stitch Fix's fashion box[8] as examples of content-based recommender systems. Also, the engines used in Amazon, Netflix (Gomez-Uribe and Hunt, 2015), YouTube (Davidson et al., 2010), Spotify or Linkedin (Wu et al., 2014) were (at the beginning of its development) examples of collaborative filtering-based approaches.[9] 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., 2007).

For the second layer, more effective methods are currently being developed by looking at similarity beyond explicit feedback as well as latent attributes (e.g., by using matrix factorisation Kroenke and Spitzer (2002) or deep learning embeddings Zhang et al. (2019)) revealing relationships that have not yet been realised. Research behind this more advanced and flexible approaches has increased exponentially in the past recent years[10] with notable examples such as those from Zillow,[11] Netflix[12] and Airbnb (Grbovic and Cheng, 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 and Basilico, 2016), to make sure, for instance, 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 may also be a step beyond (Sar 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 behaviour. All this would correspond with the third layer 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 layer, we are including further innovations for recommendation systems such as recommending new items that do not exist and should be created to fill missing needs. Recommendation content generation is relatively new, but already includes proof-of-concept systems validated in the lab (TRL 2 to TRL 4) in areas such as automatic food menu generation (Bianchini et al., 2017), music generation (Johansen, 2018), simple fashion design (Kang et al., 2017; Kumar et al., 2019) or even artificially generated comments (Lin et al., 2019).

4.2.2. Technology: apprentices 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 and Dollard, 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, from 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 video game with a finite number of "action keys" is an example of this technology, or a spreadsheet automator that learns a simple program snippet to perform an operation. A full robotic 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 and Halbert, 1993) or programming by example

[8] https://algorithms-tour.stitchfix.com/.

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

[10] E.g., the leading international conference on recommendation systems (RecSys) started to organise regular workshops on deep learning in 2016.

[11] https://www.zillow.com/tech/embedding-similar-home-recommendation/.

[12] https://help.netflix.com/en/node/100639.

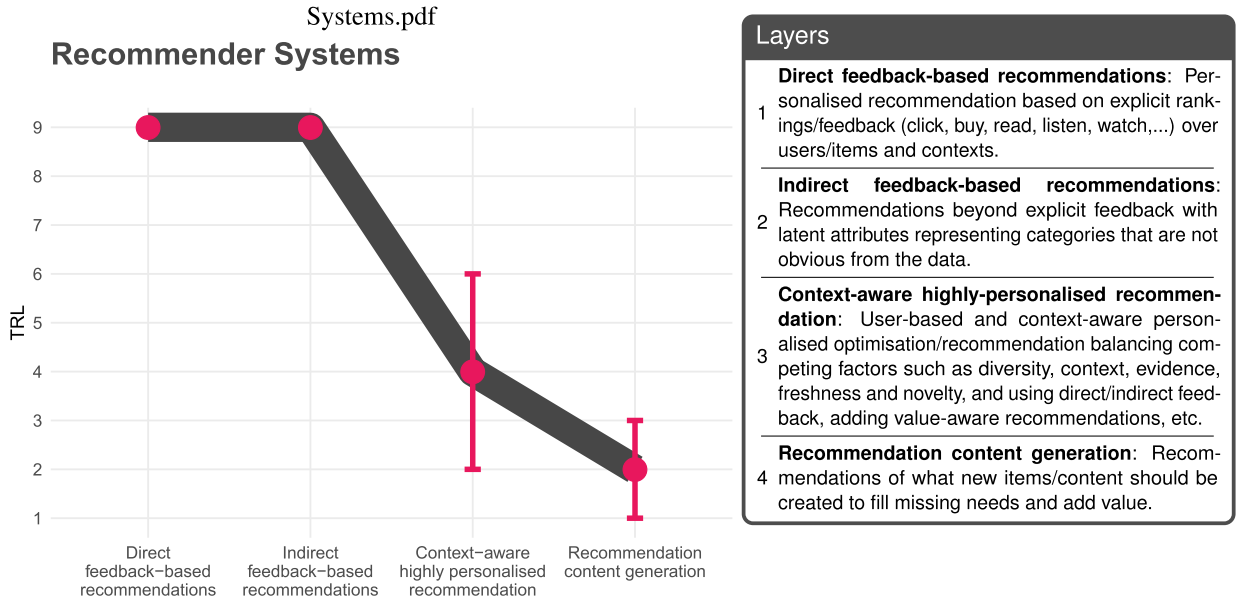


Fig. 4. Readiness-vs-generality chart for recommender engines technology. TRL 9 reached for those very well-known recommender systems based on user feedback and used in a variety of areas such as playlist generators for video and music services or product recommenders. Current developments going beyond 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 unfair personalisations using various dimensions of data. Finally, recommendation content generation would be a future direction in the field, with still little or no research nowadays.

(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 and Barto, 1998) or recorded interactions performed by a human or artificial expert (Silver et al., 2016; Mnih et al., 2016; Harb and Precup, 2017).

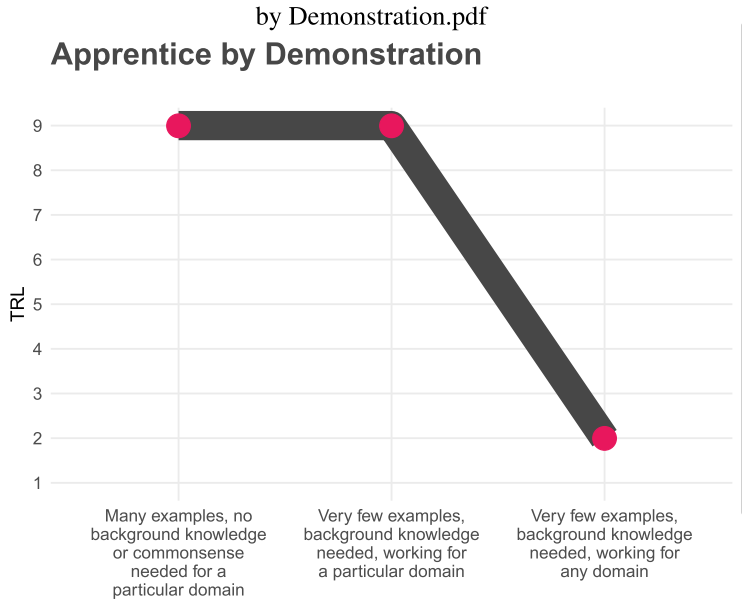
The *x*-axis of Fig. 5 uses three different generality layers, as described in the right box. For layer 1 there is evidence of the progress of deep reinforcement learning from traces. For instance, AlphaGo (Silver et al., 2016) was able to learn how to play go but used some hints from human traces. Similarly, many deep reinforcement learning algorithms use traces (Mnih et al., 2016; Harb and Precup, 2017). Because new variants of these algorithms are open source and already implemented,[13] with more modest resources than in the original paper, this puts us in TRL 9, at least for the 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 layer 2, the challenge comes from the limited number of examples. This is possible in humans 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 and Gulwani, 2015).

Finally, for layer 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 layer is still in its inception, even if there has been research for decades (Muggleton, 1992; Olsson, 1995; Ferri-Ramírez 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 layer would have a major impact in many daily tasks that are repetitive and would not need programming scripts or code snippets by hand. General apprentices by demonstration would have a transformative effect on the labour market, especially for programmers, among other professions. Because the challenge may depend on symbolic knowledge representations, and this has been explored for decades, we do not expect a breakthrough to high TRL 9 in the near-term future.

[13] See, e.g., https://github.com/openai/baselines.



Layers

1 **Many examples, no background knowledge or commonsense 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) with thousands of traces of humans/systems succeeding or failing at the task. The database records cases such as protocols, treatments, etc. Learning with traces is supposed to be more efficient than without them, or even necessary in some environments for which we lack a simulator.

2 **Very few examples, background knowledge needed, working for a particular domain**: When few examples are available, learning needs to rely on background knowledge. We assume that only one domain 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.

3 **Very few examples, background knowledge needed, working for any domain**: In this case we want the system to handle virtually any domain. This needs switching the background knowledge from one domain to another, or wide knowledge about different areas, so that the system can understand traces, videos, demos, etc., for different domains. For instance, the system should be able to automate a task, in a sales office or in a newspaper editorial office.

Fig. 5. Readiness-vs-generality chart for learning by demonstration. We see that layer 1 reaches TRL 9, especially because of the possibilities of deep reinforcement learning using human traces. Layer 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, layer 3 requires learning systems that can process background knowledge in any domain, which is still at a very preliminary stage (TRL 2) with no more than the principles and their envisaged applications.

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 AI has been the development of tools that allow humans and machines to communicate more smoothly in natural language, and more generally 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. Machine translation

Machine translation (MT) is the automatic translation of texts from one language into another language. While human translation is the subject of applied linguistics, machine translation is seen as a subarea of artificial intelligence and computational linguistics. 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.

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

For the first two layers, it is clear we already reached TRL 9, with a myriad of translation products[14] including dictionaries[15] and thesaurus,[16] helping to combine machine and human-based translations.

In terms of current developments in FAMT (third layer 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 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[17] are best suited when we need quick, one-off translations and accuracy is not of importance. Also, MT applications are particularly effective in

[14] See, e.g., https://www.sdl.com/,https://www.memoq.com/ orhttps://www.wordfast.net/.
[15] See e.g., https://www.wordreference.com/.
[16] See, e.g., https://www.thesaurus.com/.
[17] See https://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications for a comparison.

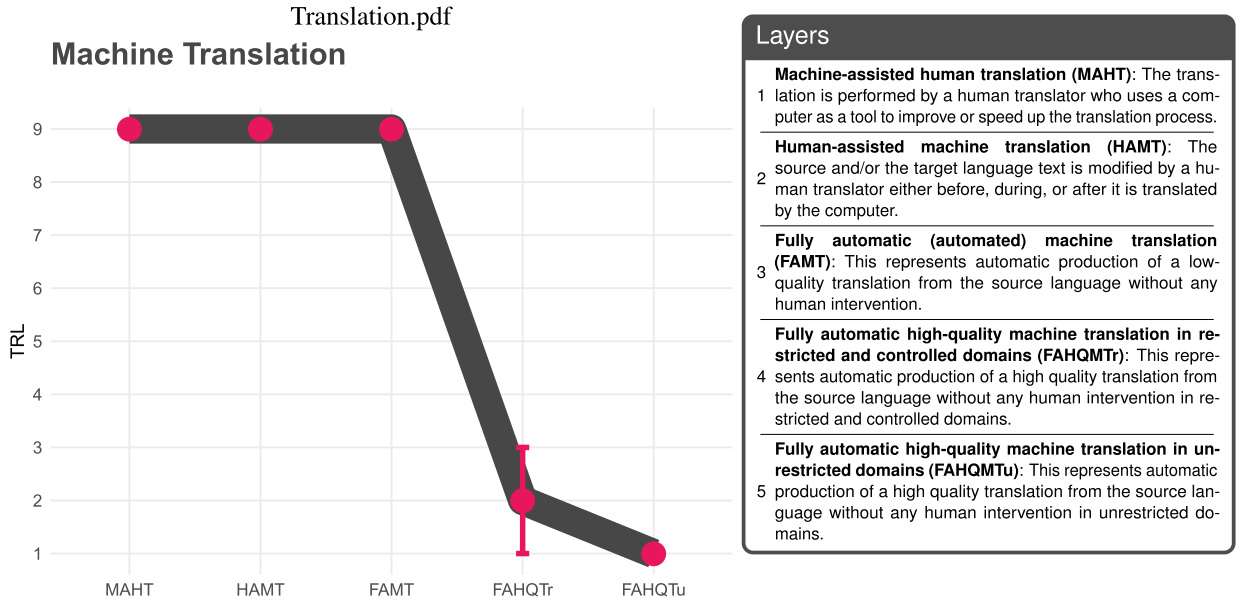


Fig. 6. Readiness-vs-generality chart for Machine Translation (MT) technology. TRL 9 has been reached for the first two types of MT (MAHT and HAMT). Currently, FATM approaches have reached a crucial moment, with powerful market-ready products such as Google Translator or DeepL, and a lively research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

domains where formal (structured) language is used. Finally, it should also be noted that although this layer 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 layers correspond with the ultimate goal of MT: FAHQMT. As already mentioned, MT produces usable outputs when doing informal translations. 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. 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 (Hillel, 1964). More recently, some research (TRL 1 to TRL 3) is being reached for restricted scenarios (see, e.g., Muegge, 2006), corresponding with the fourth layer. Layer 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 subareas such as Speech Transcription (Seide et al., 2011) and Spoken Language Understanding (Tur and De Mori, 2011), among others, but we focus on the former. 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% accuracy rate (Juang and Rabiner, 2005). Nowadays, systems 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).

Because of the evolution of expectations and capabilities of speech recognition technology, the *x*-axis of Fig. 7 uses four different generality layers, as described in the right box. For the first layer, 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[18] in the 1980s) or in-car voice recognition systems (Tashev et al., 2009).

For the second layer, common applications today include voice interfaces in robots, digital assistants or specific software such as voice dictation, voice dialling or call routing, home 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 systems are still limited to restricted domains, also having problems with noisy environments, different accents, disorganised conversations, echoes, speaker distance from the microphone, etc.

[18] http://www.robotsandcomputers.com/robots/manuals/Julie.pdf.

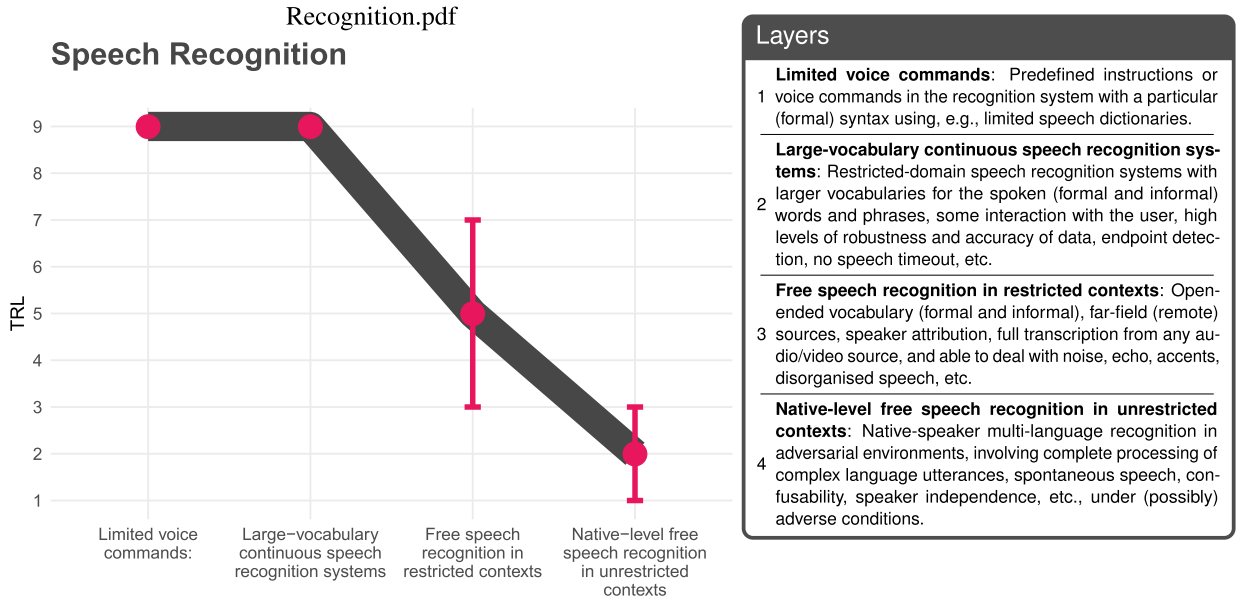


Fig. 7. Readiness-vs-generality chart for speech recognition technology. TRL 9 has clearly been reached for narrow systems with limited voice commands or rigid conversational interfaces such as those shown by the widespread virtual assistants such as Amazon's Alexa, Apple's Siri, etc. Current research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

Layer 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 misinterpretations being common. 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, so they are given TRLs from 1 to 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., modelling speaker intention) are clearly more long-term goals in the field.

4.4. Perception

Perception is a capability that we find in many 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 granted this predominance to the field machine vision.[19] 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 identity 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 significant attention, becoming an active research area that covers 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 this technology, the *x*-axis of Fig. 8 uses three different generality layers of facial recognition systems. Regarding the first layer, most current facial recognition systems excel in matching one image of an

[19] This predominance is perhaps exaggerated, as there are many other ways in which AI can interact with the world and achieve intelligent behaviour, as clearly illustrated in the case of humans by blind people from birth.

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,[20] find missing people,[21] etc.); retail and advertising (e.g., enabling more targeted advertising [22]), social media (e.g., to automatically recognise when its members appear in photos), financial services (e.g., digital payments or online account access[23]) among others. At present development levels, these systems are also able to detect people's gender (see, e.g., Mansanet et al., 2016), age[24] and even emotions (see, e.g., Ko, 2018) with accuracy levels of over 99%.[25] However, these systems still rely on full frontal face images with little or no change in illumination and orientation angle.

As for the second layer, 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., Li and Da, 2012), and age (see 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) such as railway stations and airports (e.g., boarding controls in airports or train stations[26]). However, at this layer of generality, the technology is having two major drawbacks: (1) performance: facial recognition is still much more effective in "constrained situations" than in more general and uncontrolled scenarios where illumination, pose, angle, position and expression are the major uncontrolled parameters that make facial recognition a hard task. Most facial-recognition algorithms also exhibit racial bias showing a wide range of accuracy across demographic differences (Grother et al., 2017) (see, e.g., the NIST Face Challenges[27]). (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[28] (approval is needed for TRL 8). Nevertheless, there are some surveillance and security systems currently operating (TRL 9) in countries such as China (see, for instance, the YITU Dragon Eye products used in Shanghai Metro[29]).

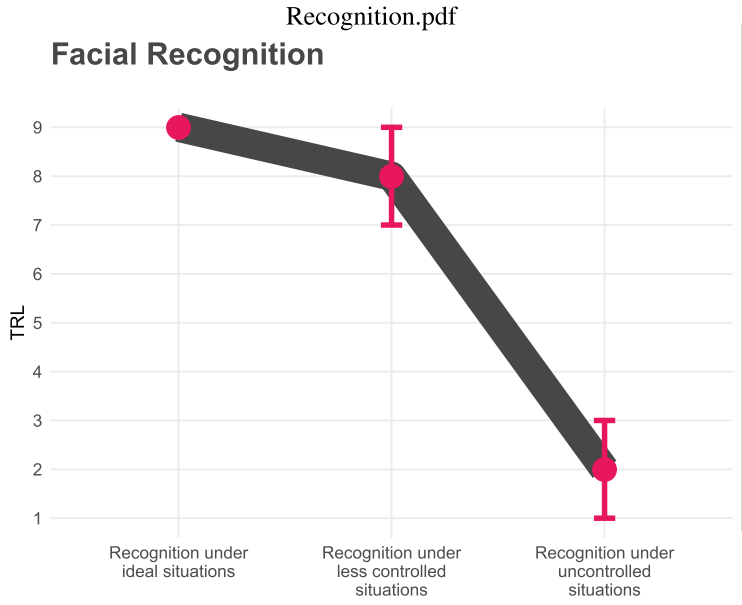
A further step in generality, corresponding with the third layer in Figure 8, involves addressing more complex factors (in uncontrolled scenarios) such as inadequate illumination, partial or low quality image or video (e.g., only one eye is visible), multiple camera angles, poses or image variations (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., 2003), 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% to 50% in some cases (Elmahmudi and Ugail, 2019).

4.4.2. Text recognition

Text Recognition is the process of digitising text by automatically identifying characters, or more generally symbols, from an image belonging to a certain alphabet, making them accessible in a computer-friendly form for text processing programs. 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 focus on the former. Large amounts of written, typographical or handwritten information exist and are 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.

Fig. 9 tries to model the evolution of expectations in terms of the different capabilities of text recognition technology, as shown in the box on the right. For the first layer we find the most simple (and common) form of character recognition: template-based OCR. This has been instrumental in automating the processing of managing physical typewritten documents. For instance, companies 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, driving licenses 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 98–99% for printed text (Holley, 2009), which is good enough for most applications, or 95% when addressing, for instance, very specific handwritten recognition tasks such as postal address interpretation (see, e.g., Srihari and Kuebert, 1997). Most

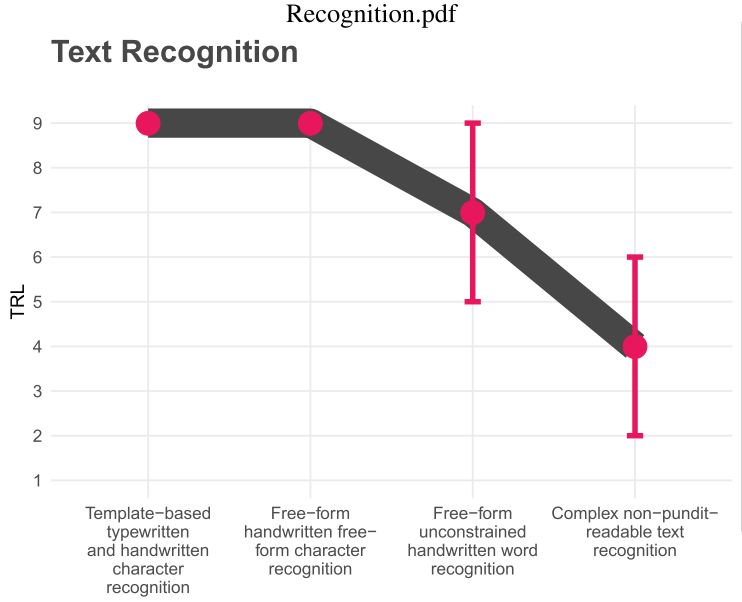
[20] https://www.interpol.int/How-we-work/Forensics/Facial-Recognition.
[21] https://www.independent.co.uk/life-style/gadgets-and-tech/news/india-police-missing-children-facial-recognition-tech-trace-find-reunite-a8320406.html.
[22] https://www.theguardian.com/business/2013/nov/03/privacy-tesco-scan-customers-faces.
[23] https://findface.pro/en/solution/finance/.
[24] https://labs.everypixel.com/api/demo.
[25] https://paperswithcode.com/task/face-recognition orhttps://neurosciencenews.com/man-machine-facial-recognition-120191/.
[26] https://www.airportveriscan.com/.
[27] https://www.nist.gov/programs-projects/face-challenges.
[28] 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 orhttps://sciencebusiness.net/news/eu-makes-move-ban-use-facial-recognition-systems.
[29] https://www.yitutech.com/en.



Layers

1. **Recognition under ideal situations**: gender, age or identity recognition from high-quality still images of full frontal faces in controlled scenarios (illumination, camera and person are controlled).
2. **Recognition under partially controlled situations**: gender, age or identity recognition from low-quality frontal faces (≅0 degrees off) possibly in less controlled scenarios: illumination and camera are controlled, but not the person (e.g., station or airport check-ins).
3. **Recognition under uncontrolled situations**: gender, age or identity recognition at pose variations, in low-resolution and poorly illuminated photos or 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, dash-cams, etc. The recognition must be robust to characteristics such as race, sex, etc., as well as changes in hairstyle, facial hair, body weight, and ageing.

Fig. 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 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 or resolution, bad lighting, etc., and with people of various race, sex and other personal characteristics (e.g., changing facial hair, body weight, accessories, etc.).



Layers

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.).
2. **Free-form handwritten character recognition**: recognition of (non-)separable/segmentable handwritten characters with automatic layout analysis in unstructured documents.
3. **Free-form unconstrained handwritten word recognition**: recognition of unconstrained non-segmentable handwritten words in unstructured documents.
4. **Complex non-pundit -readable text recognition**: recognition, interpretation and deciphering of non-readable media (e.g., ancient or badly damaged which cannot be deciphered by non-expert humans) or unconstrained texts in any format.

Fig. 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.

commercial products and software are of this type (TRL 9).[30]

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, presenting new challenges. This represents thus the second layer of capabilities in Fig. 9. Because this process is involved in recognising handwriting text, accuracy levels may, in some circumstances, not be very good, but the systems still can achieve 97–99% accuracy rates in structured forms when handling capital letters and numbers (Ptucha et al., 2019) that are easily segmentable. They 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[31] (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) with new alternatives (e.g., neural approaches) being developed and assessed.

The third layer of capabilities represents further advancements in this sort of technology involving recognition of unconstrained (i.e., non-easily segmentable) and free-form handwritten word (instead of "character") recognition[32]). "Intelligent word recognition" (IWR) technologies[33] may fall into this layer. 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 recognises 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-written 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[34] used in Google Docs[35] and Google Lens app[36]), as well as a number of prototypes being tested and validated in relevant environments (Yuan et al., 2012; Acharyya et al., 2013).

Finally, much more advanced text recognition systems would be needed, for instance, to interpret ancient or badly damaged texts that can only be deciphered by pundits or even not deciphered by humans. There are some research efforts in this direction (see, e.g., Lavrenko et al., 2004; Sánchez et al., 2013; Granell et al., 2019; 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

This AI category covers a continuum from planning, usually dealing with choosing the best sequence of actions according to some utility function, and scheduling, dealing with 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 exemplar of an AI technology that represents *systems that plan*.

4.5.1. 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). Transport scheduling solutions (described using action languages Lifschitz, 1999) are usually obtained by means of iterative trial and error processes including, mainly, combinatorial and search optimisation (Matyukhin et al., 2017), but also dynamic (Dai et al., 2020) and constraint programming (El Hachemi et al., 2011) approaches. Note that these sort of solutions can be found and evaluated (1) prior to execution (known environments), or (2) needs to be revised and adapted online (unknown environments). Furthermore, unlike classical control problems, the solutions must be discovered or optimised in a multidimensional space. 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), taking several parameters as input, 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 Fig. 10 uses three different generality layers, explained in the box on the right. Although, traditionally, transport timetables have been manually generated (e.g., using time-distance diagrams Chakroborty and Das, 2017), the process becomes unfeasible when dealing with high-load transport networks. At the first layer of generality, computer-based scheduling and planner systems based on optimisation heuristics have appeared over the last decades to provide automated and optimised transport scheduling for vehicles and drivers. These

[30] https://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software.

[31] Seehttp://www.cvisiontech.com/library/ocr/text-ocr/intelligent-character-recognition-software.html,https://abbyy.technology/en:features:ocr:icr orhttps://www.scanstore.com/Forms_Processing_Software/ICR_Software/.

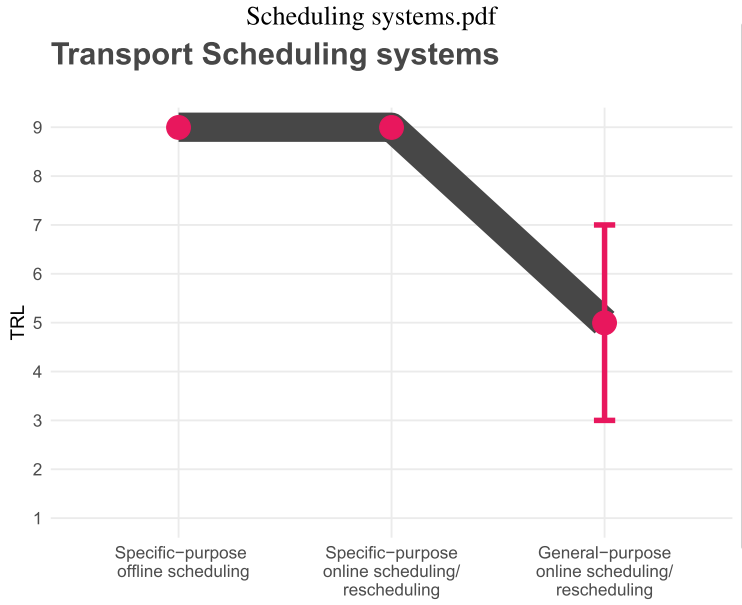
[32] Note that the transcription at further levels (e.g., line or paragraph) goes beyond this technology as it involves other technologies such as (joint) line segmentation (Bluche, 2016.

[33] 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.

[34] https://cloud.google.com/vision/docs/handwriting.

[35] https://docs.google.com/.

[36] https://lens.google.com/.



Layers

1. **Specific-purpose offline scheduling**: all the information is available beforehand with no uncertainty, serving as a batch input that produces the optimised schedule as output. The particularities of the domain are embedded into the system so that only the data is given as an input.
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 should have arrived at 3:30 but arrives at 3:40). Still, the particularities of the domain are embedded into the system.
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.

Fig. 10. 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 layers. More general-purpose scheduling systems have a lower TRL, between 3 and 7.

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,[37] OTT[38] or MULTIRAIL[39]; (b) flights (Feo and Bard, 1989), also with a myriad of commercial applications such as FLIGHTMANAGER,[40] OASIS[41] or TAKEFLIGHT[42]; (c) buses and shuttles (Gavish et al., 1978), with software platforms such as GOALBUS,[43] TRIPSPARK[44] or REVEAL[45]; (d) maritime transport (Meng et al., 2014) with commercial software such as MJC2,[46] or MES[47]: and (e) road transport (Törnquist et al., 2006), with software products such as PARAGON[48] or PARADOX.[49] 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 layer, we consider that the input information can be provided online, so that an automated scheduling system needs to process it in real time. The systems should therefore have two parts: off-line scheduling (for known information) and on-line rescheduling. While the former is in charge of scheduling vehicles and crews from known information, the latter has to be applied in response to the new specific needs and incidents that may appear. The schedules have to be dynamically updated balancing the resources (vehicles, time-slots, crew, etc.) available. Real-time incidents may required, for instance, to meet specific travel demands or requests of passengers (e.g., new stops), to adapt to perturbations or problems regarding resources or demand (e.g., failures in vehicles), or manage new schedule intervals between new events (e.g., as volcano eruptions or heavy weather-related issues). Dealing with real-time needs also entails that scheduling systems have to be able to confront different layers 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). As in the first layer, we are able to find plenty of research in this regard (see, e.g., Eberlein et al., 1999; D'Ariano et al., 2008; Verderame et al., 2010; Reiners et al., 2012) as well as market-ready applications (e.g., TPS[50] for trains, OPTIBUS[51] for bus and shuttles) applied to different transport scenarios, thus implying a TRL 9 for this sort of more capable scheduling systems. Note that these systems make use of search

[37] https://www.rmcon-int.de/railsys-en/.
[38] https://www.via-con.de/en/development/opentimetable/.
[39] https://www.oliverwyman.com/our-expertise/insights/2013/jan/multirail-pax-_-integrated-passenger-rail-planning-.html.
[40] https://www.topsystem.de/en/flight-scheduling-1033.html.
[41] http://www.osched.com/.
[42] https://tflite.com/airline-software/Passenger-Service-System/flight-schedule/.
[43] https://www.goalsystems.com/en/goalbus/.
[44] https://www.tripspark.com/fixed-route-software/scheduling-and-routing.
[45] http://reveal-solutions.net/bus-routing-scheduling-software/bus-scheduling-software-101/.
[46] https://www.mjc2.com/transport-logistics-management.htm.
[47] https://cirruslogistics.com/products/marine-enterprise-suite/.
[48] https://www.paragonrouting.com/en-gb/our-products/routing-and-scheduling/integrated-fleets/.
[49] https://www.paradoxsci.com/transportation-logistics-software-rst.
[50] https://www.hacon.de/en/solutions/train-capacity-planning/.
[51] https://www.optibus.com/.

approaches for rerouting optimization purposes as well as several other procedures and techniques such as stochastic, parametric, fuzzy or constraint programming, robust optimization techniques, or conditional value-at-risk, among others (Verderame et al., 2010).

Finally, for the third layer, 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 (scheduling a fleet of trucks based on road-traffic characteristics is not the same 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 an optimised schedule for any kind of service or a combination of them. In Hassold and Ceder (2014), Liu and Ceder (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 to be given instructions (e.g., where to go and what to clean) and follow them autonomously. The following two exemplars are good examples of AI technologies that represent *systems that interact physically*.

4.6.1. 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 the greatest forces transforming modern transportation systems. For a vehicle to be autonomous, it needs particular underlying AI capabilities mostly related to (1) navigation (e.g., transfer of objects and oneself from one place to another at different scales, such as rooms, buildings, towns, landscape, roads, etc., through different modalities and approaches such as landmarking, geolocations, etc.); and (2) naive physics (e.g., identification and tracking of pedestrians and other vehicles, object movement prediction, obstacle avoidance, keeping safe trajectories and inter-vehicle distance, etc., using different approaches such as Hidden Markov Models (Kelley et al., 2008), Kalman Filters (Siegwart et al., 2011), (neural-based) heuristic approaches (Smolyanskiy et al., 2017), etc.)

(Fuzzy logic, Neural Network, Neuro-Fuzzy, Genetic Algorithm, Particle Swam Optimization, Ant Colony Optimization and Artificial Immune System) specially hybridized technique (Neuro-Fuzzy) gave suitable and effective results for mobile robot navigation (targetreaching and obstacle-avoidance

However, and despite extraordinary efforts from many of the leading names in tech and in automaking, fully-autonomous[52] cars are still out of reach except in special trial programs,[53] and their potential impact with respect to timing, uptake, and penetration remains uncertain[54] (Silberg et al., 2012).

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 layers of driving automation based on the amount of driver intervention and attentiveness required. This is just loosely related to the capabilities and generality of the self-driving engine. In particular, the US National Highway Traffic Safety Administration (NHTSA) defines six levels of car autonomy, which can be used to evaluate the self-driving capabilities of cars.[55] They released this guidance to both push forward and standardise autonomous vehicle testing. We therefore use these 'NHTSA levels' as (see the right box in Fig. 11 for their definition).

For the first layer (NHTSA level 0), there is no automation at all as the human does all the driving at all times. Most vehicles were at this level until very recently (TRL 9). In the second layer (NHTSA level 1) we find some assistance systems for driving and maintenance such as the Adaptive Cruise Control (ACC), which 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 layer (NHTSA 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 centring (or auto steer, a mechanism that keeps a car centred in the lane) capabilities met the definition of Level 2. Tesla's Auto-Pilot[56]

[52] 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.

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

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

[55] 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.

[56] https://www.tesla.com/autopilot.

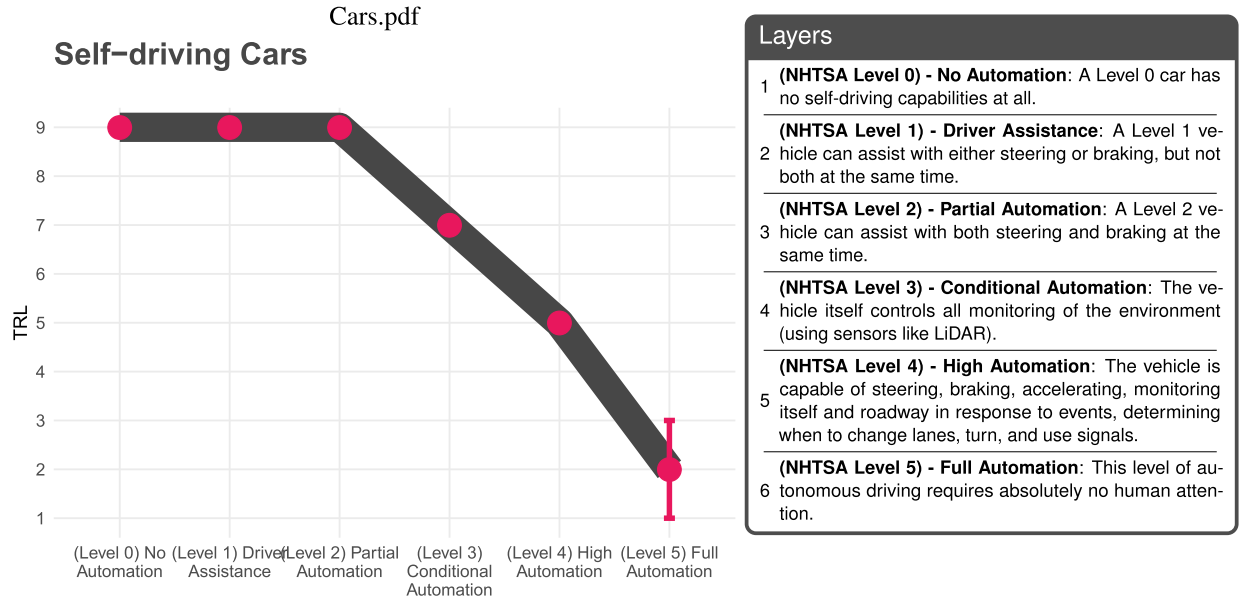


Fig. 11. Readiness-vs-generality chart for self-driving cars technology. TRL 9 has been clearly reached by many cars in our roads in the layers between NHTSA levels 0 and 2 of automation. For NHTSA levels 3 and 4, current developments from automobile companies are currently performing research, prototyping 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.

feature, as seen on the Model S, X, and 3, currently falls into this Level 2 category (TRL 9).

As for the fourth layer (NHTSA 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 already announced Level 3 autonomous cars to launch in 2018, but it did not actually happen due to the restrictive regulatory, technical, safety, behavioural, legal and business-related complications (TRL 8).

At the fifth layer (NHTSA 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.[57] 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 program,[58] but they are limiting the conditions in which the vehicles are allowed to drive (e.g., in dry weather areas).

Finally, for the sixth layer (NHTSA 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 a successful proof-of-concept we find Nuro,[59] which 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.[60] Given this partial evidence, fully self-driving cars are not here yet and we have to assign a TRL between 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. As mentioned above, there are some exceptions, such as certain Tesla models and Google's Waymo featuring a

[57] https://hondanews.com/releases/honda-targeting-introduction-of-level-4-automated-driving-capability-by-2025.
[58] https://waymo.com/apply/.
[59] https://www.reviewgeek.com/1717/two-ex-googlers-want-nuro-a-new-self-driving-car-to-handle-your-deliveries/.
[60] https://www.theverge.com/2019/12/9/21000085/waymo-fully-driverless-car-self-driving-ride-hail-service-phoenix-arizona.

limited set of self-driving capabilities (e.g., enabling the car to steer, accelerate and brake on behalf of the driver), but these are still research projects in initial or testing or trial programs.[61] 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, 2013) or the media announcements made just a few years ago[62] claiming that by 2020 we all would be permanent backseat drivers.[63] This provides very gloomy evidence about the complexity and difficulty of the driving tasks (with the high layer of reliability required König and Neumayr, 2017), with even those simplest subtasks (e.g., tracking other vehicles and objects around a car on the road) actually being much trickier than they were thought to be.

4.6.2. 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.[64] 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. In general terms, this technology uses location algorithms and coverage optimisation approaches (e.g., random walk-based algorithms, spiral algorithms, path transforms, genetic algorithms, etc.) for solving the problem of indoor cleaning tasks (Sörme and Edwards, 2018).

However, despite their popularity, we can analyse how far we are from reaching the original goals if we analyse these technologies by their layer of generality, as described by the box on the right of Fig. 12. For layer 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 AI categories (such as speech recognition for voice commands or even basic conversation capabilities). However, they are still at this layer 1, as they are not able to manipulate objects. A similar situation happens with other specific tasks such as window cleaning,[65] pool cleaning, lawn mowing or car washing.

The second layer involves the manipulation of objects, which requires more advanced recognition of the environment and dexterity. There are current prototypes[66] to fold laundry (Bersch et al., 2011; Miller et al., 2012 or iron clothes[67] Estevez et al. (2020). More complex tasks such as making the bed or clean the bathroom[68] 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 layer is still in very early stages, and we do not have evidence to assign a value beyond TRL 1. About the near future, it is clear that innovations are required at layer 2, before moving to significant progress at layer 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 layer 1. For example, they add video conferencing and voice assistants to their devices rather than the ability to actually manipulate objects or a diversification of 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 layer 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 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

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

[62] See, e.g., https://www.wired.com/story/gms-cruise-rolls-back-target-self-driving-cars/, https://www.theatlantic.com/technology/archive/2018/03/the-most-important-self-driving-car-announcement-yet/556712/, https://www.wsj.com/articles/toyota-aims-to-make-self-driving-cars-by-2020-1444136396 orhttps://www.autotrader.com/car-shopping/self-driving-cars-honda-sets-2020-as-target-for-highly-automated-freeway-driving-266836.

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

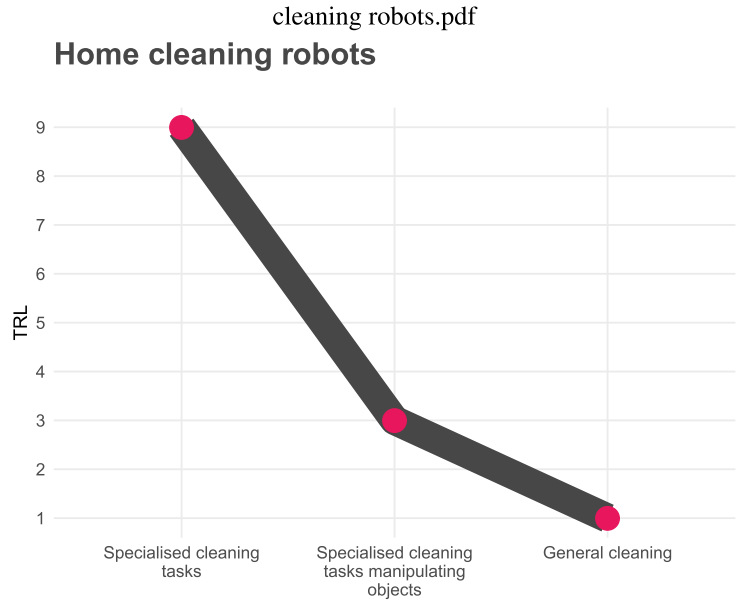
[64] An early example of this is the 2001 Electrolux robot vacuum cleaner (https://www.electroluxgroup.com/en/trilobite-advert-elubok115-2/).

[65] https://www.digitaltrends.com/home/best-window-cleaning-robots/.

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

[67] https://helloeffie.com/).

[68] 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/).



Layers	
1	**Specialised cleaning tasks**:The robot 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 well-defined objects such as walls and windows), and avoiding obstacles. The robot may have a physical configuration that exploits the particularities of the task, such as a roundish device moving on the floor or a small autonomous drone doing the windows.
2	**Specialised cleaning tasks manipulating objects**: 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.
3	**General cleaning**: the same robot does many home chores. The robot may have a more flexible physical configuration (probably with some kind of robotic limbs) and a sophisticated interplay between perceptors and actuators. We do not imply though that these robots have to be humanoid, clean exactly as humans do or use the same instruments (broom, vacuum cleaners, etc.).

Fig. 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 layer.

negotiation agents. This AI technology is representative of *systems that collaborate socially*.

4.7.1. 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 systems (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 cover the layers 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* (Baarslag et al., 2015), although the latter has incorporated new challenges over the years[69] (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), being transparent about the techniques[70] that are used (argumentation techniques 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).

Following this increasing generality mostly due to the complexity and diversity of the trading rules and the other agents, the layers we use for negotiation agents are defined in the right box in Fig. 13. Early negotiation agents can be found at layer 1 using the basics of decision theory (Parsons et al., 2012), and at this level many negotiating agents do not even need AI (Lin et al., 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 al., 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

[69] https://web.tuat.ac.jp/katfuji/ANAC2020/.

[70] 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 and Parsons, 2002).

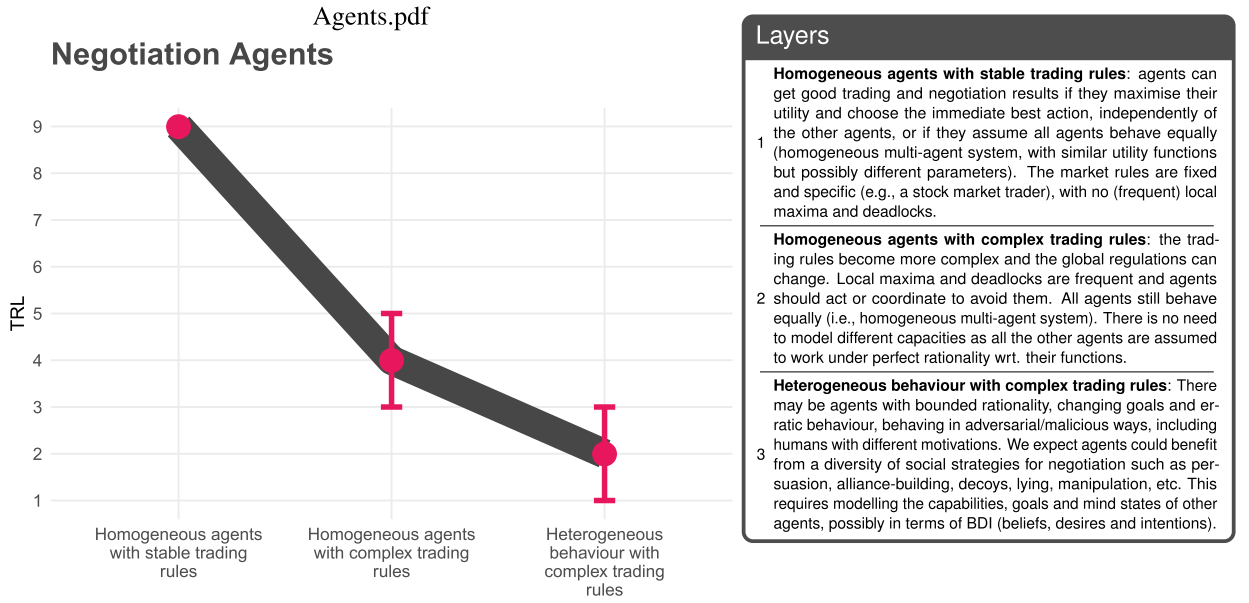


Fig. 13. Readiness-vs-generality chart for learning by demonstration. Layer 1 reaches TRL 9, with some negotiation bots running in simple scenarios. Layer 2 is more challenging, and TRL ranges between 3 and 5. Finally, layer 3 is still far ahead in the future, with an estimated TRL between 1 and 3.

markets, and more recently in cryptocurrencies. While they reach high TRLs at this layer, there is the question of whether they really help their users (or owners) make profits. 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 layer.

Layer 2 expects global rules to change and the utility functions to have different values. These two aspects are sometimes referred together as "domain knowledge and preference elicitation". As per 2017, it was considered a "challenge" (Baarslag et al., 2017), with active research in on-line or incremental preference extraction (Baarslag and Gerding, 2015; Baarslag and Kaisers, 2017), as well as in domain modelling (Hindriks et al., 2008; Sanders and Stappers, 2008; Simonsen and Robertson, 2012). However, in some scenarios such as e-commerce between companies, there have been some patents being filed (Krasadakis, 2017). 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 of the above, we consider a range between TRL 3 and TRL 5 for this layer as all the activity is still in the research and prototyping phases.

When it comes to layer 3, we have seen much activity at the research levels, with bounded rationality and heterogeneous utility functions (theoretically Sofy and Sarne, 2014 or in simulations for specific contexts Rosenfeld and Kraus, 2009), 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), 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 layer.

Layer 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 layer, such as 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 layer.

5. Discussion: rearranging the generality

Once the series of exemplars of AI technologies have been analysed in the previous section, falling into one of the seven AI categories corresponding to areas of AI, we can now recapitulate with what we observe in the readiness-vs-generality plots more globally.

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

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 Figs. 4 - recommender systems, 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 layer on the *x*-axis (see Figs. 3 - knowledge inference engines, 5 - apprentices by demonstration, 9 - text recognition, 12 - home cleaning robots and 13 - negotiation agents). Importantly, the shape of the curves depends on the definition of layers in the *x*-axis (all charts are summarised in the following subsection, see Fig. 15). Refining one layer into two or three finer layers may well flatten the curve. This is true, of course, but also a good indication of a way in which an insurmountable layer 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 on the following steps, or further refinements through the *x*-axis.

Opposed to this disaggregation direction, there is also a trend to consider technologies that, by definition, are expected to integrate many capabilities. A very good example of these integrating AI technology 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 integrating 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 (Commission, 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 yet again 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 somewhat similar with machines.

As said above, domain generality is very important, because we want these systems to do a wide range of things. However, this is more a desiderata than a reality, or even a need for some applications. This is similar to the cases with other AI technologies analysed in the previous section, such as knowledge inference engines. In particular, making an assistant for a narrow domain (a telecommunication company assistant or a ticket-purchasing service avatar) is easier than a more open-ended assistant (an executive assistant in the workplace). One important characteristic of these more open-ended systems would be to recognise what they cannot do, which is usually paraphrased by the famous "I'm afraid I can't do that" (Kubrick and Clarke, 1968). Given these considerations, we introduce a three-layer scale for generality of virtual assistants as shown in the box of the right of Fig. 14, which may of course be refined in the future.

In terms of capabilities, the simplest VAs (layer 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 the form of simple chatbots in customer-service applications on websites and other apps for restricted (simple) domains (e.g., ticket purchase assistants, legal counselling chatbots, etc.). Consequently, we can assign TRL 9 to these assistants (e.g., VAs for question and answering (Q&A) of Coronavirus-related content[71]).

Focusing on layer 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 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 able to take some decisions and fully support 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 (Commission, 2018; Hoy, 2018) being the main exponents. 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 to the market. Because of this, we assess a range of values between TRL 7 and 9, as shown in the figure.

Finally, layer 3 VAs are foreseen 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.[72] In this layer, 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 layer 3, VAs are not expected to do complex rationales or make sophisticated decisions. This is covered by technologies such as knowledge inference engines or planning. Of course, if high TRLs were obtained in these technologies they could end up being integrated in VAs, as they are usually shipped as amalgamators of AI services.

[71] See, e.g., https://avaamo.ai/projectcovid/ orhttps://www.hyro.ai/covid-19.
[72] https://www.cnet.com/news/facebook-ai-chief-we-want-to-make-smart-assistants-that-have-common-sense/.

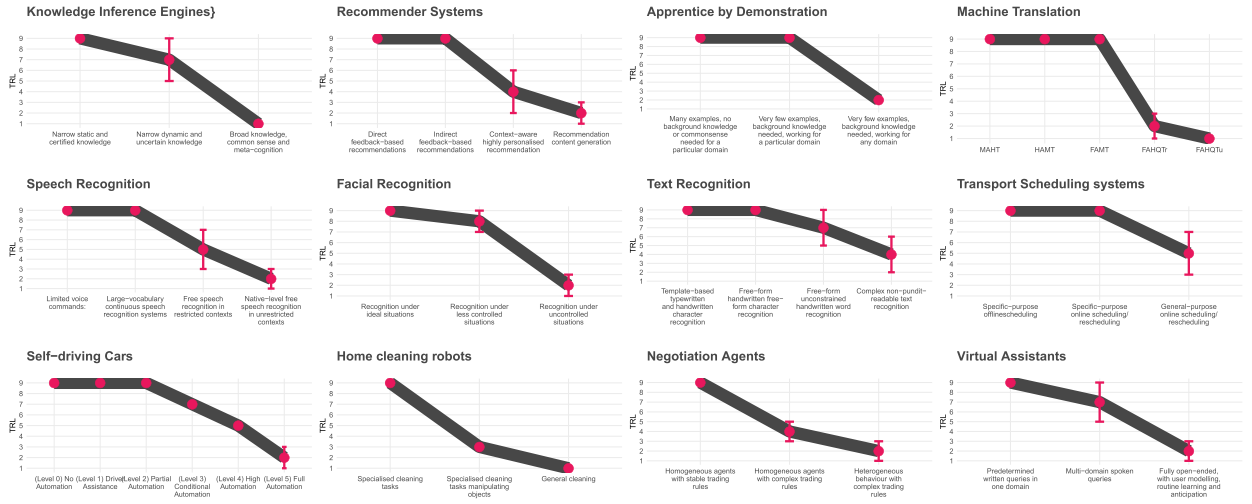


Fig. 15. A composition of all readiness-vs-generality charts from Figs. 3–14.

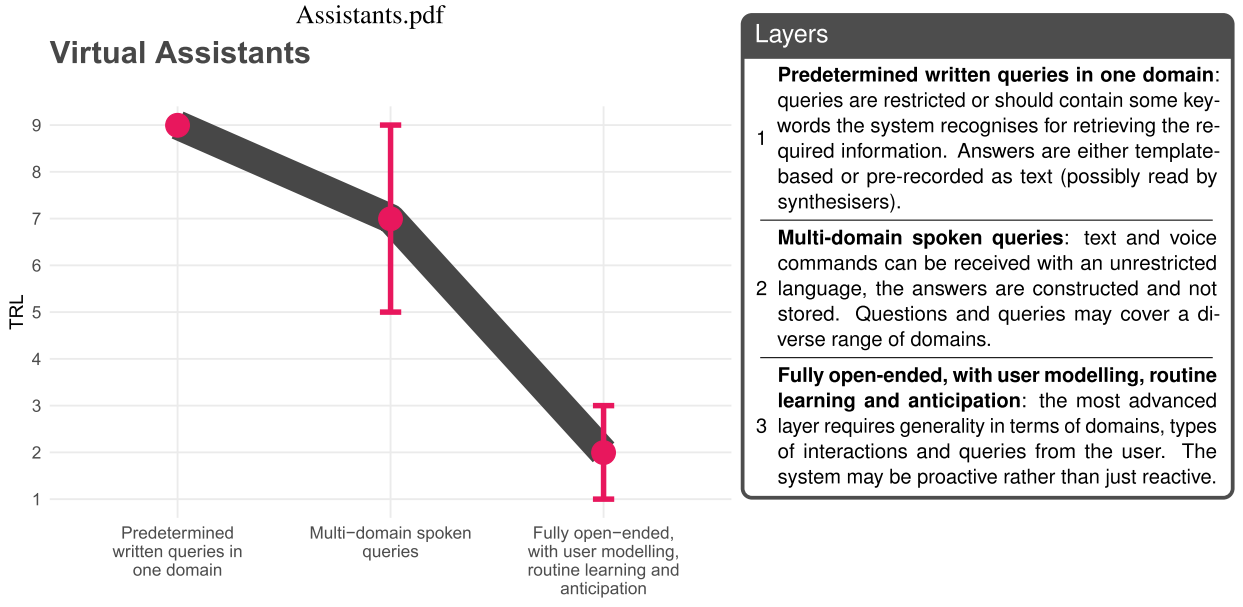


Fig. 14. Readiness-vs-generality chart for virtual assistant technology. TRL 9 has been reached for systems that work with predetermined written queries, high TRL are more diverse with open-ended spoken queries. Finally, the most advanced layer requires generality in terms of domains, types of interactions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

5.2. *Contouring technologies more precisely*

From the previous discussion we see how important it is to refine the *x*-axes such that layers are sufficiently crisp for a really accurate assessment of TRLs. Note that assessing TRLs correctly is crucial for, among others, standardisation organisations, such as NIST[73] (USA), with a long tradition on the evaluation of performance of intelligent systems (Huang et al., 2005; Messina and Jacoff, 2006; Marvel et al., 2012), or the "evaluation of artificial intelligence systems" programme at the *French Laboratoire National de Métrologie et d'Essais*.[74] However, this assessment becomes more difficult as the technology is broader, especially those that are defined by integrating capabilities from different categories (subareas) 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.

[73] https://www.nist.gov/.
[74] https://www.lne.fr/en/testing/evaluation-artificial-intelligence-systems.

Fig. 15 includes a summarised view of all readiness-vs-generality charts. As a simple way of comparing the technology, we could just look for the highest layer with TRL 9, or we could just focus on the narrowest (leftmost) layer. However, this would give a rosy view of the state of the art of AI technologies.

While it is true that these specialised early stages paved the way for more general versions of the technology, there are situations where a radical change was needed. In practice, the research pressure favours specialisation followed by incremental improvements (Adams et al., 2016) rather than radical general solutions. Funding agencies and companies usually reward projects where the technology is shown to work, even if it is just in one restricted domain. Accordingly, for the purpose of high TRLs, some research projects may be tempted to solve simplified versions of the problem for very narrow domains, with many ad hoc tweaks, rather than solving the general problem. Similarly, 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., 2017a; Silver et al., 2017b; Schrittwieser et al., 2019).

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; Bhatnagar et al., 2017; Martinez-Plumed et al., 2020a; Bhatnagar et al., 2017; Martinez-Plumed et al., 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 et al., 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 (Sun et al., 2020) of AI and anticipate the key transformations of the future. We explore this in more detail in the following section.

5.3. *Assessing TRLs more precisely*

In this work, we have assessed the TRL of each technology (at a particular layer) by asking experts (including ourselves) to follow the rubric in A to estimate the particular level in the scale (see C for further information. 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. Could we do otherwise?.

There are some sources of technology readiness 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 when we want to do this at the present moment and not in the past. Coverage on the media could also be a relevant source, and we could use repositories such as AI topics (Martinez-Plumed et al., 2018; Hernández-Orallo et al., 2020). However, there is an important source of quantitative information about the progress in AI: benchmarks and competitions (Hernández-Orallo et al., 2017).

The relation between benchmarks and TRLs is more complex than it looks initially. Some AI benchmarks (e.g., Atari 2600 games Bellemare et al. (2013)) would qualify as "simulated environments", as mentioned in the TRL 5 or TRL 6, depending on whether only some components are evaluated with them or a complete autonomous system. Other benchmarks, such as those used for self-driving vehicles, would qualify as "operational testing platforms" for TRL 7. Yet, other benchmarks, e.g., some Kaggle[75] competitions, are about real cases and their models could be applied directly, showing evidence of a TRL 8. 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 level they could be associated with, could lead to a more quantitative approach to estimating the TRLs.

To this end, we could use platforms such as *OpenML*,[76] *Papers with Code*[77] or the *AIcollaboratory* (Martinez-Plumed et al., 2020a; Martinez-Plumed et al., 2020b), which collect information for the analysis, evaluation, comparison and classification of different types of AI and machine learning systems. For the moment, we leave this mapping and quantitative analysis as future work. It is not just because the sheer volume of the endeavour but also because there are some issues to discuss and solve first to do this meaningfully and reliably. For instance, most benchmarks are not just pass or no-pass but are accompanied with one or more metrics, such as the level of performance. We should determine the minimum level of the accuracy for a benchmark that could be considered sufficient evidence for the associated TRL to be met. But this could create conflicts with the generality dimension. For instance, 70% performance on a face recognition benchmark could be considered useful for some applications and a proof of TRL 7, but it may well happen that most of the remaining 30% errors could focus on a particular niche of the technology (e.g., noisy pictures). Would this be evidence for a TRL 7 at that generality layer or at an inferior layer? 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 some other issues, such as systems being specialised to the benchmark but not to the real problem (so a TRL 7 would never translate into a TRL 9).

In the same line, it may also be the case that no appropriate (public) benchmarks or competitions exist for all sorts of AI technologies as some (if not all) of them involve different "skills" that may have (or not) specific benchmarks for assessing their performance. For instance, Virtual Assistants integrate, among others, abilities such as conversation/dialoguing, Q&A, information retrieval, speech recognition, etc. Likewise, Q&A involves abilities such as semantic parsing, answer generation, sentiment analysis, interaction,

[75] https://www.kaggle.com/.
[76] https://www.openml.org/.
[77] https://www.paperswithcode.com/.

etc. This hierarchy makes things more difficult when trying to address quantitative analyses. Another caveat is when human-equivalent (or super-human) performance is reached for AI benchmarks or competitions: they are often discontinued and replaced, or extended through the inclusion of more challenging benchmarks, in a kind of 'challenge-solve-and-replace' evaluation dynamic (Schlangen, 2019), or a 'dataset-solve-and-patch' adversarial benchmark co-evolution (Zellers et al., 2019). For instance, CIFAR10 (image classification) is accompanied by the more challenging CIFAR100 (Krizhevsky, 2009), SQuAD1.1 (Q&A) has been replaced by SQuAD2.0 (Rajpurkar et al., 2018), GLUE (language understanding) by SUPERGLUE (Wang et al., 2019), Starcraft (real-time strategy) by Starcraft II (Vinyals et al., 2017) and the Atari Learning Environment (ALE) (Machado et al., 2018) by the PlayStation Reinforcement Learning Environment (PSXLE) (Purves et al., 2019). However, and despite the speed and success of AI in many challenges which are beaten in a matter of months sometimes, progress in the field, especially in terms of TRL, seems unrelated.

Despite all these challenges, we do think that connecting benchmark results and TRLs is a very promising avenue of research, which we hope to dissect and develop soon. This could also lead to an overhaul of the current replacement dynamics and interpretation of AI benchmarks.

6. AI progress through TRL: 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, how useful are they 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 Fig. 4 – apprentice by demonstration) suggests that there may be a long way to go from one layer of the technology to the next one. The gap may include significant discoveries, results or inventions at some low TRLs, which may be linked to fundamental research, usually linked to slower progress. A flatter curve (such as in Fig. 8 – 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 layer 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. But in some sectors, the market already existed before automation. For self-driving cars, there is already 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.[78]

To better understand the speed of progress, we also need to consider the notion of technology "hyper adoption", which is related to the Hype Cycle from Gartner (Linden and Fenn, 2003). This theory states that people adapt to and adopt new technologies much faster than they used to do in the past. This may be partially caused by the so-called "democratisation" of new technology innovations, as they are available to anyone very soon. For instance, electricity took 70 years for mass adoption, but in the case of the Internet, it took just 20 years. The same is happening with AI technologies. A clear example is the current hyper-adoption of voice-related technology,[79] with all the tech giants such as Amazon, Google and Microsoft launching new products every few months. It may be the case that the advancements in this sort of technology has enhanced the adoption rates of voice assistants, and vice versa. This phenomenon may also have something to do with the shape of the TRL curves, but we will not delve into this further, as "hyper adoption" has also had some criticisms (Steinert and Leifer, 2010). The trend may even stop because of ageing populations in many countries, which are more reluctant to technological innovations.

In order to have more ground from extrapolations we would need a less static picture of the evolution of AI technologies. Having information about the charts in the past years would give us data about how curves evolve, and what TRL transitions are faster than others. Of course, trends may simply not exist or trends may cease to hold because of some radical changes in the AI playground or society (e.g., a big financial crisis, a pandemic or another AI winter). Nevertheless, to see the potential we will 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 some historical perspective? This is what we do next.

6.1. Readiness trends

As an example of the evolution of technology readiness, Fig. 16 shows three readiness-vs-generality curves for the case of virtual assistants on the same plot, as snapshots in 2000, 2010 and 2020. In the 1990s, digital speech recognition technology became a feature of personal computers of brands such as Microsoft, IBM or Philips, with no conversational or Q&A capabilities yet. In 1994, IBM launched the very first smartphone (IBM Simon Sager, 2012) with some assistant-like capabilities: sending emails, setting up calendars and agenda, taking notes (with a primitive predictive text system installed) or even downloading programs. However, it was a menu-based interaction, very different from the way assistants are known today. In this regard we may estimate that some research in this sense was being performed (TRL 1 to TRL 3), mostly focused in 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 (Herzog and Rollinger, 1991) on the domain of tourist information. These projects never went past the stage of successful demonstrations in

[78] https://www.blog.google/products/search/search-language-understanding-bert/.
[79] https://www.forbes.com/sites/forbestechcouncil/2018/06/08/the-hyper-adoption-of-voice-technology.

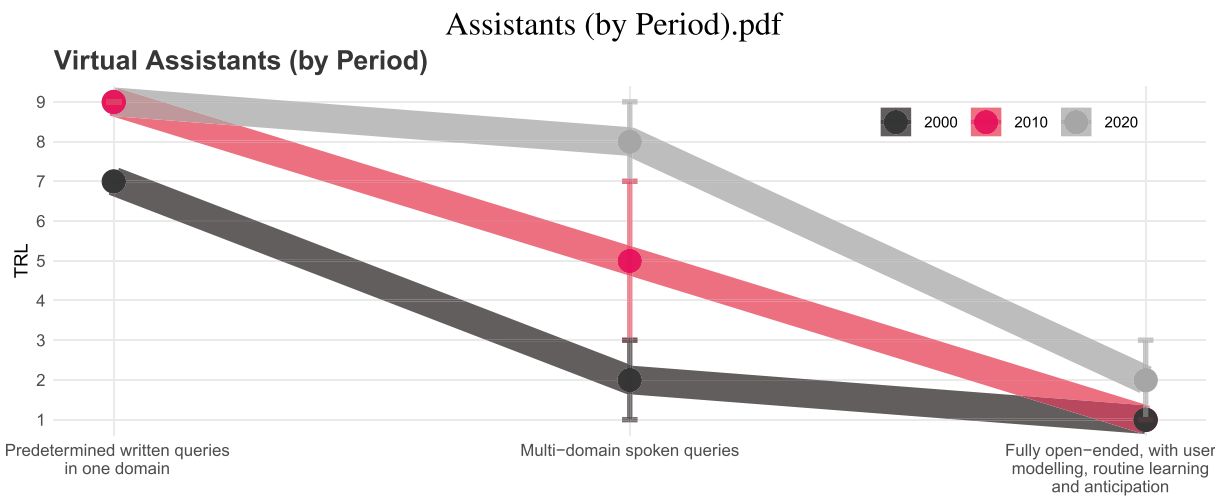


Fig. 16. Readiness-vs-generality chart for virtual assistant technology at different moments in time (grey: 2020, pink: 2010, black: 2000). We see how the "curve" has evolved from big step in the year 2000 between the first and second layers to another step between the second and third layers in 2020.

relevant scenarios (TRL 7). It should be also noted that a system at TRL9 for layer 1 would not be called a virtual assistant today, as the expectations have changed, but we keep the layers stable to see the historical progress.

By the decade of the 2000s, not only were there relevant advances in speech recognition technology, but also in Question Answering (with market-ready products such as Wolfram Alpha Wolfram, 2009), Information Retrieval and knowledge-Based Systems that paved the way for the future VA systems. One important milestone in this decade was the launch of Google Voice Search in 2002 (Franz and Milch, 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 for launching their 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 their prediction models of what a person is actually saying. At the same time, IBM also pushed their research in Q&A and information retrieval during this decade (from 2005 onward) with a 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 over the years 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) but still, 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, 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 evolved 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.

Note that, for layer 3, VAs are expected to have much more advanced capabilities (e.g., background knowledge, open-domain conversations, commonsense reasoning, etc.) that have not yet been found in the research agenda (TRL 1 to TRL 3) of natural language processing, planning, learning or reasoning. One may assume this will not happen until high TRLs are obtained for the second layer of generality which will 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.).

Even if there are many uncertainties when assessing and inspecting the evaluation of these curves with time, we think that this view is more robust than the observation of a single moment. And it is much better than one single point (the technology at the same layer) or mixing layers on the *x*-axis. With the usual mistake of conflating or collapsing layers, we could wrongly say that there has been no progress in smart phones in the past ten years, once the penetration of devices reached near 100%. But progress is evident: the percentage of time we use them has increased, because they have increased the generality of tasks and activities they can do, and their transformation goes on.

6.2. *AI futures*

There are many ways in which AI futures can be extrapolated, from expert panels (Müller et al., 2016; Grace et al., 2018; Betz et al., 2019) 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 specifications or simply not meeting any AI forecasting desiderata (Dafoe, 2018) [Ap. A]. Trying to rely on measurable indicators, we can connect the progress in AI with some economic indicators (such as the PREDICT dataset[80]). In this paper we have taken a different approach based on technology readiness levels. Our methodology serves both to describe the state of the art of a discipline (for applications such as project assessment or product development) and to use it for forecasting.

The truth is that we are still terribly bad at predicting the capabilities and products that 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 (Martinez-Plumed et al., 2018). We often fail to understand why some expectations are not met, 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 many explanations about why AI is not meeting expectations are related to the lack of generality of current AI technologies. This is one reason for expressing generality as a dimension in representations and measurements, and are 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 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 for 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 generality but massive penetration. But when a widespread system becomes more general (e.g., a mobile phone, useful for calls and messages, 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 wide generality of tasks, may represent a major transformation in the years to come. Hence the interest by the tech giants in investing in this technology.

If the dimensions are right, high TRLs for high layer 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 that one should consider to refine these predictions. Therefore, apart from task performance, one also needs to consider the availability of other supporting technologies, resources and costs incurred in the development of AI technologies (some of them usually neglected): data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, load, energy and (what kind of) time (Martinez-Plumed et al., 2018). Actually, these costs are distributed over the life cycle of the system, and may place differing demands on different developers and users. However, it is not straightforward to quantify what exactly can be attributed to some (or all) of these dimensions.

In this regard, a given technology may be ready but the costs of deployment may not be affordable for the mass. For instance, self-driving car technology can be based on 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. Also, even if some particular devices are massively flooding the market, that does not mean that they are used extensively. This happens with many gadgets that serve as toys (fancied the first days but forgotten afterwards). Some virtual assistants will have this fate in many homes. Sometimes products are sold before they are 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 backslash from consumers (e.g., Microsoft Clippy created aversion against assistants Veletsianos, 2007), or human labour costs may fluctuate, accelerating or slowing the adoption of certain technology (Borghans and Ter Weel, 2006; Suri, 2011; Sohn and Kwon, 2020). In other words, technological readiness does not mean technological success. Analysing all the factors contributing to the latter 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.

What we have covered in this paper is an exemplar-based methodology where (1) we identify the technology, its category and its scope, (2) we recognise and define the layers 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, if drawn historically, extrapolate its future trends. 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.

As future work, there are many avenues we would like to see explored. First, the reliability of the assessments in each of the charts could be increased by using external experts for each of them. With a larger and wider group of experts we could use methods such as Delphi. We could also derive the levels from the results of the associated benchmarks for each technology, as discussed at the end of the previous section. Second, covering many more 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 layers of generality. In some cases there may be different scales or even multidimensional (e.g., hierarchical) scales to explore.

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. The use the technology readiness levels and combining them with layers of generality, as we have done in this paper, may represent a

[80] https://ec.europa.eu/jrc/en/publication/2018-predict-dataset.

powerful and refreshing take on the state of the art of artificial intelligence, and how it is expected to affect our society in the near future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to the members of the panel of experts that provided valuable comments, suggestions and useful critiques for this work (in alphabetical order): Carlos Carrascosa, Blagoj Delipetrev, Paul Desruelle, Salvador España, Cèsar Ferri, Ross Gruetzemacher, Stella Heras, Alfons Juan, Carlos Monserrat, Daniel Nepelsky, Eva Onaindia, Barry O'Sullivan, M[a]José Ramírez-Quintana, Miguel Ángel Salido and Laura Sebastià.

This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018-094403-B-C3, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed acknowledges funding of the AI-Watch project by DG CONNECT and DG JRC of the European Commission. J. Hernández-Orallo is funded by an Future of Life Institute (FLI) grant RFP2-152.

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",[81] 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 of a technology's basic properties.
  - "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:** *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.
  - Academic environment.
  - 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).

[81] See https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2003/systems/nolte2.pdf, https://faaco.faa.gov/index.cfm/attachment/download/100020 or http://aries.ucsd.edu/ARIES/MEETINGS/0712/Waganer/TRL%20Calc%20Ver%202_2.xls, the latter from the US Air Force Research Laboratory.

  - 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 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.
  - 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**[82]**:** *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.
  - 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 a technology's demonstrated readiness. 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:** *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 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

[82] 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).

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

Appendix B. Categories of AI technologies

In this section we 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 initiative (Samoili et al., 2020) from the European Commission (EC), which proposes a multi-perspective analysis and operational definition to structure the AI taxonomy. In more detail, the starting point of the categorisation and operational definition is the (less technical) definition of AI adopted by the EU High Level Expert Group in 2019 (Artificial Intelligence, 2019). From there, the authors propose a flexible scientific methodology to collect, scan and analyse a large set of AI literature (from 1955 to 2019) by using Natural Language Processing methods as well as different sort of qualitative analyses. As output, the methodology provides a taxonomy and a list of keywords that characterise the core domains of the AI research field. This AI taxonomy has been designed to serve as a basis for analysing the global landscape of AI actors, and also for detecting AI applications in related technological domains, such as, robotics, neuroscience, internet of things, etc. Finally, the methodology can be updated over time to capture the rapid AI evolution, updating thus the operational definition of AI.

We have this taxonomy as reference because of the accuracy, exhaustiveness and topicality of its content, as well as the validity of the methodology followed for its development. 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**: This AI discipline 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 the actions to be planned or the solutions to be optimised are usually more complex than the outputs obtained in classification or regression problems. This extra complexity of planning is 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 and Walser, 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; Silver et al., 2017a; Brown and Sandholm, 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 (ABT), and Swarm Intelligence, where collective behaviours emerge from the interaction, cooperation and coordination of decentralised self-organised agents (Shoham and Leyton-Brown, 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 the collaboration (Chowdhury et al., 2010).

Note that this categorisation is sufficiently comprehensive of the areas of AI (and the cognitive capabilities that are being developed by the discipline) 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, but we do not expect to have technologies that we cannot be assigned to any category.

Appendix C. Panel of experts

Our initial assessment underwent a profound evaluation by an independent panel of specialists, recognised in at least one of the technologies (or areas) addressed.

All the experts (15) were selected based on their expertise in particular technologies or in AI in general. Most of them were selected from different institutions (large AI research labs and AI policy institutions). They did not have access to paper drafts or the discussion of TRLs in AI beforehand. All were contacted by email. The experts were asked to follow the rubric in A to estimate the particular level in the scale for specific technologies. Furthermore, experts provided 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 or regarding the whole analysis.

Many of them (13) were asked to send us feedback about one or two particular technologies, while two experts went through the whole document. When they provided more evidence of higher TRL we immediately implemented the change and added that evidence to the paper. When they suggested that higher TRLs were expected (but provided no evidence), we only took this into account if this was a general remark for the experts covering the particular technology. In a few cases we asked for clarifications.

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)
- Blagoj Delipetrev (European Commission)
- Paul Desruelle (European Commission)
- Salvador España (Universitat Politècnica de València)
- Cèsar Ferri (Universitat Politècnica de València – Machine Learning)
- Ross Gruetzemacher (Auburn University)
- Stella Heras (Universitat Politècnica de València)
- Alfons Juan (Universitat Politècnica de València)
- Carlos Monserrat (Universitat Politècnica de València)
- Daniel Nepelsky (European Commission)

- Eva Onaindia (Universitat Politècnica de València)
- Barry O'Sullivan (University College Cork)
- M[a]José Ramírez-Quintana (Universitat Politècnica de València)
- Miguel Ángel Salido (Universitat Politècnica de València)
- Laura Sebastià (Universitat Politècnica de València).

References

Hintz, A., Dencik, L., Wahl-Jorgensen, K., 2018. Digital Citizenship in a Datafied Society. John Wiley & Sons.

Brynjolfsson, E., Rock, D., Syverson, C., 2017. Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics (Technical Report). National Bureau of Economic Research.

Garcia-Murillo, M.A., MacInnes, I.P., Bauer, J.M., 2018. Techno-unemployment: a framework for assessing the effects of information and communication technologies on work. Telematics Inform.

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: Markham, A.N., Powles, J., Walsh, T., Washington, A.L. (Eds.), AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7–8, 2020. ACM, pp. 94–100.

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

Martınez-Plumed, F., Loe, B.S., Flach, P., O hEigeartaigh, S., Vold, K., Hernández-Orallo, J., 2018. The facets of artificial intelligence: a framework to track the evolution of ai. In: International Joint Conferences on Artificial Intelligence. pp. 5180–5187.

Mankins, J.C., 1995. Technology readiness levels, White Paper, April 6, 1995.

Russell, S., Norvig, P., 2020. Artificial Intelligence: A Modern Approach, fourth ed.

Charalambous, G., Fletcher, S.R., Webb, P., 2017. The development of a human factors readiness level tool for implementing industrial human-robot collaboration. Int. J. Adv. Manuf. Technol. 91, 2465–2475.

Buchner, G.A., Stepputat, K.J., Zimmermann, A.W., Schomacker, R., 2019. Specifying technology readiness levels for the chemical industry. Ind. Eng. Chem. Res. 58, 6957–6969.

Eljasik-Swoboda, T., Rathgeber, C., Hasenauer, R., 2019. Assessing technology readiness for artificial intelligence and machine learning based innovations. In: DATA. pp. 281–288.

Ellefsen, A.P.T., Oleśków-Szłapka, J., Pawłowski, G., Toboła, A., 2019. Striving for excellence in ai implementation: Ai maturity model framework and preliminary research results. LogForum 15.

Lavin, A., Renard, G., 2020. Technology readiness levels for machine learning systems, arXiv preprint arXiv:2006.12497.

Gadepally, V., Goodwin, J., Kepner, J., Reuther, A., Reynolds, H., Samsi, S., Su, J., Martinez, D., 2019. Ai enabling technologies: a survey, arXiv preprint arXiv: 1905.03592.

Salakhutdinov, R., 2015. Learning deep generative models. Annu. Rev. Stat. Appl. 2, 361–385.

Gruetzemacher, R., Whittlestone, J., 2019. Defining and unpacking transformative ai, arXiv preprint arXiv:1912.00747.

Hernández-Orallo, J., 2017. The Measure of All Minds: Evaluating Natural and Artificial Intelligence. Cambridge University Press.

Samoili, S., Cobo, M.L., Gomez, E., De Prato, G., Martinez-Plumed, F., Delipetrev, B., et al., 2020. AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence, Technical Report, Joint Research Centre (Seville site).

Kwok, R., 2019. Junior ai researchers are in demand by universities and industry. Nature 568, 581–584.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al., 2018. Never-ending learning. Commun. ACM 61, 103–115.

Gonçalves, R., Dorneles, C.F., 2019. Automated expertise retrieval: a taxonomy-based survey and open issues. ACM Comput. Surveys 52, 1–30.

Wagner, W.P., 2017. Trends in expert system development: a longitudinal content analysis of over thirty years of expert system case studies. Expert Syst. Appl. 76, 85–96.

Reinfrank, M., 1988. Reason maintenance systems. In: Begründungsverwaltung. Springer. pp. 1–26.

Shortliffe, E., 2012. Computer-based medical consultations: MYCIN, vol. 2. Elsevier.

Banks, G., 1986. Artificial intelligence in medical diagnosis: the internist/caduceus approach. Crit. Rev. Med. Informatics 1, 23–54.

Gibson, Carl S., 1990. Vax 9000 series. Digital Tech. J. Digital Equipment Corporation 2, 118–129.

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

Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., 2017. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics Inform. 34, 133–144.

Jayaraman, V., Srivastava, R., 1996. Expert systems in production and operations management. Int. J. Oper. Prod. Manage.

Dymova, L., Sevastianov, P., Kaczmarek, K., 2012. A stock trading expert system based on the rule-base evidential reasoning using level 2 quotes. Expert Syst. Appl. 39, 7150–7157.

Rasmussen, A.N., 1990. The inco expert system project: clips in shuttle mission control. First CLIPSConference 305.

Ahmed, M.N., Toor, A.S., O'Neil, K., 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, 4–9.

Ricci, F., Rokach, L., Shapira, B., 2011. Introduction to recommender systems handbook. In: Recommender Systems Handbook. Springer, pp. 1–35.

Howe, M., 2009. Pandora's music recommender, A Case Study, I. pp. 1–6.

Gomez-Uribe, C.A., Hunt, N., 2015. The netflix recommender system: algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6, 1–19.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al., 2010. The youtube video recommendation system, in. In: Proceedings of the fourth ACM Conference on Recommender Systems, pp. 293–296.

Wu, L., Shah, S., Choi, S., Tiwari, M., Posse, C., 2014. The browsemaps: Collaborative filtering at linkedin. In: RSWeb@ RecSys, Citeseer.

Covington, P., Adams, J., Sargin, E., 2016. Deep neural networks for youtube recommendations, in. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 191–198.

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

Chen, H.-H., Ororbia, I., Alexander, G., Giles, C.L., 2015. Expertseer: a keyphrase based expert recommender for digital libraries, arXiv preprint arXiv:1511.02058.

Felfernig, A., Isak, K., Szabo, K., Zachar, P., 2007. The vita financial services sales support environment. In: Proceedings of the national conference on artificial intelligence. vol. 22. AAAI Press; MIT Press; Menlo Park, CA; Cambridge, MA; London. p. 1692.

Kroenke, K., Spitzer, R.L., 2002. The phq-9: a new depression diagnostic and severity measure. Psychiatric Ann. 32, 509–515.

Zhang, S., Yao, L., Sun, A., Tay, Y., 2019. Deep learning based recommender system: a survey and new perspectives. ACM Comput. Surveys 52, 1–38.

Grbovic, M., Cheng, H., 2018. 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.
Amatriain, X., Basilico, J., 2016. Past, present, and future of recommender systems: an industry perspective, in. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 211–214.
Sar Shalom, O., Koenigstein, N., Paquet, U., Vanchinathan, H.P., 2016. Beyond collaborative filtering: the list recommendation problem, in. In: Proceedings of the 25th international conference on world wide web, pp. 63–72.
Leonhardt, J., Anand, A., Khosla, M., 2018. User fairness in recommender systems. In: Companion Proceedings of the The Web Conference 2018, pp. 101–102.
Ahmed, A., Teo, C.H., Vishwanathan, S., Smola, A., 2012. 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.
Kang, W.-C., Kim, E., Leskovec, J., Rosenberg, C., McAuley, J., 2019. Complete the look: scene-based complementary product recommendation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10532–10541.
Bianchini, D., De Antonellis, V., De Franceschi, N., Melchiori, M., 2017. Prefer: a prescription-based food recommender system. Comput. Stand. Interfaces 54, 64–75.
Johansen, E.S., 2018. Personalized Content Creation using Recommendation Systems (Master's thesis). The University of Bergen.
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). IEEE, pp. 207–216.
Kumar, S., Gupta, M.D., 2019. ĉ+gan: Complementary fashion item recommendation, arXiv preprint arXiv:1906.05596.
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 Trans. Knowl. Data Eng.
Schaal, S., 1997. Learning from demonstration. In: Advances in Neural Information Processing Systems. pp. 1040–1046.
Miller, N.E., Dollard, J., 1941. Social learning and imitation.
Cypher, A., Halbert, D.C., 1993. Watch What I Do: Programming by Demonstration. MIT Press.
Lieberman, H., 2001. Your Wish is My Command: Programming by Example. Morgan Kaufmann.
Sutton, R.S., Barto, A.G., et al., 1998. Introduction to reinforcement learning, vol. 135. MIT Press, Cambridge.
Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529, 484.
Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. pp. 1928–1937.
Harb, J., Precup, D., 2017. Investigating recurrence and eligibility traces in deep q-networks, arXiv preprint arXiv:1704.05495.
Gulwani, S., Harris, W.R., Singh, R., 2012. Spreadsheet data manipulation using examples. Commun. ACM 55, 97–105.
Polozov, O., Gulwani, S., 2015. 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.
Muggleton, S., 1992. Inductive logic programming, 38. Morgan Kaufmann.
Olsson, R., 1995. Inductive functional programming using incremental program transformation. Artif. Intell. 74, 55–81.
Ferri-Ramírez, C., Hernández-Orallo, J., Ramírez-Quintana, M.J., 2001. Incremental learning of functional logic programs. In: International Symposium on Functional and Logic Programming. Springer. pp. 233–247.
Gulwani, S., Hernández-Orallo, J., Kitzelmann, E., Muggleton, S.H., Schmid, U., Zorn, B., 2015. Inductive programming meets the real world. Commun. ACM 58, 90–99.
Hutchins, W.J., Somers, H.L., 1992. An Introduction to Machine Translation, vol. 362. Academic Press London.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112.
Hillel, Y.B., 1964. Language and Information: selected essays on their theory and application. vol. 373. Addison-Wesley Publ. Comp.
Muegge, U., 2006. Fully automatic high quality machine translation of restricted text-a case study. Transl. Comput. 28, 15.
Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks, in. In: Twelfth Annual Conference of the International Speech Communication Association.
Tur, G., De Mori, R., 2011. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. John Wiley & Sons.
Juang, B.-H., Rabiner, L.R., 2005. Automatic Speech Recognition – A Brief History of the Technology Development, 1. Georgia Institute of Technology. Atlanta Rutgers University and the University of California, Santa Barbara, p. 67.
Wallia, C., 1994. Talking and listening to a mac quadra 840 av. Tech. Commun. 41, 130–131.
Tashev, I., Seltzer, M., Ju, Y.-C., Wang, Y.-Y., Acero, A., 2009. Commute ux: Voice enabled in-car infotainment system.
Mansanet, J., Albiol, A., Paredes, R., 2016. Local deep neural networks for gender recognition. Pattern Recogn. Lett. 70, 80–86.
Ko, B.C., 2018. A brief review of facial emotion recognition based on visual information. Sensors 18, 401.
Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.-H., Xiang, Y., He, J., 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors 19, 1863.
Li, X., Da, F., 2012. Efficient 3d face recognition handling facial expression and hair occlusion. Image Vis. Comput. 30, 668–679.
Park, U., Tong, Y., Jain, A.K., 2010. Age-invariant face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 32, 947–954.
Grother, P., Grother, P., Ngan, M., Hanaoka, K., Boehnen, C., Ericson, L., 2017. The 2017 IARPA Face Recognition Prize Challenge (FRPC), US Department of Commerce, National Institute of Standards and Technology.
Grother, P., Ngan, M., Hanaoka, K., 2003. Face recognition vendor test (frvt) part 3: Demographic effects, NIST IR 8280.
Elmahmudi, A., Ugail, H., 2019. Deep face recognition using imperfect facial data. Future Gener. Comput. Syst. 99, 213–225.
Holley, R., 2009. How good can it get? Analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine 15.
Srihari, S.N., Kuebert, E.J., 1997. 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. IEEE. pp. 892–896.
Ptucha, R., Such, F.P., Pillai, S., Brockler, F., Singh, V., Hutkowski, P., 2019. Intelligent character recognition using fully convolutional neural networks. Pattern Recogn. 88, 604–613.
Bai, J., Chen, Z., Feng, B., Xu, B., 2014. Image character recognition using deep convolutional neural network learned from different languages. In: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2560–2564.
Oyedotun, O.K., Olaniyi, E.O., Khashman, A., 2015. Deep learning in character recognition considering pattern invariance constraints. Int. J. Intell. Syst. Appl. 7, 1.
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 Recogn. 58, 190–203.
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.
Yuan, A., Bai, G., Yang, P., Guo, Y., Zhao, X., 2012. Handwritten english word recognition based on convolutional neural networks. In: 2012 International Conference on Frontiers in Handwriting Recognition. IEEE, pp. 207–212.
Acharyya, A., Rakshit, S., Sarkar, R., Basu, S., Nasipuri, M., 2013. Handwritten word recognition using mlp based classifier: a holistic approach. Int. J. Comput. Sci. Issues 10, 422.
Lavrenko, V., Rath, T.M., Manmatha, R., 2004. Holistic word recognition for handwritten historical documents. In: First International Workshop on Document Image Analysis for Libraries. Proceedings. IEEE. pp. 278–287.
Sánchez, J.A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R.M., Vidal, E., de Does, J., 2013. Transcriptorium: a european project on handwritten text recognition. In: Proceedings of the 2013 ACM Symposium on Document Engineering. pp. 227–228.

Granell, E., Romero, V., Martínez-Hinarejos, C.-D., 2019. Image–speech combination for interactive computer assisted transcription of handwritten documents. Comput. Vis. Image Underst. 180, 74–83.
Toselli, A.H., Vidal, E., Puigcerver, J., Noya-García, E., 2019. Probabilistic multi-word spotting in handwritten text images. Pattern Anal. Appl. 22, 23–32.
Boyle, D.K., 2009. Controlling System Costs: Basic and Advanced Scheduling Manuals and Contemporary Issues in Transit Scheduling. vol. 135. Transportation Research Board.
Lifschitz, V., 1999. Action languages, answer sets, and planning. In: The Logic Programming Paradigm. Springer. pp. 357–373.
Matyukhin, V., Shabunin, A., Kuznetsov, N., Takmazian, A., 2017. Rail transport control by combinatorial optimization approach. In: 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT). IEEE, pp. 1–4.
Dai, Z., Liu, X.C., Chen, X., Ma, X., 2020. Joint optimization of scheduling and capacity for mixed traffic with autonomous and human-driven buses: a dynamic programming approach. Transp. Res. Part C: Emerg. Technol. 114, 598–619.
El Hachemi, N., Gendreau, M., Rousseau, L.-M., 2011. A hybrid constraint programming approach to the log-truck scheduling problem. Ann. Oper. Res. 184, 163–178.
Chakroborty, P., Das, A., 2017. Principles of Transportation Engineering. PHI Learning Pvt. Ltd.
Ghoseiri, K., Szidarovszky, F., Asgharpour, M.J., 2004. A multi-objective train scheduling model and solution. Transp. Res. Part B: Methodol. 38, 927–952.
Ingolotti, L., Barber, F., Tormos, P., Lova, A., Salido, M.A., Abril, M., 2004. An efficient method to schedule new trains on a heavily loaded railway network. In: Ibero-American Conference on Artificial Intelligence. Springer. pp. 164–173.
Abril, M., Barber, F., Tormos, P., Lova, A., Ingolotti, L., Salido, M., 2006. A decision support system for railway timetabling (mom): the spanish case. Comput. Railways X: Computer System Design and Operation in the Railway and Other Transit Systems 10, 235.
Feo, T.A., Bard, J.F., 1989. Flight scheduling and maintenance base planning. Manage. Sci. 35, 1415–1432.
Gavish, B., Schweitzer, P., Shlifer, E., 1978. Assigning buses to schedules in a metropolitan area. Comput. Oper. Res. 5, 129–138.
Meng, Q., Wang, S., Andersson, H., Thun, K., 2014. Containership routing and scheduling in liner shipping: overview and future research directions. Transp. Sci. 48, 265–280.
Törnquist, J., 2006. Computer-based decision support for railway traffic scheduling and dispatching: a review of models and algorithms. In: 5th Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS'05). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
Eberlein, X.J., Wilson, N.H., Bernstein, D., 1999. Modeling real-time control strategies in public transit operations. In: Computer-aided Transit Scheduling. Springer, pp. 325–346.
D'Ariano, A., Corman, F., Pacciarelli, D., Pranzo, M., 2008. Reordering and local rerouting strategies to manage train traffic in real time. Transp. Sci. 42, 405–419.
Verderame, P.M., Elia, J.A., Li, J., Floudas, C.A., 2010. Planning and scheduling under uncertainty: a review across multiple sectors. Ind. Eng. Chem. Res. 49, 3993–4017.
Reiners, T., Pahl, J., Maroszek, M., Rettig, C., 2012. Integrated aircraft scheduling problem: an auto-adapting algorithm to find robust aircraft assignments for large flight plans. In: 2012 45th Hawaii International Conference on System Sciences. IEEE. pp. 1267–1276.
Hassold, S., Ceder, A.A., 2014. Public transport vehicle scheduling featuring multiple vehicle types. Transp. Res. Part B: Methodol. 67, 129–143.
Liu, T., Ceder, A., 2016. Synchronization of public transport timetabling with multiple vehicle types. Transp. Res. Rec. 2539, 84–93.
Kelley, R., Tavakkoli, A., King, C., Nicolescu, M., Nicolescu, M., Bebis, G., 2008. Understanding human intentions via hidden markov models in autonomous mobile robots. In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, pp. 367–374.
Siegwart, R., Nourbakhsh, I.R., Scaramuzza, D., 2011. Introduction to Autonomous Mobile Robots. MIT Press.
Smolyanskiy, N., Kamenev, A., Smith, J., Birchfield, S., 2017. Toward low-flying autonomous mav trail navigation using deep neural networks for environmental awareness. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4241–4247.
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, 132–146.
Narla, S.R., 2013. The evolution of connected vehicle technology: from smart drivers to smart cars to...self-driving cars. Ite J. 83, 22–26.
König, M., Neumayr, L., 2017. Users' resistance towards radical innovations: the case of the self-driving car. Transp. Res. Part F 44, 42–52.
Sörme, J., Edwards, T., 2018. A comparison of path planning algorithms for robotic vacuum cleaners.
Bersch, C., Pitzer, B., Kammel, S., 2011. Bimanual robotic cloth manipulation for laundry folding. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 1413–1419.
Miller, S., Van Den Berg, J., Fritz, M., Darrell, T., Goldberg, K., Abbeel, P., 2012. A geometric approach to robotic laundry folding. Int. J. Robot. Res. 31, 249–267.
Estevez, D., Victores, J.G., Fernandez-Fernandez, R., Balaguer, C., 2020. Enabling garment-agnostic laundry tasks for a robot household companion. Robot. Autonomous Syst. 123, 103330.
Walker, N., Peng, Y.-T., Cakmak, M., 2019. Neural semantic parsing with anonymization for command understanding in general-purpose service robots. In: Robot World Cup. Springer. pp. 337–350.
Wooldridge, M., 2009. An Introduction to Multiagent Systems. John Wiley & Sons.
Jonker, C.M., Hindriks, K.V., Wiggers, P., Broekens, J., 2012. Negotiating agents. AI Magazine 33, 79, 79.
Steele, K., Stefánsson, H.O., 2016. Decision theory. In: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, winter 2016 ed. Metaphysics Research Lab, Stanford University.
Myerson, R.B., 2013. Game Theory. Harvard University Press.
Janssen, M., 2002. Complexity and Ecosystem Management: The Theory and Practice of Multi-agent Systems. Edward Elgar Publishing.
Jennings, N.R., Faratin, P., Lomuscio, A.R., Parsons, S., Sierra, C., Wooldridge, M., 2001. Automated negotiation: prospects, methods and challenges. Int. J. Group Decis. Negotiation 10, 199–215.
Baarslag, T., Aydoğan, R., Hindriks, K.V., Fujita, K., Ito, T., Jonker, C.M., 2015. The automated negotiating agents competition, 2010–2015. AI Mag. 36, 115–118.
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 Commun. 11, 5–19.
Wellman, M.P., 2011. Trading agents. Synthesis Lectures Artif. Intell. Mach. Learn. 5, 1–107.
McBurney, P., Parsons, S., 2002. Games that agents play: a formal framework for dialogues between autonomous agents. J. Logic Lang. Inf. 11, 315–334.
Ossowski, S., 2012. Agreement Technologies, vol. 8. Springer Science & Business Media.
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. Prog. Artif. Intell. 1, 277–290.
Parsons, S.D., Gymtrasiewicz, P., Wooldridge, M., 2012. Game Theory and Decision Theory in Agent-based Systems, vol. 5. Springer Science & Business Media.
Perez, J.B., Rodríguez, J.M.C., Mathieu, P., Campbell, A., Ortega, A., Adam, E., Navarro, E.M., Ahrndt, S., Moreno, M.N., Julián, V., 2014. Trends in practical applications of heterogeneous multi-agent systems. The PAAMS Collection, Springer.
Rosenfeld, A., Kraus, S., 2009. Modeling agents through bounded rationality theories. In: Twenty-First International Joint Conference on Artificial Intelligence.
Von Der Osten, F.B., Kirley, M., Miller, T., 2017. The minds of many: opponent modeling in a stochastic game. IJCAI 3845–3851.
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. Springer. pp. 57–80.
Ramchurn, S.D., Vytelingum, P., Rogers, A., Jennings, N.R., 2012. Putting the'smarts' into the smart grid: a grand challenge for artificial intelligence. Commun. ACM 55, 86–97.
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. Springer. pp. 9–17.
Hu, W., Bolivar, A., 2008. Online auctions efficiency: a survey of ebay auctions, in. In: Proceedings of the 17th international conference on World Wide Web, pp. 925–934.

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. Int. Jt. Conf. Artif. Intell.
Baarslag, T., Gerding, E.H., 2015. Optimal incremental preference elicitation during negotiation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence.
Baarslag, T., Kaisers, M., 2017. The value of information in automated negotiation: a decision model for eliciting user preferences, in. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, pp. 391–400.
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.
Sanders, E.B.-N., Stappers, P.J., 2008. Co-creation and the new landscapes of design. Co-design 4, 5–18.
Simonsen, J., Robertson, T., 2012. Routledge International Handbook of Participatory Design. Routledge.
Krasadakis, G., 2017. Artificial intelligence negotiation agent. US Patent App. 15/087,870.
Fatima, S., Kraus, S., Wooldridge, M., 2014. Principles of Automated Negotiation. Cambridge University Press.
Sofy, N., Sarne, D., 2014. Effective deadlock resolution with self-interested partially-rational agents. Ann. Math. Artif. Intell. 72, 225–266.
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. Springer. pp. 103–110.
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., et al., 2019. Machine behaviour. Nature 568, 477–486.
Commission, E., 2018. Digital transformation monitor: the rise of virtual personal assistants. European Commission Report.
Hoy, M.B., 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. Med. Reference Services Q. 37, 81–88.
Turing, A., 1950. Computing machinery and intelligence. Mind 59, 433.
Kubrick, S., Clarke, A., 1968. Screenplay for 2001: A Space Odyssey.
Huang, H.-M., Pavek, K., Novak, B., Albus, J., Messin, E., 2005. A framework for autonomy levels for unmanned systems (alfus). In: Proceedings of the AUVSI's Unmanned Systems North America, pp. 849–863.
Messina, E., Jacoff, A., 2006. Performance standards for urban search and rescue robots. In: Unmanned Systems Technology VIII. vol. 6230. International Society for Optics and Photonics. p. 62301V.
Marvel, J.A., Saidi, K., Eastman, R., Hong, T., Cheok, G., Messina, E., 2012. Technology readiness levels for randomized bin picking. In: Proceedings of the Workshop on Performance Metrics for Intelligent Systems, pp. 109–113.
Adams, S.S., Banavar, G., Campbell, M., 2016. I-athlon: towards a multidimensional turing test. AI Magazine 37, 78–84. https://doi.org/10.1609/aimag.v37i1.2643. URL: https://wvvw.aaai.org/ojs/index.php/aimagazine/article/view/2643.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al., 2017a. Mastering the game of go without human knowledge. Nature 550, 354–359.
Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al., 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, arXiv preprint arXiv:1712.01815.
Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al., 2019. Mastering atari, go, chess and shogi by planning with a learned model, arXiv preprint arXiv:1911.08265.
Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B.S., Martínez-Plumed, F., et al., 2017. Mapping intelligence: Requirements and possibilities. In: 3rd Conference on Philosophy and Theory of Artificial Intelligence. Springer. pp. 117–135.
Martínez-Plumed, F., Gómez, E., Hernández-Orallo, J., 2020. Tracking ai: The capability is (not) near. In: Proceedings of the Twenty-fourth European Conference on Artificial Intelligence. IOS Press. pp. 2915–2916.
Martínez-Plumed, F., Gómez, E., Hernández-Orallo, J., 2020. Tracking the evolution of ai: The aicollaboratory. In: Proceedings of the 1st International Workshop: Evaluating Progress in Artificial Intelligence (EPAI 2020).
Hernández-Orallo, J., Martınez-Plumed, F., Avin, S., Whittlestone, J., Seán, O., 2020. Ai paradigms and ai safety: mapping artefacts and techniques to safety issues. In: Proceedings of the Twenty-fourth European Conference on Artificial Intelligence. pp. 2521–2528.
Sun, S., Zhai, Y., Shen, B., Chen, Y., 2020. Newspaper coverage of artificial intelligence: a perspective of emerging technologies. Telematics Inform. 101433.
Hernández-Orallo, J., Baroni, M., Bieger, J., Chmait, N., Dowe, D.L., Hofmann, K., Martínez-Plumed, F., Strannegård, C., Thórisson, K.R., 2017. A new ai evaluation cosmos: ready to play the game? AI Magazine 38, 66–69.
Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M., 2013. The arcade learning environment: an evaluation platform for general agents. J. Artif. Intell. Res. 47, 253–279.
Schlangen, D., 2019. Language tasks and language games: on methodology in current natural language processing research, arXiv preprint arXiv:1908.10747.
Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y., 2019. Hellaswag: can a machine really finish your sentence? arXiv preprint arXiv:1905.07830.
Krizhevsky, A., 2009. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/kriz/cifar.html.
Rajpurkar, P., Jia, R., Liang, P., 2018. Know what you don't know: Unanswerable questions for squad, arXiv preprint arXiv:1806.03822.
Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2019. Superglue: a stickier benchmark for general-purpose language understanding systems, arXiv preprint arXiv:1905.00537.
Vinyals, O., et al., 2017. Starcraft II: a new challenge for reinforcement learning, arXiv preprint arXiv:1708.04782.
Machado, M.C., Bellemare, M.G., Talvitie, E., Veness, J., Hausknecht, M., Bowling, M., 2018. Revisiting the arcade learning environment: evaluation protocols and open problems for general agents. J. Artif. Intell. Res. 61, 523–562.
Purves, C., Cangea, C., Veličković, P., 2019. The playstation reinforcement learning environment (psxle), arXiv preprint arXiv:1912.06101.
Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
Linden, A., Fenn, J., 2003. Understanding Gartner's hype cycles, Strategic Analysis Report No R-20-1971. Gartner, Inc 88.
Steinert, M., Leifer, L., 2010. Scrutinizing Gartner's hype cycle approach. In: Picmet 2010 Technology Management for Global Economic Growth. IEEE. pp. 1–13.
Sager, I., 2012. Before iphone and android came simon, the first smartphone. Bloomberg Businessweek 29.
Wilensky, R., 1987. The berkeley unix consultant project. In: Wissensbasierte Systeme. Springer. pp. 286–296.
Herzog, O., Rollinger, C.-R., 1991. Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence: Final Report on the IBM Germany LILOG-Project. Springer.
Wolfram, S., 2009. Wolfram— alpha, On the WWW. URL: http://www.wolframalpha.com.
Franz, A., Milch, B., 2002. Searching the web by voice. In: Proceedings of the 19th international conference on Computational linguistics-vol. 2. Association for Computational Linguistics. pp. 1–5.
Ferrucci, D.A., 2012. Introduction to "this is watson". IBM J. Res. Dev. 56, 1, 1.
Müller, V.C., Bostrom, N., 2016. Future progress in artificial intelligence: a survey of expert opinion. In: Fundamental Issues of Artificial Intelligence. Springer. pp. 555–572.
Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O., 2018. When will ai exceed human performance? Evidence from ai experts. J. Artif. Intell. Res. 62, 729–754.
Betz, U.A., Betz, F., Kim, R., Monks, B., Phillips, F., 2019. Surveying the future of science, technology and business–a 35 year perspective. Technol. Forecast. Soc. Chang. 144, 137–147.
Avin, S., 2019. Exploring artificial intelligence futures. J. AI Humanities. Available at 10.17863/CAM 35812.
Kurzweil, R., 2005. The Singularity is Near: When Humans Transcend Biology. Penguin.
Dafoe, A., 2018. Ai Governance: A Research Agenda, Governance of AI Program, Future of Humanity Institute. University of Oxford, Oxford, UK.
Marcus, G., 2020. The next decade in ai: four steps towards robust artificial intelligence, arXiv preprint arXiv:2002.06177.

Schilling, M.A., 1998. Technological lockout: an integrative model of the economic and strategic factors driving technology success and failure. Acad. Manage. Rev. 23, 267–284.
Veletsianos, G., 2007. Cognitive and affective benefits of an animated pedagogical agent: considering contextual relevance and aesthetics. J. Educ. Comput. Res. 36, 373–377.
Borghans, L., Ter Weel, B., 2006. The division of labour, worker organisation, and technological change. Econ. J. 116, F45–F72.
Suri, T., 2011. Selection and comparative advantage in technology adoption. Econometrica 79, 159–209.
Sohn, K., Kwon, O., 2020. Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. Telematics Inform. 47, 101324.
H.-L. E. G. on Artificial Intelligence, 2019. A definition of artificial intelligence: main capabilities and scientific disciplines.
Fuchs, J., Heller, I., Tolpilsky, M., Inbar, M., 1999. Cadet, a computer-based clinical decision support system for early cancer detection. Cancer Detect. Prevent. 23, 78.
Langley, P., 1996. Elements of Machine Learning. Morgan Kaufmann.
Flach, P., 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press.
Alpaydin, E., 2020. Introduction to Machine Learning. MIT Press.
Clark, A., Fox, C., Lappin, S., 2013. The Handbook of Computational Linguistics and Natural Language Processing. John Wiley & Sons.
Goldberg, Y., 2017. Neural network methods for natural language processing. Synthesis Lect. Human Lang. Technol. 10, 1–309.
Szeliski, R., 2010. Computer Vision: Algorithms and Applications. Springer Science & Business Media.
Gold, B., Morgan, N., Ellis, D., 2011. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley & Sons.
Kautz, H., Walser, J.P., 2000. Integer optimization models of ai planning problems. Knowl. Eng. Rev. 15, 101–117.
Segler, M.H., Preuss, M., Waller, M.P., 2018. Planning chemical syntheses with deep neural networks and symbolic ai. Nature 555, 604–610.
Spyropoulos, C.D., 2000. Ai planning and scheduling in the medical hospital environment.
Brown, N., Sandholm, T., 2019. Superhuman ai for multiplayer poker. Science 365, 885–890.
Murphy, R.R., 2019. Introduction to AI Robotics. MIT Press.
Shoham, Y., Leyton-Brown, K., 2008. Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations. Cambridge University Press.
Chowdhury, S.R., Rodríguez, C., Daniel, F., Casati, F., 2010. Wisdom-aware computing: on the interactive recommendation of composition knowledge. In: International Conference on Service-Oriented Computing. Springer. pp. 144–155.