Information Fusion 103 (2024) 102135

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance

Isaac Triguero [a,d,*], Daniel Molina [a], Javier Poyatos [a], Javier Del Ser [b,c], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain
[b] TECNALIA, Basque Research & Technology Alliance (BRTA), Derio, 48160, Spain
[c] University of the Basque Country (UPV/EHU), Bilbao, 48013, Spain
[d] School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, United Kingdom

ARTICLE INFO

Keywords:
General-purpose AI
Meta-learning
Reinforcement learning
Neuroevolution
Few-shot learning
AutoML
Transfer learning
Generative AI
Large language models

ABSTRACT

Most applications of Artificial Intelligence (AI) are designed for a confined and specific task. However, there are many scenarios that call for a more general AI, capable of solving a wide array of tasks without being specifically designed for them. The term General Purpose Artificial Intelligence Systems (GPAIS) has been defined to refer to these AI systems. To date, the possibility of an Artificial General Intelligence, powerful enough to perform any intellectual task as if it were human, or even improve it, has remained an aspiration, fiction, and considered a risk for our society. Whilst we might still be far from achieving that, GPAIS is a reality and sitting at the forefront of AI research.

This work discusses existing definitions for GPAIS and proposes a new definition that allows for a gradual differentiation among types of GPAIS according to their properties and limitations. We distinguish between closed-world and open-world GPAIS, characterising their degree of autonomy and ability based on several factors such as adaptation to new tasks, competence in domains not intentionally trained for, ability to learn from few data, or proactive acknowledgement of their own limitations. We then propose a taxonomy of approaches to realise GPAIS, describing research trends such as the use of AI techniques to improve another AI (commonly referred to as *AI-powered AI*) or (single) foundation models. As a prime example, we delve into generative AI (GenAI), aligning them with the terms and concepts presented in the taxonomy. Similarly, we explore the challenges and prospects of multi-modality, which involves fusing various types of data sources to expand the capabilities of GPAIS. Through the proposed definition and taxonomy, our aim is to facilitate research collaboration across different areas that are tackling general purpose tasks, as they share many common aspects. Finally, with the goal of providing a holistic view of GPAIS, we discuss the current state of GPAIS, its prospects, implications for our society, and the need for regulation and governance of GPAIS to ensure their responsible and trustworthy development.

1. Introduction

The recent advances in Large Language Models (LLMs) [1], such as ChatGPT, may be perceived as a step towards getting to Artificial General Intelligence (AGI) [2], in which a machine could think for itself, matching or exceeding human capabilities. This has created a lot of hype and fears about the development of AI [3]. While these models seem to be able to perform some tasks which they were not directly trained for, it is yet unclear whether such models do display emergent abilities on unseen tasks [4–6]. The intelligence of current AI systems significantly differs from human intelligence [7], as they may be lacking other abilities (e.g. complex reasoning, consciousness, or initiative, among others), but they possess others (e.g. assembling patterns) that may be great to supporting humans in complex tasks. In the literature, the term AGI is often used but ill-defined [8]. This has made it difficult to say how far we are currently from realising AGI, if at all possible [9,10].

Actually, thus far, the most extensive use of AI models today falls within those that are capable of performing confined and specific tasks

* Corresponding author at: Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain.
E-mail addresses: triguero@decsai.ugr.es (I. Triguero), dmolina@decsai.ugr.es (D. Molina), jpoyatosamador@ugr.es (J. Poyatos), javier.delser@tecnalia.com (J. Del Ser), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.inffus.2023.102135
Received 26 July 2023; Received in revised form 3 November 2023; Accepted 6 November 2023
Available online 8 November 2023

(also known as narrow or fixed-purpose AI). Typically, these models are manually designed by experts, following numerous steps from data collection, through data modelling, to deployment, in a pipeline referred to as the Machine Learning (ML) Lifecycle [11]. These models are necessary and are being established in many areas of our society with a high volume of data, where designing and exploiting AI algorithms by experts is doable. Unfortunately, these models usually lack the generalisation abilities to perform well on unseen tasks, so that, their practical application remains bound to the tasks it was trained for. Nevertheless, numerous General-Purpose AI systems (GPAIS[1]) have recently emerged to solve more than one task and to be able to generalise to unseen tasks with very few data (or even none) and/or adjustments.

While AGI remains an aspiration, the term GPAIS can be regarded as a more modest and realistic expectation of AI, which would not expect other abilities inherent to humans, as mentioned above. In the best-case scenario, GPAIS would also be capable of dealing with tasks without being directly programmed to do so, as soon as it has seen data that could help it solve them. To enable GPAIS, one option is to add a new layer of abstraction that would use an AI algorithm to design or enrich another AI algorithm. In a nutshell, we *construct* or *enhance* AI with an additional AI stage. A common example of this is to make AI learn to work like an AI expert, determining which algorithms and/or components are most suitable for a given problem [12]. Alternatively, we may aim to create a single model capable of exploiting large amounts of data for multiple tasks, as multi-task learning [13] does, or include a fine-tuning step to adapt to new tasks, as foundation models implement [14]. In the literature, the term **foundation models** is frequently employed interchangeably with GPAIS. Similarly, with the growing prevalence of LLMs, there is a tendency to conflate the concept of generative AI (GenAI) with that of general purpose AI. In this work, we contend that while these terms are integral to GPAIS, they do not encompass their entirety.

Among others, one may expect a GPAIS to be able to transfer knowledge from similar tasks, learn with as little data as possible, and rapidly adapt to changes and/or new tasks. These properties are typically needed in a wide range of research areas such as AutoML [15], few-shot learning [16], or continual learning [17], some of which are solved by meta-learning [18], reinforcement learning [19] or evolutionary computation [20]. To regulate and manage the risks of generally capable AIs [21,22], different authors have recently defined what a GPAIS may be [8,23]. Apart from some level of disagreement across definitions, those works provided high-level definitions without landing their definitions based on the existing research.

The goal of this paper is to become a primer for researchers and practitioners interested in a holistic vision of GPAIS, mapping out the desiderata of a GPAIS and describing different approaches to build those models. Our analysis tackles the main aspects of GPAIS:

- *Proposing a new definition together with its properties*: First, we will discuss what a GPAIS is, looking in detail at what is considered to be a GPAIS in the literature and how it has evolved, analysing four previous definitions. Then, we articulate the differences between GPAIS, traditional fixed-purpose AI, and AGI to provide a complete definition and categorisation of GPAIS. We will describe the most relevant features of GPAIS, considering different levels of ability and autonomy, and distinguishing between various degrees of GPAIS. Taking into consideration the variable time, we differentiate between closed-world GPAIS and open-world GPAIS.
- *Breaking it down onto a taxonomy*: We devise a taxonomy of approaches to building GPAIS, describing some of the key research trends and problems in which GPAIS are being developed. We identify how these approaches may fit within the proposed definition and the capabilities they may provide. This includes GPAIS which use a single model to generalise to many tasks and AI-powered AI approaches.
- *Linking the proposed taxonomy with trending topics in AI: GenAI and multi-modality*. First, we delve into GenAI [24] as the first type of GPAIS perceived as general intelligence by the public. Next, we talk about the development of GPAIS capable of exploiting (and fusing) multiple types of data sources (e.g. image, text, audio) to learn a more cohesive representation of concepts in the same fashion that humans do.
- *Providing a detailed discussion around GPAIS*: We debate the state of GPAIS and their challenges, including their limitations and prospects, and explore how some of the approaches identified in the proposed taxonomy could help advance this field. Finally, we reflect on the impact of GPAIS on our society and emphasise the urgent need for trustworthy and sustainable systems together with appropriate regulations that dictate how these models must be audited and accounted for. We claim that an understanding of the design principles behind these techniques might aid in establishing appropriate regulations and necessary governance for the safe deployment of GPAIS.

As depicted in Fig. 1, the remainder of this paper is organised as follows. Section 2 covers the existing definitions for GPAIS, and provides our definition and categorisation as well as the expected properties of a GPAIS. Section 3 presents the proposed taxonomy of approaches to creating GPAIS. Section 4 focuses on generative AI models as the most notable example of GPAIS nowadays. Section 5 elaborates on the use of multi-modality in emerging GPAIS. Section 6 explores the present status of GPAIS, their challenges ahead, societal implications and the need for regulation and governance. Finally, Section 7 concludes the paper with a summary of our contribution.

2. General-purpose artificial intelligence systems: Definitions and properties

This section shows and discusses the existing definitions of GPAIS in the literature (Section 2.1). After that, we establish a new definition and categorise GPAIS together with their key properties (Section 2.2).

2.1. Definitions related to general purpose artificial intelligence systems in the literature

Many have described what a GPAIS may entail, but very few formal definitions exist. We briefly present and analyse recent formal definitions of the concept of GPAIS, or closely related terms, that exist in the literature, most of which revolve around legislation. The risks about the potential of AI, and more recently about GPAIS, have promoted various definitions of what these systems may be and what their capabilities are. Our goal is not to track down all possible definitions of GPAIS, but to identify the most relevant ones and their differences.

In terms of regulation and legislation, the AI Act of the European Union is the first proposed law on AI in the world. The first article referring to GPAIS defines them as follows:

Definition 1 (*Article 3(1b) AI Act (December 6, 2022) [25]*)**.** Intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI system.

This definition has been heavily criticised [26], as it is overly inclusive, so that, simple image or speech recognition systems would qualify as GPAIS. Within this context, the Future of Life Institute put forward some recommendations to shape the different articles of the proposed European law. They put forward the following definition of GPAIS:

[1] Throughout the document, GPAIS may represent either a singular system or multiple systems depending on the context.

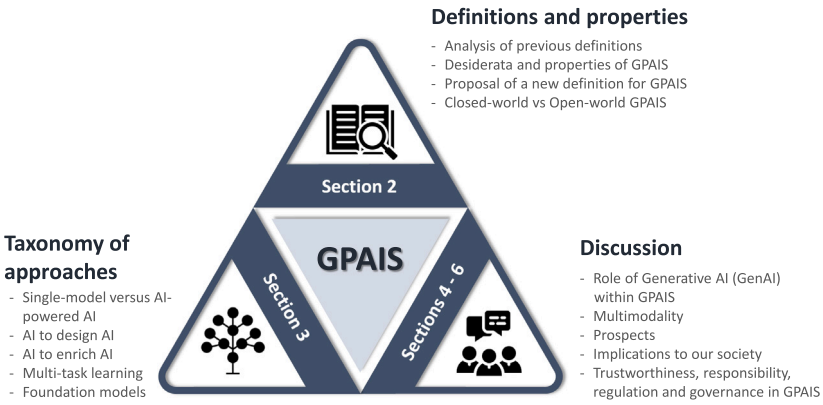

Fig. 1. Schematic diagram representing the contributions of this work and their distribution over the sections of the manuscript.

Definition 2 ([27]). 'General purpose AI system' means an AI system that is able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc, and is able to have multiple intended and unintended purposes.

The most relevant aspects of this definition include the idea of performing many tasks, and the ability of these systems to perform tasks they were not directly trained for. In addition, the Future of Life Institute indicates that GPAIS are characterised by their scale (large memory, abundant data, and powerful hardware) as well as their reliance on transfer learning (applying knowledge from one task to another). Moreover, they also highlight that GPAIS are not limited to a single type of information input, that is, they are multi-modal.

With regard to the use of transfer learning, they refer to the concept of foundation models coined in [14]. Foundation models are based on standard deep learning and transfer learning, and they are defined as a model that "*is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks*". In this work, we argue that GPAIS goes beyond transfer learning and deep learning, and therefore, foundation models become a subset of GPAIS.

In line with the previous definition, and also in the context of the AI Act, Gutierrez et al. presented in [8] another formal definition for GPAIS. This definition states:

Definition 3 ([8]). An AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained.

As such, the definition is somewhat equivalent to the previous one, emphasising the possibility of completing tasks outside of those it is specifically trained for. Prior to the definition, the authors claimed that GPAIS resembles the idea of AGI, but their definition does not specify the differences between GPAIS and AGI. The authors also mentioned that GPAIS may have different degrees of autonomy, with or without human intervention. They also mentioned that GPAIS may be trained in different manners (e.g. Gato [28] uses supervised learning, whereas MuZero [29] is based on reinforcement learning).

Expanding on this definition, [23] recently suggested some edits to Definition 3, yielding:

Definition 4 ([23]). An AI system that can accomplish a range of distinct *valuable* tasks, including some for which it was not specifically trained.

Specifically, they deleted the words "be adapted to" and "intentionally", and added the word "valuable" to the definition. The nature of those edits is all about the safety and risk-based frameworks that can be used to mitigate the challenges of GPAIS. For example, they considered that the ability of a model to be adapted to accomplish a task provides too many degrees of freedom. The authors exemplified this with the minimal risks associated with a fine-tuned BERT model [30]; according to [23], a specialised translation model would not easily be able to do any other tasks other than translation. Whilst we agree on the safety aspect covered in this work, narrowing the definition of GPAIS down to the riskiest models would not help advance the development of the different research areas that encompass GPAIS.

Definitions 2 and 3 are very much focused on the adaptation to perform tasks of different types, while Definition 4 is centred on the most generally capable models, neglecting simpler or more primitive methods. Thus, we find Definitions 2 and 3 suitable to generally speak about GPAIS. However, differently from the existing definitions, our focus is not on the legislation of GPAIS, but on the way they are designed, the problems they can solve, and the challenges they are to address to continue progressing. Therefore, the main distinct points to approach a new definition for GPAIS are:

- Current definitions speak about AGI and GPAIS interchangeably. We consider AGI a much more pretentious goal, in which a machine would have the autonomy of a human being. Conversely, GPAIS is a more practical term that allows us to consider that an AI could fulfil the term at different levels of generality.
- GPAIS may follow other approaches that are not pure transfer learning only. Thus, we deem foundation methods as an important category within GPAIS, but other strategies exist which do not explicitly use broad data prior to a fine-tuning step. An example of that could be AutoML-zero, which aims to automatically discover novel AI algorithms using a set of tasks without following the pre-training preceded by fine-tuning approach of foundation models.
- We categorise GPAIS, so that we may understand their degree of autonomy and expected capabilities. We consider that the proposed categorisation helps provide a global view, highlighting the role that different techniques can play for these systems.

2.2. Definition and properties of general purpose artificial intelligence systems

With the goal of providing a definition of GPAIS based on current research and articulating its differences with respect to traditional fixed-purpose AI and futuristic AGI, this section describes the main properties and functionalities one would expect for GPAIS, distinguishing different levels of ability and autonomy. Fig. 2 represents the transition from fixed-purpose AI to AGI.

Let T be an AI/ML task, for which there is some data available and can be modelled in various ways depending on the expected outcome. For example, a classical ML task could consist of classifying images to distinguish among different types of objects (e.g. *car* vs *motorbike*).

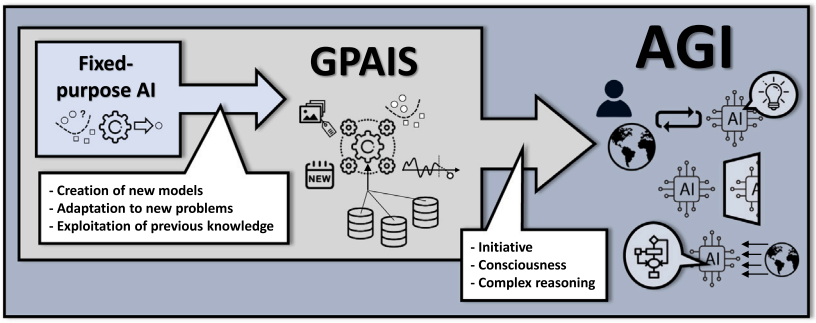

Fig. 2. From Narrow AI (fixed-purpose) to AGI.

It is important to note that different ML tasks could be derived from the same source of data. A **fixed-purpose AI system** would consider a single task at a time, and would normally require having sufficient data to train a model. To solve a task, experts would typically follow the ML lifecycle, considering various strategies to learn effectively from the data (e.g. finding a suitable pipeline of preprocessing and learning algorithms, together with the tuning of their hyper-parameters). Having said that, the methodology proposed by an expert could be further applied to other tasks of different nature, as soon as they can be modelled within the same ML setting (e.g. as a single-output classification problem). For example, it is common to apply the same classification technique to various datasets with very minor modifications (e.g. some hyper-parameter tuning). Nevertheless, each task is tackled in a holistic way, training with enough data, and without any interaction/knowledge extracted from other tasks.

When shifting towards GPAIS, we simultaneously consider N tasks $(T_1, \ldots, T_N)$. Thus, a GPAIS is expected to be able to solve one or more problems. While intuitively this means a more complex learning environment, dealing with a variety of tasks may allow us to exploit synergies among them (if they are sufficiently related). A classical example of this is multi-task learning [13], which aims to improve the performance on multiple tasks by leveraging shared representations and knowledge transfer between them.

Apart from dealing with more than one task at a time, to further characterise GPAIS let us involve time in the definition. Let t be the current point in time in which we are given a set of tasks $(T_1, \ldots, T_N)$. In GPAIS, a new set of K tasks $(T_{N+1}, \ldots, T_{N+K})$ may arise at a future time $t+\Delta t$, for which we do not necessarily expect to have much data to learn from. Within this setting, we differentiate between **closed-world GPAIS** and **open-world GPAIS:**

- **Closed-world GPAIS**: In this type of GPAIS, we assume that we have data for a fixed number of tasks at a time t, and those will always be the only tasks to solve. It assumes that all the tasks that an AI system will encounter are predetermined and accounted for during the training phase. In a nutshell, we consider closed-world as the simplest form of GPAIS, which is capable of dealing with more than one task. However, these systems may lack the ability to generalise to novel tasks outside their training scope. Therefore, we would not expect a closed-world GPAIS to accomplish any task that it was not directly trained to do. If a new task arrived at the time $t+\Delta t$, the whole system would require to be retrained in order to perform well.
- **Open-world GPAIS**: This type of GPAIS refers to a scenario where the AI system operates in a more dynamic and evolving environment, encountering new tasks at a time $t+\Delta t$ and data that was not necessarily included at the time t (namely, in its initial training set). Unlike closed-world GPAIS, open-world GPAIS acknowledges the presence of unknown and unforeseen tasks that may arise over time. This requires the AI system to possess a degree of generalisation, flexibility, and the ability to learn from limited or scarce data. To do so, a GPAIS would typically have to exploit what it learned previously at time t to adapt faster to new tasks. Similar to the way human perception is inherently limited about what they know of the world, GPAIS operating in open-world scenarios also possess a substantial but constrained understanding of their context. This knowledge of the unknown could be treated as "meta-information" or knowledge about the environment, which could be leveraged to develop adaptation strategies to proactively confront the emergence of new tasks.

Therefore, open-world GPAIS provides a higher degree of generality compared to closed-world. Looking at the current progress in the literature, we identify systems with various properties in terms of what they can or cannot do, or how they respond to the end user during certain tasks. In what follows, we make a few remarks with respect to data availability in both closed and open worlds, and the performance on new tasks:

- *Data availability*: Although it may not always be the case, in a closed-world GPAIS system we would normally expect sufficient data to train models on all the tasks. That does not imply that data augmentation and other pre-processing techniques may not be needed. However, in an open-world GPAIS, data scarcity may become the norm for new tasks that arise at time $t+\Delta t$. This is not always the case as, for example, many AutoML techniques usually assume enough data at time $t+\Delta t$, where an ML model is actually created. Further discussion on this is provided in Section 3.
- *Performance on new tasks*: When asked to perform new tasks, an open-world GPAIS may display difficulties while trying to solve them, but may not be aware of having such a difficulty. For example, in LLMs, the system may hallucinate and output untrue statements with high confidence [31], or in computer vision, a generative model may produce images resembling objects that contain unrealistic elements [32]. While hallucination in AI is not yet fully understood, several factors play an important role, including lack of data or biased training data [33]. Conversely, we might also encounter a situation in which the system would successfully (yet serendipitously) perform tasks for which it was not intentionally and specifically trained (the so-called *emergent* abilities) [4]. These two issues raise concerns about the trust we may place on this kind of general-purpose system. Section 6 discusses in more depth the implications of these yet to be solved issues of open-world GPAIS.

Fig. 3 shows a graphical representation of the differences between closed-world and open-world GPAIS, focusing on the time difference in which they work on, and noting some of the potential properties and characteristics of these systems. Note that some of those characteristics are not compulsory to qualify as open-world or closed-world GPAIS. For example, not all methods will have emerging abilities, or would be able to work well under data scarcity, but they may still be classed as

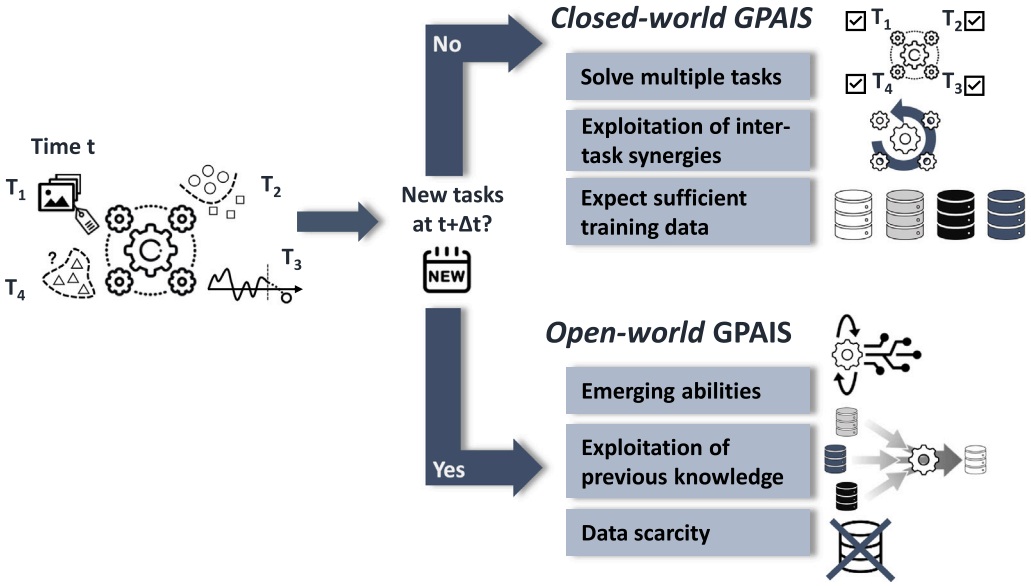

Fig. 3. Closed-world vs. Open-world GPAIS: Some of their potential properties and characteristics.

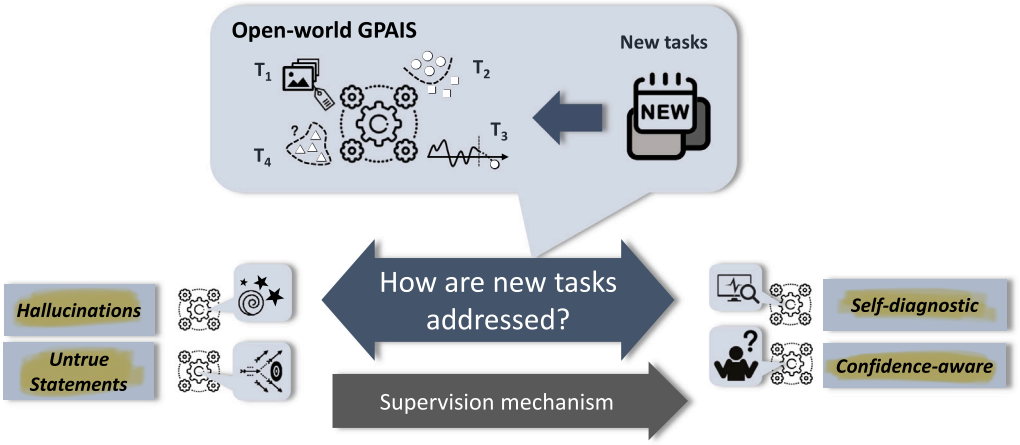

Fig. 4. Open-world GPAIS: challenges and potential solutions to become advanced models.

open-world. The key distinguishing feature between closed-world and open-world lies in the expectation of new tasks at time $t + \Delta t$.

For a GPAIS to become advanced, we would expect them to respond more positively to new tasks, since they either have the information to perform the task in question or recognise that they do not know how to perform it. Advanced GPAIS would be expected to self-diagnose themselves, be aware of the confidence in their responses and ask (or look) for feedback and improve themselves within an active learning framework [34] (See Section 3.3). To do so, a supervision mechanism could be introduced, so that, it learns to distinguish between what is correct and what is not. To date, none of the existing GPAIS provide such a level of autonomy. Fig. 4 illustrates the challenges posed by new tasks in an open-world setting, and how a supervision mechanism may help boost their confidence and capabilities.

The final step in the evolution of these systems is to know under what conditions an advanced GPAIS would become an AGI system. An AGI system is characterised by being able to perform any task efficiently and as intelligently as a human being. It is precisely this quality of being human that advanced GPAIS need to become AGI systems since they must be able to do more complex reasoning, exploit causality, have a curiosity to explore new tasks, or have a conscience of themselves, among other abilities.

The properties and expected functionalities we have described before for GPAIS will be used to present a formal definition for GPAIS that considers different degrees of autonomy and ability:

Definition 5 (*GPAIS*)**.** A General Purpose Artificial Intelligence System (GPAIS) refers to an advanced AI system capable of effectively performing a range of distinct tasks. Its degree of autonomy and ability is determined by several key characteristics, including the capacity to adapt or perform well on new tasks that arise at a future time, the demonstration of competence in domains for which it was not intentionally and specifically trained, the ability to learn from limited data, and the proactive acknowledgement of its own limitations in order to enhance its performance.

This definition fills the gap left by the definitions presented in the previous section, distinguishing between various degrees and properties of GPAIS. This definition may serve as a guide to transforming an existing AI system into a GPAIS.

3. General purpose artificial intelligence systems: A multidimensional taxonomy

The above definitions allowed us to distinguish between various degrees of GPAI. The aim of this section is to present a taxonomy

of approaches to realise GPAIS of any kind, describing some of the key research trends and problems in which GPAIS are being developed. Although we will highlight how existing solutions fit within our definition and categorisation of GPAIS, it is important to early note that there may be some degree of overlapping between categories, and certain strategies can facilitate the transition of methods across categories (e.g. adding some elements to evolve a typically closed-world AI system into an open-world one). It should also be stressed that in this work, our primary target is on the ML aspect of AI, which encompasses other research areas such as optimisation or simulation.

3.1. Breaking it down onto a multidimensional taxonomy

With the aim of making AI more self-sufficient and capable of learning without human intervention, researchers in the AI community have followed many strategies to provide generalisation abilities. Broadly speaking, we observe two distinct approaches in the literature: **AI-powered AI** vs **single-model** approaches.

- **AI-powered AI**: We can consider an additional layer of abstraction that would use an AI algorithm to either design or enrich another AI algorithm. As an example, the hybridisation of ML techniques and optimisation algorithms, by interacting with each other or themselves [35], has recurrently been exploited to foster generalisation (among other objectives). In essence, we are boosting the performance and robustness of an underlying AI model via another AI technique. Throughout the following subsections, we will discuss the different ways in which AI may power other AI systems, categorising AI-powered AI as per their objective: designing an AI algorithm or helping/enriching an AI algorithm to learn/perform better.
- **Single-model**: On the other hand, not all the advances in GPAIS would always need to use an extra AI model to help generalise. Instead, their generalisation abilities come from learning from various tasks and/or vast amounts of data. We identify two relevant research areas that at their original definition would typically use a single AI model, namely, multi-task learning and foundation models. It is important to clarify that while we categorise these techniques differently from the AI-powered AI ones, that does not mean they offer fewer abilities, nor could they be combined with the above in many cases.

Fig. 5 shows the proposed multidimensional taxonomy for different categories of GPAIS, approaches and research areas that we discuss in the following subsections. Sections 3.2 and 3.3 summarise some of the ways in which AI can be used to make another AI system more autonomous and general to design AI algorithms and to enrich other AI models, respectively. Sections 3.4 and 3.5 cover multi-task learning and foundation models, respectively, as prevalent examples of single AI models that generalise to multiple tasks, but do not necessarily involve an additional AI layer.

3.2. AI to design AI

To solve a task with an AI system, experts are typically confronted with various design decisions that range from customising the hyper-parameters of a well-known algorithm, selecting an appropriate method and its hyper-parameters, to designing the components of – or even the entire – AI algorithm. It is common for the stages of the ML Lifecycle to become somewhat repetitive, so experts often base their decision on their own experience with previous problems. AI algorithms can be used to help with all of those design decisions, extracting general knowledge about how to implement an ML process easier and faster within a set of ML tasks.

- **Hyper-parameter optimisation**: Determining the best hyper-parameters for a given task is of extreme importance to achieve good results. While traditional hyper-parameter tuning strategies have frequently been used to tackle a specific dataset and that would not constitute general-purpose, they can also be used to tune hyper-parameters for related datasets [36]. To do this, meta-heuristics and Bayesian optimisation are well-known examples of commonly used techniques for black-box function optimisation [37]. Among others, general frameworks such as Optuna [38] or Hyperopt [39] greatly facilitate the search for adequate hyper-parameters.
  In GPAI, we could find the best set of hyper-parameters for a number of tasks, or even determine which of them work best for different tasks, and use that knowledge for when a new task arises. Thus, this can be seen as exploiting previous knowledge and generalising beyond the original tasks over which hyper-parameter tuning was done. Nevertheless, if the new task has very little data, adapting/fine-tuning those hyper-parameters could become very challenging.
- **Automated algorithm selection**: As there are a wide variety of AI algorithms, selecting the best model(s) is also critical in solving a task. In the world of GPAIS, the previous hyper-parameters optimisation is often coupled together with the automated selection of an AI algorithm (also known as the combined algorithm selection and hyper-parameter problem [40] in ML). Depending on the kind of AI we aim to further automate, e.g. ML or optimisation, we may find different terminologies to refer to this kind of strategy. In ML, they are typically referred to as **AutoML** [12], while in optimisation, they would usually be known as hyper-heuristics [41]. In ML, it is common for the AutoML to optimise an entire pipeline, involving the automated selection of the best combination of a pre-processing technique (e.g. one-hot encoding, dimensionality reduction, or feature selection, among others), as well as an ML algorithm [42]. In the area of reinforcement learning, the idea of AutoML may also be considered to automate the reinforcement learning pipeline with some specific challenges (e.g. non-stationarity, environment and algorithm design, or the generation of complex and diverse behaviours) [43].
  Many authors talk about AutoML [44] when the aim is to find the best pipeline for a deep learning approaching, covering, feature engineering, hyper-parameter optimisation and neural architecture search [45,46]. In this field, metaheuristics have made a good contribution through the use of evolutionary algorithms, aiming to both reduce the complexity of the deep learning model and obtain better results [47,48]. As an alternative to finding the best pipeline, [49] has recently shown that a transformer [50] can be used to perform classification without designing explicitly a machine learning pipeline.
  While the common objective of these techniques is to take the human out of learning/optimisation process [51], different methods may be only capable of dealing with closed-world GPAI, and others may work in open-world scenarios. Some may use previous knowledge available at time t (e.g. those based on meta-learning), whereas others may focus on adaptation to the new task, aiming to find the best algorithm at time $t + \Delta t$. As happened before for hyper-parameter tuning, when tackling open-world tasks, current automated algorithm selection techniques would normally work under the assumption that sufficient data is still provided to create a model for the new task. One could argue, however, that a meta-learning approach could be viable to find the best algorithm and hyper-parameters within a zero-shot learning approach, which assumes that some meta-information about the new task is available and that it can be related to previously known tasks. Similar to the way auto-sklearn works [42], this could be in the form of meta-features for the new dataset which describes its main properties.

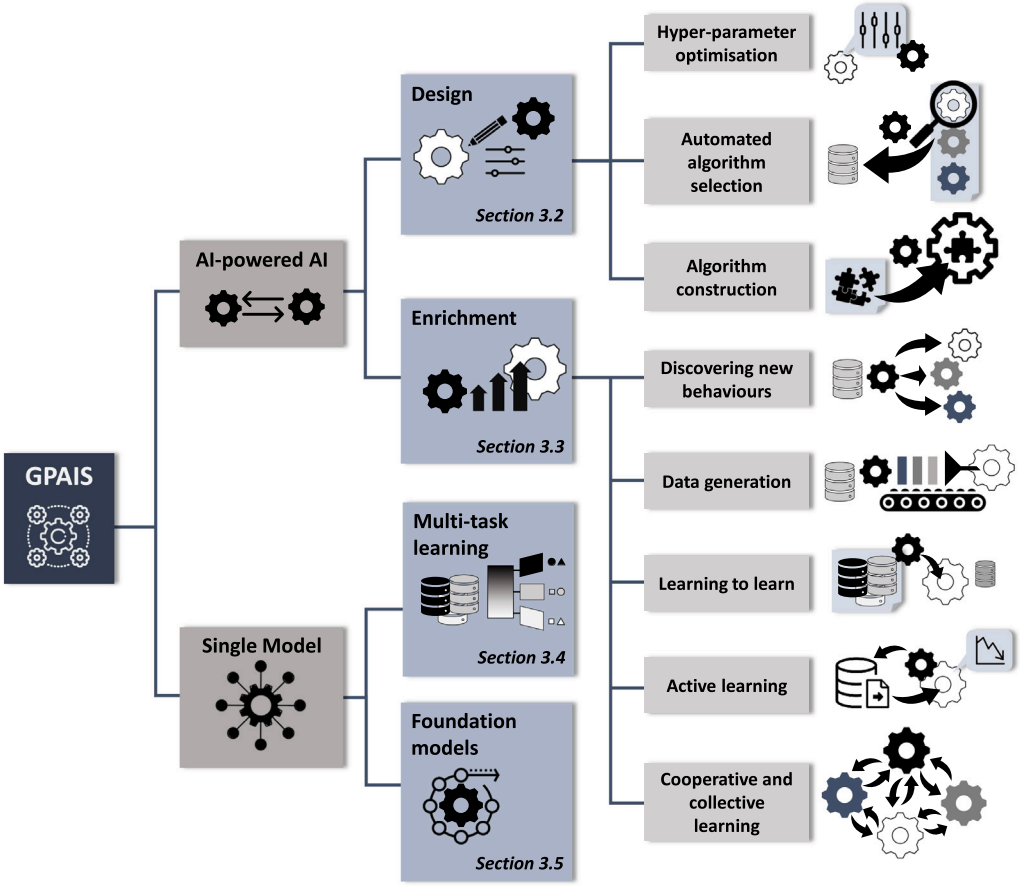

Fig. 5. Taxonomy of approaches for GPAIS.

- **Algorithm construction**: The previous two approaches are focused on the configuration of existing algorithms. However, in addition to these approaches, various AI strategies have been proposed to go beyond algorithm configuration and actually construct new algorithms from scratch or determine the necessary low-level components for a given task.
  AutoML-zero [15] and Hyper-heuristics [52] are two examples of AI-based strategies to generate new algorithms for ML and optimisation, respectively. Neuroevolution [20] is also an alternative to designing entire deep learning algorithms from scratch. AutoML-zero explores the space of possible algorithms and their configurations using search algorithms. This strategy enables the creation of novel algorithms tailored to specific tasks or problem domains. Current developments are focused on designing algorithms for a set of tasks within a specific domain, which would class as closed-world within the proposed definition. However, there is potential in the future to become open-world, by evolving the generated algorithms in new tasks, or by formulating new objective functions in existing algorithmic construction networks that seek generalisation over unseen tasks.

3.3. AI to enrich AI

Instead of focusing solely on the design of AI algorithms, AI is often coupled with other AI techniques to enrich or help with its learning process. By harnessing the power of various AI techniques and their collaborative potential, we can empower AI systems to become more robust and capable intelligent systems. In this context, this section identifies some of the most relevant areas in which AI may help make another AI to be more general-purpose and adaptable.

- **Discovering new behaviours**: One of the challenges for an AI is to dynamically adapt to changes in the environment, such as shifts in the underlying data distribution. **Continual learning** [17,53] is an ML paradigm that involves learning a model able to solve several tasks from a continuous stream of data, without forgetting knowledge obtained from the preceding tasks. Data related to the old tasks are not available during the training of the new tasks, forcing the adaptability of the system and the preservation and retrieval of valuable knowledge over time. By leveraging other AI techniques, we can explore innovative approaches to discovering new behaviours. Research on data stream mining has largely addressed concept drift detection and adaptation methods [54] to help identify instances where the data distribution has changed significantly, and to efficiently adapt the knowledge captured by the model to such changes. Thus, it can be viewed as an indicator of new behaviours or patterns in the data. Other ways to facilitate discovering and adapting, avoiding catastrophic forgetting, involve reinforcement learning and meta-learning [55]. For instance, reinforcement learning [56] could help explore and learn from interactions with the environment, allowing for the emergence of novel and adaptive behaviours that could help the agent performs better in varying tasks and/or new environments. Another area in optimisation research with ample potential to discover emerging novel behaviours in data-based models is quality-diversity optimisation [57,58]. Techniques belonging to this research area permit to efficiently explore complex search spaces by simultaneously accounting for the quality (fitness) and diversity (similarity) of their improvised solutions. When applied to the optimisation of the parameters of a given model (as in Deep Learning), quality-diversity optimisation can

yield diversified GPAIS that, when combined together, may effectively cope with unseen modelling tasks in few- or even zero-shot open-world settings, especially when such new tasks are not radically different from the ones providing the training data for the GPAIS. These sorts of GPAIS techniques may belong to both the closed-world and open-world settings depending on their abilities. On the one hand, when the tasks themselves do not really change, but the distribution of the data does (i.e. a new concept emerges), it would be classified as closed-world. On the other hand, they could also belong to the open-world setting when they are capable of dealing with entirely new tasks [59], as in robotic tasks in which the action space changes over time.

- **Data generation**: One of the cornerstones in AI to generalise well is the lack of data that could resemble what might happen at time $t + \Delta t$. With the emergence of GenAI, we find generative models [24] that are capable of creating data that follows the distribution of a given training dataset, achieving unprecedented scales and levels of fidelity over complex data modalities. This makes it possible to simulate and generate potential data samples (and environments in a reinforcement learning context). This opens up new avenues for improving AI's ability to generalise well by augmenting the existing dataset and expanding the range of scenarios the model can learn from. Additionally and linked to the previous challenge, creating synthetic data can facilitate the discovery of new patterns, behaviours, or anomalies that may not have been explicitly present in the original training, further enhancing the AI's capacity to generalise effectively. In the context of reinforcement learning, Paired Open-Ended Trailblazer (POET) [60] is a prime example of environment generation to improve the generality of a model in potential open-world scenarios. In [61], the author highlights that learning how to automatically generate effective training data is one of the most important pillars to realise general purpose intelligence, and it is currently the most underdeveloped area. More about generative models will be discussed in Section 4.
- **Learning to learn**: Having many (sometimes related) tasks to deal with and the lack of data have motivated the research on AI systems that learn how to learn. As mentioned before, meta-learning [18] finds applications in the design of AI (e.g. hyper-parameter optimisation and AutoML), but it is also applicable to other scenarios in which it is designed to help other AI models learn. The ways in which meta-learning can help other AI methods may range from learning an efficient optimisation algorithm for the sake of faster convergence and better performance [62], which would not class a general purpose system, to learning how to transfer knowledge effectively from a set of tasks to a new task [63]. In GPAIS, **few-shot learning** [16] is a prime example of beneficiaries of meta-learning approaches. The challenge in the few-shot regime is to learn effectively (without overfitting) from an extremely limited number of labelled examples (or even none, known as *zero-shot* learning [64]). Although few-shot learning can be tackled without meta-learning [65,66], many methods do use it to mimic human learning. These methods extract meta-knowledge from a collection of few-shot learning tasks, and then transfer this meta-knowledge to unseen few-shot learning tasks comprising novel categories. In this way, there are two distinct AI models at play: the meta-learning model and the base model. The meta-learning model is responsible for acquiring meta-knowledge from multiple tasks, while the base model leverages that meta-knowledge to perform few-shot learning on new tasks. Thus, the majority of the methods that follow this learning-to-learn paradigm may naturally class as open-world.
- **Active learning**: Another of the weaknesses of AI is to realise when they are wrong and seek human intervention. Active learning [67] has traditionally been used to reduce annotation effort by intelligently selecting the most informative instances for labelling. Next, the algorithm being trained asks an *oracle* (which is usually human assistance) to label new data that may help to improve its performance, with the goal of maximising the performance of the algorithm with the least amount of data. This idea could potentially be very useful in GPAIS when the demand for new annotated data is autonomously decided by the model itself [68]. Thus, it can also facilitate adaptability by enabling an AI system to actively seek guidance from humans or oracles in unfamiliar or uncertain situations. It allows the AI system to acquire new knowledge and fine-tune its performance to different tasks, enhancing its generalisation capabilities. This has the potential to develop advanced open-world GPAIS that are closer to AGI. To realise an effective active learning approach, GPAIS should have sufficient reasoning tools to identify the need for guidance. In this context, knowledge graphs and semantic databases could provide the understanding needed to query for the right data labelling [69].
- **Cooperative and collective learning**: To mitigate the weaknesses of different AI systems alone, we could get them to help each other to constitute a larger, general purpose system. Collaborative AI refers to the concept of multiple AI agents working together, often in a coordinated manner, to achieve common goals or solve complex problems. While this is not a new research area [70], in the context of GPAIS, collaborative AI agents/models can be employed to enhance the capabilities and performance of the system. These agents could specialise in different tasks or domains and collaborate with each other to provide a more comprehensive and versatile solution. For example, within a GPAIS, different AI models could focus on tasks such as natural language understanding, image recognition, recommendation systems, or decision-making, and they can exchange information or cooperate to address complex queries or problems.

  Related to the idea of collaboration between AI systems, we may find **Federated Learning** [71], as a decentralised ML approach that enables multiple devices or nodes to collectively train a shared model without sharing their raw data. In the field of robotics, the concept of **Swarm Intelligence** [72] has gained attention for its potential in achieving general intelligence [73]. Swarm Intelligence involves the emulation of collective behaviours observed in social insects, where individual agents interact locally and make decisions based on simple rules or local information. The majority of existing federated learning or swarm intelligence methods would not class as GPAIS as they are usually developed to perform a single task. Nevertheless, the ability of these distributed AI models is not limited to fixed-purposed models, and GPAIS with other attributes such as being capable of dealing with multiple tasks and/or continual learning [74] could be built using a decentralised approach. Therefore, the categorisation of a cooperative and collective learning approach as closed-world or open-world depends on the features of the underlying model.

3.4. Multi-task learning

Multi-task learning is an ML technique that allows a model to be trained simultaneously on different tasks, exploiting the synergies between them. Consequently, knowledge and representation are shared across all tasks with the goal of maximising model performance [13]. When learning from different domains of data, a more robust behaviour for new data is expected [75,76], even against adversary attacks [77].

However, multi-task models may face a higher risk of error and lower efficiency in tackling all tasks compared to models trained individually for each task. This problem has been studied through different approaches, as shown in [13,78]. Although there have been certain techniques that have proven to be good options such as metaheuristics and Reinforcement Learning [79], currently deep learning models like

Gato [28] and the more recent Meta-transformer [80] have shown to be the most successful approaches for multi-task learning, especially with different data modalities and diverse learning problems.

Traditional multi-task learning algorithms would typically fall under the closed-world GPAIS category because a multi-task system is trained in a stationary environment, i.e. it is trained to solve a set of tasks at a time t. However, it is not usually conceived to tackle new tasks. If a new task emerged, data related to that task would have to be added. This would normally imply retraining the model as a whole, although other strategies are possible. These strategies to enable multi-task learning in an open-world setting would require a good amount of data for the new task in order to perform well. In contradistinction to retraining the model, we could follow a pre-training plus fine-tuning (as in [81]) to make a multi-task learning model perform well in the open world, giving rise to the next category discussed in the next section.

3.5. Foundation models

As discussed previously, the underlying idea of foundation models is that a single model is trained on a very large amount of data (typically for multiple tasks), and in a later stage (time $t+\Delta t$), such a model can be fine-tuned to tackle new tasks. As described in [14], these models are based on **deep learning** and **transfer learning**. These methods have become very popular in LLMs, demonstrating strong general purpose abilities to generate human-like text, which will be the main topic of the next section.

It is important to clarify that both meta-learning and transfer learning involve leveraging knowledge from previous tasks. However, meta-learning focuses on developing learning algorithms or models that can learn how to learn, adapt, and generalise, while transfer learning aims to transfer specific knowledge or representations from a source task to improve performance on a target task. Although the purpose does not directly relate to learning to learn, the training and adaptation framework of foundation models facilitates the separation of concerns. For example, in the context of natural language processing, the training process of an LLM like ChatGPT may involve exposure to a vast range of text sources, such as books, articles, and websites. During this training process, the model learns various aspects of language, including grammar, syntax, semantics, and common word usage. This meta-knowledge acquired during training allows the model to understand and generate coherent and contextually appropriate text.

One of the keys to successfully training foundation models lies in how to use vast amounts of unlabelled data effectively. Semi-supervised learning – and more prominently, self-supervision – have been key research areas to deliver high-quality foundation models. Self-supervised learning [82] allows us to reduce the necessity for labels by creating its own labelled data and solving auxiliary (pretext) tasks to extract knowledge therefrom. For example, in LLMs, it is common to mask a word out in the text and predict the surrounding words. This permits us to model the relationships among consecutive words without the need for external labels. These relationships can later be used for downstream tasks, such as text generation or translation to other languages.

Once the foundation model is trained, it can be adapted to perform specific tasks or domains. For instance, it can be fine-tuned on a dataset of movie reviews to learn how to classify sentiments (e.g., positive or negative) expressed in film reviews. The adaptation process involves providing task-specific data and adjusting the model's parameters to make it more specialised and accurate for the particular task at hand, retaining as much previous knowledge as was found to be valuable for the specific task/domain under the target.

Foundation models are inherently open-world, but they may have some limitations. For example, they require enough quality data for the adaptation phase to be successful. When the amount of data is low, meta-learning comes into play. We could leverage a foundation model to quickly adapt to new tasks with limited data. Even with the addition of meta-learning to tackle new tasks with very small amounts of data, we would not class foundation models as advanced GPAIS. For them to become advanced models, they should be hybridised with other AI models, such as active learning.

4. A closer look into generative AI: a kind of foundation models in GPAIS

As mentioned in the introduction, modern GenAI models like ChatGPT are the first AI models that the general public has begun to recognise as GPAIS. Therefore, due to their relevance, and to align them with the terms and concepts presented in the taxonomy of Section 3, we now provide a more in-depth vision of these approaches.

As commented above, GenAI models [24] are an outstanding example of how AI may help enrich other AI systems. ***Generative modelling*** or ***GenAI*** consists of a set of algorithms designed to learn the distribution of a dataset, so that, its underlying patterns can be characterised, and samples that resemble the original data can be generated [83]. Queries for new data instances can be performed unconditionally or conditioned on a query to produce samples with specific traits (e.g. class belongingness, the induction of high-level properties on the output, or a descriptive prompt, among many other possibilities).

In the literature, one of the earliest generative models that we may encounter are generative adversarial networks (GANs) [84], which have showcased a great ability to generate highly specific data, including realistic images [32] and medical images [85]. However, by design GANs are only trained to improve a specific type of results, because the generation process is evaluated/trained with a discriminator for a very specific task [86]. As such, they cannot be considered enablers for general purpose intelligence, but they could help improve AI systems generating additional data to learn as described in Section 3.

Recently, various AI models have been proposed that, differently from GANs, are able to generate a wide variety of images, ranging from paintings to photo-realistic images, based on textual descriptions. At their core, these text-to-image models are diffusion probabilistic models or diffusion models (DMs) that capture a higher-level semantic meaning of a group of images [87], so they are considered generative models. LLMs are also considered generative models because they are trained using many texts for which they learn the probability distribution over its vocabulary. Once trained, LLMs are able to generate content similar to original texts, so they seem to have been written by a human. In essence, both DMs and LLMs are **Foundations Models** as they are trained on large quantities of unlabelled text containing up to trillions of tokens (for LLMs) or millions of pictures (for DMs), using self-supervised or semi-supervised learning. These models, in contrast to GANs, are not trained for a very specific task, and they are able to solve many problems [24], so they can be considered GPAIS. It is important to highlight that DMs may not necessarily be GPAIS if their focus is merely on generating images, and if that is their only task. LLMs, however, can easily be used for multiple tasks (e.g. translation, summarising, etc.).

Some of the most relevant DMs include DALL-E [88,89], Stable Diffusion models (as 2 [87] or XL [90]), and the commercial tool Midjourney. These models have demonstrated their capability to create professional images solely from textual descriptions, making them advanced tools for image generation. Their creations are of such high quality that experts might even consider using them to win an art competition.[2]

Currently, the more relevant generative models in languages are Bidirectional Encoder Representations from Transformers (BERT) [30] and Generative Pre-trained Transformer (GPT). In terms of architecture, BERT is a bidirectional model that can consider both the preceding and succeeding text when generating output. It consists of encoder and decoder components. In contrast, GPT is a unidirectional model that generates subsequent text based on the context and preceding text. Although BERT can be used for question-answering tasks, particularly

[2] https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html (Last access: 2023/10/28).

for specific questions, it was not primarily designed for generating free-form text with coherent and contextually relevant content, unlike GPT models. Therefore, we will focus on these latest models. GPT models are able to generate human-like text based on prompts given to them. They can participate in conversations, answer questions, provide explanations, and assist with various language-related tasks. GPT has the ability to adjust its output based on the context of the conversation, enabling it to provide more accurate answers. The following LLMs (in all their versions) share similarities with GPT, but they differ in terms of the underlying models used for text generation and the training processes employed.

The most relevant LLMs are ChatGPT [91], GPT-4 [92] (both proposed by OpenAI in collaboration with Microsoft), LLAMA 1 [93] (proposed by Meta) which has produced a list of improved models with additional fine-tuning like Vicuna [6,94]), LLAMA 2 [95] or Google AI Bard [96]. These models achieve unprecedented performance due to a two-phase approach: a training process involving a large amount of textual information (obtained from the web and other resources, some of which are not specified), and a reinforcement learning stage involving human input. This combination enables AI models to generate text that closely resembles human-authored writing. LLAMA 1 lacks the aforementioned reinforcement learning phase, resulting in less natural output. To address this limitation, tuned models have been developed, such as Vicuna. However, in the case of Vicuna, reinforcement learning was conducted by a GPT model instead of humans due to resource constraints. The use of an LLM as ChatGPT to evaluate the output of other LLMs can be considered an interesting case of an AI model to enrich AI, although some authors do not recommend this evaluation methodology [97]. LLAMA 2, however, has a version tuned for chat conversations, which has been coined as LLAMA-2 Chat [95].

Unfortunately, the actual GPT-4, ChatGPT or Bard models are not really available to the end user. Therefore, it is not possible to have a copy of the models to use them locally or to adapt them using fine-tuning. The only way to use them is through a remote interface that enables users to interact with the AI model hosted on a remote and private server. LLAMA 1 and their fine-tuned version are available for local use in various sizes, but they are not permitted for commercial purposes. Nevertheless, there is an increasing number of models that allow it, like the new LLAMA 2 [95] (for a limited number of active users), Falcon LLM [98], or Dolly, from Databricks [99].

One of the most notable features of these generative models is their ability to synthesise information from various domains and their capacity to apply knowledge and skills in diverse contexts and disciplines. Some models, such as GPT-4, have demonstrated a certain level of intelligence [6]. First, these models have shown a high level of proficiency in various domains such as literature, mathematics, programming law, and others. This enables their use not only for general tasks but also for highly specialised ones. However, these multi-domain models coexist with other specific models for some domains of special interest. As an example, Github CoPilot is an AI model specifically trained to be capable of generating source code, used as an advanced development tool that can improve the productivity of developers [100].

Another significant feature of these generative AI models is their multi-modality, which enables them to combine multiple modalities and generate outputs that are more diverse and nuanced. This implies the ability to combine output and relationships between different formats such as text, images, audio, and more. For example, a multi-modal AI model can generate both a description of an image and an image based on text input. Furthermore, if it is trained on both text and audio, it can generate speech based on text input or generate text from speech input. This feature indicates that these generative AI models are not limited to a single type of information, thereby enhancing their potential to solve complex tasks. The next section will provide more information about multimodality in GPAIS.

As foundation models, generative models are part of the open-world category, typically requiring enough data and a fine-tuning step to adapt to new tasks. Due to their striking performance and assertiveness, they can easily be confused with human intelligence [101]. However, these models may *hallucinate*, providing very confident responses that do not seem to be justified by its training data, hence they could be completely false. The main issue with hallucinations is that the outputs provided by a model may sound plausible but are either factually incorrect or even unrelated to the given context, raising concerns about how to trust these models and ethical implications [102]. This problem can be even more severe in cases where generative models are used to produce training data to learn other models [103]. These techniques are specially interesting to generate synthetic medical images [104–106], for privacy and anonymity reasons [106,107]. However, due to the severity of the medical errors, the danger of hallucinations may be greater, requiring specialist supervision to minimise it. One may also argue that hallucinations may be related to speculation and therefore similar to how humans reason in case of a lack of information. Additionally, there is a debate whether this kind of model shows emergent abilities [4–6], meaning that they are capable of solving tasks that they were not explicitly trained for. Both hallucinations and emergent abilities are two current hot topics in LLMs and DMs (See Section 6).

A very interesting and promising approach is the use of LLM-powered autonomous agent systems, in which LLMs could be used as an intelligent tool to divide a problem using task decomposition [108]. They use several LLM agents to solve different partial problems or even decision-making tasks [109], like AutoGPT[3] or HuggingGTP [110] or MetaGPT [111], or in which the AI model is able to interactively ask more questions, like GTP-Engineer.[4] This cooperative approach, which aligns with the potential of *cooperative and collective learning* discussed in Section 3.3, has already shown interesting results in complex domains like chemistry [112].

5. Multi-modality in GPAIS

A significant challenge for real-world systems is the integration of data coming from multiple sources, called *data fusion* [113]. AI has extensively been used for this task [114]. Traditionally, this fusion implies merging different data from the same kind of source, for instance, medical images resulting from several tests should be merged when related to the same organs [115,116]. However, the variety of data available for a given problem is rapidly increasing, including images, text, audio, etc [117]. This multi-modality of data types provides an opportunity to GPAIS, which can learn a more cohesive representation of a concept by having multiple "views", very much in the same way that humans do. For example, a human learns what a cat is (i.e. the concept) by seeing, hearing, touching and even smelling it. GPAIS that only use a single source of data may be losing an important part of the concept (e.g. LLMs would only know what a cat is from what it has "read"). Conversely, having such a variety of data poses a challenge to learning effectively.

In recent years, there has been a lot of effort in using AI models capable of processing multimodal data [117,118]. Modern deep learning models have also been researched in the context of multi-modality [119]. A good example of multi-modality in GPAIS is Gato [28], a general purpose agent that may use text and images. Recent models, like graph deep learning models [120,121], can represent more structured data, enabling AI systems to tackle new prospects and applications [122].

In prompt-to-image GenAI models, there is always a certain multi-modality, because they receive some text as input and it must create an image. Consequently, two distinct media are involved. For LLMs, although they achieve great success only by working with text, the

[3] https://github.com/Significant-Gravitas/Auto-GPT (Last access: 2023/10/28).

[4] https://github.com/AntonOsika/gpt-engineer (Last access: 2023/10/28).

ability to process images could improve their functionality and, in particular, their abilities to communicate with the user (e.g. answering questions like describing a picture). Nowadays, this functionality has been recently added to ChatGPT, and is expected to improve the potential usages of this technology (e.g. to help people with vision impairment [123,124] or even improve the training of other generative models, like DALL-E 3 [89]). However, the real challenge with multi-modal models is the integration of information coming from different sources about the same concept in a cohesive format, in a way that allows an AI model to have a more general view of the target concept. For instance, joining visual recognition (recognising an object) and conversational commands (processing it) can be very useful to improve the interaction with people [125] and even in robotic environments [126].

Multi-modality as a way to improve AI models is the approach for Google's new advancement in AI, spearheaded by their proposal called *Generalised Multimodal Intelligence Network Interface model* (Gemini AI).[5] Gemini AI has been designed to be versatile, to be able to tackle diverse tasks and to learn from a myriad of domains without the current constraints. To this end, Gemini embodies an interconnected network composed of individual modular ML models, which are trained on specific tasks. The different modules will produce varied outputs, and the Gemini AI encoders will transform these diverse data forms into a cohesive and common format. Then, decoders will produce outputs in diverse modalities, contingent on the received encoded inputs and the task on focus. It is expected to greatly improve current multimodal AI models.

Gemini AI was not available at the time of writing this work. However, it is clear that Google has decided to focus its strategy on incorporating improved multi-modality in their AI models as a way to outperform current state-of-the-art GenAI models. This movement suggests that for renowned experts (it is the collaboration work of well-known AI labs, such as DeepMind and Google Brain) this could be a strategic field that could significantly expand the realm of possibilities for future AI models.

6. A discussion on the prospects, implications and regulation and governance of GPAIS

The previous sections have aimed to provide a definition for GPAIS considering various degrees of autonomy (Section 2) and a taxonomy of methods to build these systems (Section 3). In this section, we briefly discuss on the current state of GPAIS prospects and limitations, the implications of it in our society, and the need for regulation and governance.

6.1. Current status and prospects

As discussed before, LLMs are currently the most prominent GPAIS we may find that display some degree of general intelligence. While they are certainly the most outstanding ones, LLMs are not the sole means of generalisation that have been extensively researched in the literature, which could significantly contribute to the development of a GPAIS. Our taxonomy offers a comprehensive perspective, encompassing various approaches and methods employed in building these systems. However, the proposed taxonomy is not intended to establish a sharp separation across methods, and we may find cases in which some AI systems may belong to more than one category at once, which could be desirable, as we will discuss next. While many considerations are being made in different research niches, we hope that our taxonomy facilitates a more holistic and global approach to GPAIS development.

To advance in the field of GPAIS, it is crucial to explore open-world approaches that enable systems to operate in dynamic and unfamiliar environments, where new tasks are dealt with limited data, exploiting as much as possible previous knowledge. In doing so, we advocate for considering not only very successful approaches such as foundation models, but also hybridising them with some of the other strategies discussed in this article. For example, foundation models such as ChatGPT have already rendered exceptional performance in complex tasks, which could be enhanced if other approaches such as *AI to enrich AI* methods are used in conjunction with them. From the approaches discussed in Section 3, we highlight *continual learning* (i.e. models capable of discovering new behaviours in the data), learning to learn, and active learning as some of the most prominent approaches that could be used in combination with foundation models.

By incorporating these approaches into GPAIS, we can enhance their adaptability, generalisation capabilities, and learning efficiency in the presence of new learning tasks:

- Continual learning can enable GPAIS to keep up with varying data distributions, while meta-learning can endow them with the ability to quickly adapt to new tasks and domains.
- Active learning and reinforcement learning empower GPAIS to actively seek information and learn from their interactions with the environment. Proactive measures such as acknowledging limitations and asking for help, or finding ways to look for additional sources of information, would help GPAIS to become more autonomous.

Together, these techniques pave the way for more advanced and versatile GPAIS that can address complex real-world challenges effectively. For current GPAIS to evolve beyond their limitations, we must explore their potential for developing emergent abilities and addressing inherent challenges. These two aspects are calling for further research with the aim of understanding whether emergent abilities are real and explaining why they occur, together with means to identify and control hallucinations.

6.2. Implications to our society

While we are not yet at the stage of achieving AGI, there are significant technical and ethical challenges that must be addressed before reaching that milestone. Current GPAIS offerings already provide numerous functionalities, but they also come with certain risks. The ability of these systems to process vast amounts of data and make automated decisions raises concerns regarding privacy, ethics, and fairness in their implementation.

Under the umbrella term of *Trustworthy AI*, recent research efforts in AI are focused on developing techniques and frameworks that enhance oversight, transparency, interpretability, and robustness, among others, with the aim of designing responsible AI systems [102]. Those efforts become even more needed when developing GPAIS. By gaining a deeper understanding of the underlying mechanisms and limitations of models qualifying as such, we can refine the design and training processes to minimise the occurrence of hallucinations, and ultimately maximise the emergence of desirable and reliable behaviours in GPAIS. This ongoing exploration and mitigation of emergent abilities and hallucinations will contribute to the responsible development and deployment of GPAIS in the future.

Another important aspect that must be taken into consideration in GPAIS is their sustainability. The training process of current GPAIS like ChatGPT exhibits an unprecedented demand for computing resources [127]. Future GPAIS generations may imply even higher environmental impact in terms of carbon footprint [128]. This is calling for more sustainable and greener approaches [129] that reduce the computational cost of GPAIS, before they can be massively adopted.

One of the main challenges in this regard continues to be the evaluation of the results and explanations of an AI model, which has been a long-lasting issue. The community should continue re-thinking these issues [130], as these metrics are the only means to determine secure and effective use. Evaluating and explaining the results of a

[5] https://www.gemini-ai.org/ (Last access: 2023/10/28).

GPAIS, particularly in the open-world setting [131], may prove to become a very challenging task, especially when they combine multiple (potentially black-box) AI models and use very broad sources of information. Nevertheless, generative AI models may help generate better interpretations, explanations, and reasoning over them [132,133].

In summary, trustworthy AI technologies, including human oversight, transparency, interpretability, and robustness, among others, must contribute to the effective management of AI risks and address emerging GPAIS prospects, ultimately reducing the uncertainty and concerns of the society about the use of these modern AI systems.

6.3. Regulation and governance in GPAIS

Vivid discussions on the regulation of GPAIS, foundation models, and GenAI are taking place within the community as advances in GPAIS are continually taking place. While regulatory efforts for closed-world GPAIS may be more manageable, the open-world setting poses unresolved difficulties, as the specific tasks to be tackled can be largely unknown while the model is audited. This uncertainty makes it difficult to anticipate potential outcomes, even with valid metrics in place. The definition of GPAIS proposed in [23] aimed to limit the set of AI models that would class as GPAIS, to those that may display emergent abilities, as they are more likely to inflict risks. Nevertheless, there are other challenges in the open-world setting, such as hallucinations or untrue statements/outputs, which call for an external audit process that would help us trust the behaviour of such an autonomous system.

The recent advancements in GPAIS have raised concerns regarding the need for prompt regulatory measures to ensure their safe deployment in society (See Section 6.5 in [102] for a brief discussion). However, different countries are taking divergent approaches in this regard.

The EU AI Act[6] envisions a distinct regulatory framework compared to the proposals under consideration in the United Kingdom. In [134], an analysis was conducted to assess the compliance of the latest LLMs with the proposed EU AI Act, revealing their non-compliance. Developing standards to support the AI Act or any other regulatory framework will be faced with the task of specifying the current best practices in trustworthy and responsible AI [135].

The recent EU AI Act discussion on GPAIS and foundations models[7] considers that they should guarantee robust protection of fundamental rights, health and safety and the environment, democracy and rule of law. They should assess and mitigate risks, comply with design, information and environmental requirements and register in the EU database. Generative foundation models, like GPT, should comply with additional transparency requirements, such as disclosing that the content was generated by AI, designing the model to prevent it from generating illegal content, and publishing summaries of copyrighted data used for training.

Regulation is always associated to the auditability and accountability during its design, development, and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used, to design a responsible AI system [102]. Auditability refers to a property sought for the AI-based system, which may require transparency (e.g., explainability methods, traceability), measures to guarantee technical robustness, etc. The auditability of a responsible AI system may not necessarily cover all requirements for trustworthy AI, but rather those foretold by ethics, regulation, specifications and protocol testing adapted to the application sector (i.e., vertical regulation).

Accountability is another crucial aspect to consider in regulation. Determining who is responsible for the outputs of a GPAIS, particularly in cases of unexpected emergent abilities, becomes essential when decisions made thereof bring about fatal consequences. From a scientific point of view, understanding the principles of design of these techniques, how they can be built, and their properties and limitations can help prescribe the regulatory directives that should be put in place for GPAIS. Even in the current developmental status of GPAIS, many ethical and legal aspects of their practical use remain without consensus, providing ample space for further debate.

A recent "Policy Brief" on AI risk management standards for GPAIS and foundation models has been adopted by the UC Berkeley Center for Long-Term Cybersecurity (CLTC).[8] It highlights key policy implications of the profile, as well as the considerations of what AI risk-related policies would be especially valuable beyond the profile. They recommend employing the following three strategies as they seek to regulate GPAIS, foundations models, and GenAI.

1. "Ensure that developers of GPAIS, foundation models, and generative AI adhere to appropriate AI risk management standards and guidance.
2. Ensure that GPAIS, foundation models, and generative AI undergo sufficient pre-release evaluations to identify and mitigate risks of severe harm, including for open source or downloadable releases of models that cannot be made unavailable after release.
3. Ensure that AI regulations and enforcement agencies provide sufficient oversight and penalties for non-compliance".

These discussions are directly linked to the importance of managing AI risks, as it is discussed in [136][9]: "We must anticipate the amplification of ongoing harms, as well as novel risks, and prepare for the largest risks well before they materialise. Climate change has taken decades to be acknowledged and confronted; for AI, decades could be too long".

Together with the discussion on the societal-scale risk and the technical developments and to advance towards trustworthy AI technology to develop responsible AI systems, another important aspect is pointed out in [137,138] and other recent publications on governance and the need for governance measures: "For AI systems with hazardous capabilities, we need a combination of governance mechanisms matched to the magnitude of their risks. Regulators should create national and international safety standards that depend on model capabilities. They should also hold frontier AI developers and owners legally accountable for harms from their models that can be reasonably foreseen and prevented. These measures can prevent harm and create much-needed incentives to invest in safety. Further measures are needed for exceptionally capable future AI systems, such as models that could circumvent human control".

Paying attention to these necessary measures, we highlight a report published by the UC Berkeley Center for Long-Term Cybersecurity (CLTC)[10] that aims to help organisations develop and deploy more trustworthy AI technologies, including 150 properties related to one of seven "characteristics of trustworthiness" as defined in the NIST AI RMF[11]: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful biases managed. Using these characteristics as a starting point, the CLTC report names 150 properties of trustworthiness, which are mapped to particular parts of the AI lifecycle where they are likely

[6] https://artificialintelligenceact.eu/ (Last access: 2023/10/28).

[7] https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence (Last access: 2023/10/28).

[8] https://cltc.berkeley.edu/publication/policy-brief-on-ai-risk-management-standards-for-general-purpose-ai-systems-gpais-and-foundation-models/ (Last access: 2023/10/28).

[9] https://managing-ai-risks.com/ (Last access: 2023/10/28).

[10] https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/ (Last access: 2023/10/28).

[11] https://www.nist.gov/itl/ai-risk-management-framework (Last access: 2023/10/28).

to be particularly critical. Each property is also mapped to specific parts of the AI RMF core, guiding readers to the sections of the NIST framework that offer the most relevant resources.

In summary, governance frameworks together with regulation and trustworthiness technologies must be developed cooperatively to ensure that an AI system in general – and a GPAIS in particular – is designed and engineered to achieve its goals, while: a) maintaining the ability to disengage or deactivate the system if necessary; and b) ensuring that an AI system would not have incentives to resist or deceive its operators. In a nutshell, they must be developed cooperatively for the safe deployment of GPAIS as responsible AI systems.

7. Conclusions

Many researchers are working on GPAIS, both to design new GPAIS and to define what they are. The principal goals of this work have been two-fold: (1) proposing a more comprehensive definition of GPAIS with a focus on their properties and functionalities, and (2) categorising different approaches to build them. In comparison with existing alternatives, our proposed definition allows for a more general view of GPAIS, considering different degrees of autonomy and expected capabilities. Together with these two principal goals, we have analysed shortly GenAI as the most important foundation models and the multimodality as a crucial aspect for managing multiple inputs. Finally, we have discussed the prospects posed by GPAIS, the implications to our society, and the regulation and governance of these emerging systems.

There are a multitude of approaches to making an AI system more general. In order to consolidate the most relevant ones in a cohesive manner, we have proposed a taxonomy of methods. This taxonomy conceptually distinguishes between AI models that rely on other AI models to achieve generalisation abilities and those that utilise a single AI model. More classical multi-task learning approaches and foundation models have been categorised as AI systems in which a single AI model exists, while alternative approaches may use two or more AI systems, what we called AI-powered AI to introduce generalisation capabilities.

The field of GPAIS is continually evolving, and the proposed taxonomy has established a robust foundation for understanding the diverse existing approaches. While LLMs are currently in the spotlight, there is a broad spectrum of approaches that can significantly contribute to the realisation of GPAIS. However, technical and ethical challenges must be necessarily discussed and addressed before stepping towards AGI. In the meantime, we must be mindful in the arrival of new open-world GPAIS models and task, on managing AI risks associated with current GPAIS, and work responsibly towards anticipating and mitigating such risks effectively via regulation and governance.

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.

Data availability

No data was used for the research described in the article.

Acknowledgements

I. Triguero is funded by a Maria Zambrano Senior Fellowship at the University of Granada. I. Triguero, F. Herrera, D. Molina, and J. Poyatos are supported by the R&D and Innovation project with reference PID2020-119478GB-I00 granted by Spain's Ministry of Science and Innovation and European Regional Development Fund (ERDF). J. Del Ser would like to thank the Basque Government, Spain for the funding support received through the EMAITEK and ELKARTEK programs (ref. KK-2023/00012), as well as the Consolidated Research Group MATHMODE (IT1456-22) granted by the Department of Education of this institution.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, Vol. 26, 2013, pp. 3111–3119.

[2] J. Grudin, R. Jacques, Chatbots, humbots, and the quest for artificial general intelligence, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–11, http://dx.doi.org/10.1145/3290605.3300439.

[3] A. Kaplan, M. Haenlein, Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence, Bus. Horiz. 62 (1) (2019) 15–25, http://dx.doi.org/10.1016/j.bushor.2018.08.004.

[4] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E.H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Trans. Mach. Learn. Res. (2022) 1–30.

[5] R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage?, 2023, arXiv:2304.15004.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M.T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, 2023, arXiv:2303.12712.

[7] J.E.H. Korteling, G. van de Boer-Visschedijk, R. Blankendaal, R. Boonekamp, A. Eikelboom, Human- versus artificial intelligence, Front. Artif. Intell. 4 (2021) 1–13, http://dx.doi.org/10.3389/frai.2021.622364.

[8] C.I. Gutierrez, A. Aguirre, R. Uuk, C.C. Boine, M. Franklin, A proposal for a definition of general purpose artificial intelligence systems, DISO 2 (36) (2023) http://dx.doi.org/10.1007/s44206-023-00068-w.

[9] H. Shevlin, K. Vold, M. Crosby, M. Halina, The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge, EMBO Rep. 20 (10) (2019) e49177, http://dx.doi.org/10.15252/embr.201949177.

[10] R. Fjelland, Why general artificial intelligence will not be realized, Humanit. Soc. Sci. Commun. 7 (1) (2020) http://dx.doi.org/10.1057/s41599-020-0494-4.

[11] R. Ashmore, R. Calinescu, C. Paterson, Assuring the machine learning lifecycle: Desiderata, methods, and challenges, ACM Comput. Surv. 54 (5) (2021) 1–39, http://dx.doi.org/10.1145/3453444.

[12] F. Hutter, L. Kotthoff, J. Vanschoren (Eds.), Automated Machine Learning - Methods, Systems, Challenges, in: The Springer Series on Challenges in Machine Learning, Springer, 2019, http://dx.doi.org/10.1007/978-3-030-05318-5.

[13] Y. Zhang, Q. Yang, An overview of multi-task learning, Natl. Sci. Rev. 5 (1) (2018) 30–43, http://dx.doi.org/10.1093/nsr/nwx105.

[14] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J.Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D.E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P.W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X.L. Li, X. Li, T. Ma, A. Malik, C.D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J.C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J.S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A.W. Thomas, F. Tramèr, R.E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S.M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models, 2022, arXiv:2108.07258.

[15] E. Real, C. Liang, D. So, Q. Le, AutoML-zero: Evolving machine learning algorithms from scratch, in: International Conference on Machine Learning, Vol. 119, 2020, pp. 8007–8019.

[16] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. 53 (3) (2020) 1–34, http://dx.doi.org/10.1145/3386252.

[17] G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Netw. 113 (2019) 54–71, http://dx.doi.org/10.1016/j.neunet.2019.01.012.

[18] C. Lemke, M. Budka, B. Gabrys, Metalearning: a survey of trends and technologies, Artif. Intell. Rev. 44 (1) (2015) 117–130, http://dx.doi.org/10.1007/s10462-013-9406-y.

[19] D. Silver, S. Singh, D. Precup, R.S. Sutton, Reward is enough, Artificial Intelligence 299 (2021) 103535, http://dx.doi.org/10.1016/j.artint.2021.103535.

[20] K.O. Stanley, J. Clune, J. Lehman, R. Miikkulainen, Designing neural networks through neuroevolution, Nat. Mach. Intell. 1 (1) (2019) 24–35, http://dx.doi.org/10.1038/s42256-018-0006-z.

[21] D. Hendrycks, M. Mazeika, T. Woodside, An overview of catastrophic AI risks, 2023, arXiv:2306.12001.

[22] A. Critch, S. Russell, TASRA: a taxonomy and analysis of societal-scale risks from AI, 2023, arXiv:2306.06924.

[23] S. Campos, R. Laurent, A definition of general-purpose AI systems: Mitigating risks from the most generally capable models, 2023, pp. 1–8, http://dx.doi.org/10.2139/ssrn.4423706, Available at SSRN 4423706.

[24] C. Stokel-Walker, R. Van Noorden, What chatGPT and generative AI mean for science, Nature 614 (7947) (2023) 214–216, http://dx.doi.org/10.1038/d41586-023-00340-6.

[25] European Commission, Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (ai act) and amending certain union legislative acts, 2021, URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, EUR-Lex.

[26] P. Hacker, A. Engel, M. Mauer, Regulating chatGPT and other large generative AI models, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1112–1123, http://dx.doi.org/10.1145/3593013.3594067.

[27] Future of Life Institute, General Purpose AI and the AI Act, Future of Life Institute, 2022, URL https://futureoflife.org/project/eu-ai-act/.

[28] S. Reed, K. Zolna, E. Parisotto, S.G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J.T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, N. de Freitas, A generalist agent, Trans. Mach. Learn. Res. (2022) 1–42.

[29] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al., Mastering atari, go, chess and shogi by planning with a learned model, Nature 588 (7839) (2020) 604–609, http://dx.doi.org/10.1038/s41586-020-03051-4.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[31] H. Alkaissi, S.I. McFarlane, Artificial hallucinations in chatGPT: Implications in scientific writing, Cureus 15 (2) (2023) e35179, http://dx.doi.org/10.7759/cureus.35179.

[32] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, Q. Huang, Toward realistic face photo-sketch synthesis via composition-aided GANs, IEEE Trans. Cybern. 51 (9) (2021) 4350–4362, http://dx.doi.org/10.1109/TCYB.2020.2972944.

[33] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of Hallucination in natural language generation, ACM Comput. Surv. 55 (12) (2023) 1–38, http://dx.doi.org/10.1145/3571730.

[34] S. Budd, E.C. Robinson, B. Kainz, A survey on active learning and human-in-the-loop deep learning for medical image analysis, Med. Image Anal. 71 (2021) 102062, http://dx.doi.org/10.1016/j.media.2021.102062.

[35] H. Song, I. Triguero, E. Özcan, A review on the self and dual interactions between machine learning and optimisation, Prog. Artif. Intell. 8 (2) (2019) 143–165, http://dx.doi.org/10.1007/s13748-019-00185-z.

[36] R.C. Barros, M.P. Basgalupp, A.A. Freitas, A.C.P.L.F. de Carvalho, Evolutionary design of decision-tree algorithms tailored to microarray gene expression data sets, IEEE Trans. Evol. Comput. 18 (6) (2014) 873–892, http://dx.doi.org/10.1109/TEVC.2013.2291813.

[37] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, Vol. 25, 2012, pp. 1–9.

[38] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2623–2631, http://dx.doi.org/10.1145/3292500.3330701.

[39] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D. Cox, Hyperopt: a Python library for model selection and hyperparameter optimization, Comput. Sci. Discov. 8 (1) (2015) 014008, http://dx.doi.org/10.1088/1749-4699/8/1/014008.

[40] C. Thornton, F. Hutter, H.H. Hoos, K. Leyton-Brown, Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 847–855, http://dx.doi.org/10.1145/2487575.2487629.

[41] E.K. Burke, M. Gendreau, M.R. Hyde, G. Kendall, G. Ochoa, E. Özcan, R. Qu, Hyper-heuristics: a survey of the state of the art, J. of the Oper. Res. Soc. 64 (12) (2013) 1695–1724, http://dx.doi.org/10.1057/jors.2013.71.

[42] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, F. Hutter, Auto-sklearn 2.0: Hands-free autoML via meta-learning, J. Mach. Learn. Res. 23 (261) (2022) 1–61.

[43] J. Parker-Holder, R. Rajan, X. Song, A. Biedenkapp, Y. Miao, T. Eimer, B. Zhang, V. Nguyen, R. Calandra, A. Faust, F. Hutter, M. Lindauer, Automated reinforcement learning (autorl): A survey and open problems, J. Artif. Int. Res. (ISSN: 1076-9757) 74 (2022) http://dx.doi.org/10.1613/jair.1.13596.

[44] X. He, K. Zhao, X. Chu, AutoML: A survey of the state-of-the-art, Knowl.-Based Syst. 212 (2021) 106622, http://dx.doi.org/10.1016/j.knosys.2020.106622.

[45] T. Elsken, J.H. Metzen, F. Hutter, Neural architecture search: A survey, J. Mach. Learn. Res. 20 (1) (2019) 1997–2017.

[46] S. Schrodi, D. Stoll, B. Ru, R. Sukthanker, T. Brox, F. Hutter, Construction of hierarchical neural architecture search spaces based on context-free grammars, 2023, arXiv:2211.01842.

[47] A.D. Martinez, J. Del Ser, E. Villar-Rodriguez, E. Osaba, J. Poyatos, S. Tabik, D. Molina, F. Herrera, Lights and shadows in evolutionary deep learning: Taxonomy, critical methodological analysis, cases of study, learned lessons, recommendations and challenges, Inf. Fusion 67 (2021) 161–194, http://dx.doi.org/10.1016/j.inffus.2020.10.014.

[48] Z.-H. Zhan, J.-Y. Li, J. Zhang, Evolutionary deep learning: A survey, Neurocomputing 483 (2022) 42–58, http://dx.doi.org/10.1016/j.neucom.2022.01.099.

[49] N. Hollmann, S. Müller, K. Eggensperger, F. Hutter, TabPFN: A transformer that solves small tabular classification problems in a second, 2023, arXiv:2207.01848.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 1–11.

[51] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang, Y. Yu, Taking human out of learning applications: A survey on automated machine learning, 2018, arXiv:1810.13306.

[52] W. Yi, R. Qu, L. Jiao, B. Niu, Automated design of metaheuristics using reinforcement learning within a novel general search framework, IEEE Trans. Evol. Comput. 27 (4) (2023) 1072–1084, http://dx.doi.org/10.1109/TEVC.2022.3197298.

[53] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A continual learning survey: Defying forgetting in classification tasks, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2022) 3366–3385, http://dx.doi.org/10.1109/TPAMI.2021.3057446.

[54] J. Gama, I. Žliobaitundefined, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 1–37, http://dx.doi.org/10.1145/2523813.

[55] K. Javed, M. White, Meta-learning representations for continual learning, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 1820–1830.

[56] J. Xu, Z. Zhu, Reinforced continual learning, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 907–916.

[57] K. Chatzilygeroudis, A. Cully, V. Vassiliades, J.-B. Mouret, Quality-diversity optimization: a novel branch of stochastic optimization, in: P.M. Pardalos, V. Rasskazova, M.N. Vrahatis (Eds.), Black Box Optimization, Machine Learning, and No-Free Lunch Theorems, 2021, pp. 109–135, http://dx.doi.org/10.1007/978-3-030-66515-9_4.

[58] A. Cully, Y. Demiris, Quality and diversity optimization: A unifying modular framework, IEEE Trans. Evol. Comput. 22 (2) (2018) 245–259, http://dx.doi.org/10.1109/TEVC.2017.2704781.

[59] C.V. Nguyen, Y. Li, T.D. Bui, R.E. Turner, Variational continual learning, in: International Conference on Learning Representations, 2018, pp. 1–18.

[60] R. Wang, J. Lehman, J. Clune, K.O. Stanley, Paired open-ended trailblazer (POET): Endlessly generating increasingly complex and diverse learning environments and their solutions, 2019, arXiv:1901.01753.

[61] J. Clune, AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence, 2020, arXiv:1905.10985.

[62] K. Li, J. Malik, Learning to optimize, in: 5th International Conference on Learning Representations, ICLR 2017, 2017, pp. 1–13.

[63] Y.-X. Wang, D. Ramanan, M. Hebert, Learning to model the tail, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 7032–7042.

[64] Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, Zero-shot learning- A comprehensive evaluation of th good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2019) 2251–2265, http://dx.doi.org/10.1109/TPAMI.2018.2857768.

[65] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop, Vol. 2, 2015, pp. 1–30.

[66] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, in: Advances in Neural Information Processing Systems, Vol. 29, 2016, pp. 3637–3645.

[67] B. Settles, Active Learning, in: Synthesis Lectures on Artificial Intelligence and Machine Learning, Springer, 2012, pp. 1–100, http://dx.doi.org/10.1007/978-3-031-01560-1.

[68] M. Fang, J. Yin, L.O. Hall, D. Tao, Active multitask learning with trace norm regularization based on excess risk, IEEE Trans. Cybern. 47 (11) (2017) 3906–3915, http://dx.doi.org/10.1109/TCYB.2016.2590023.

[69] A. Khan, Knowledge graphs querying, 2023, arXiv:2305.14485.

[70] L. Panait, S. Luke, Cooperative multi-agent learning: The state of the art, Auton. Agents Multi Agent Syst. 11 (3) (2005) 387–434, http://dx.doi.org/10.1007/s10458-005-2631-2.

[71] P.R. Silva, J. Vinagre, J. Gama, Towards federated learning: An overview of methods and applications, WIREs Data Min. Knowl. Discov. 13 (2) (2023) e1486, http://dx.doi.org/10.1002/widm.1486.

[72] M. Dorigo, G. Theraulaz, V. Trianni, Swarm robotics: Past, present, and future [point of view], Proc. IEEE 109 (7) (2021) 1152–1165, http://dx.doi.org/10.1109/JPROC.2021.3072740.

[73] H.L. Kwa, J.L. Kit, N. Horsevad, J. Philippot, M. Savari, R. Bouffanais, Adaptivity: a path towards general swarm intelligence? Front. Robot. AI 10 (2023) 1163185, http://dx.doi.org/10.3389/frobt.2023.1163185.

[74] A. Zweig, G. Chechik, Group online adaptive learning, Mach. Learn. 106 (9–10) (2017) 1747–1770, http://dx.doi.org/10.1007/s10994-017-5661-5.

[75] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, Domain generalization for object recognition with multi-task autoencoders, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2551–2559, http://dx.doi.org/10.1109/ICCV.2015.293.

[76] E.A. Sosnina, S. Sosnin, M.V. Fedorov, Improvement of multi-task learning by data enrichment: application for drug discovery, J. Comput.-Aided Mol. Des. 37 (4) (2023) 183–200, http://dx.doi.org/10.1007/s10822-023-00500-w.

[77] X. Chen, C. Liu, Y. Zhao, Z. Jia, G. Jin, Improving adversarial robustness of Bayesian neural networks via multi-task adversarial training, Inform. Sci. 592 (2022) 156–173, http://dx.doi.org/10.1016/j.ins.2022.01.051.

[78] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (12) (2022) 5586–5609, http://dx.doi.org/10.1109/TKDE.2021.3070203.

[79] A.D. Martinez, J. Del Ser, E. Osaba, F. Herrera, Adaptive multifactorial evolutionary optimization for multitask reinforcement learning, IEEE Trans. Evol. Comput. 26 (2) (2022) 233–247, http://dx.doi.org/10.1109/TEVC.2021.3083362.

[80] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, X. Yue, Meta-transformer: A unified framework for multimodal learning, 2023, arXiv:2307.10802.

[81] Y. Chebotar, K. Hausman, Y. Lu, T. Xiao, D. Kalashnikov, J. Varley, A. Irpan, B. Eysenbach, R.C. Julian, C. Finn, S. Levine, Actionable models: Unsupervised offline reinforcement learning of robotic skills, in: Proceedings of the 38th International Conference on Machine Learning, Vol. 139, 2021, pp. 1518–1528.

[82] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A.G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, M. Goldblum, A cookbook of self-supervised learning, 2023, arXiv:2304.12210.

[83] A. Oussidi, A. Elhassouny, Deep generative models: Survey, in: 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 2018, pp. 1–8, http://dx.doi.org/10.1109/ISACV.2018.8354080.

[84] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, Vol. 27, 2014, pp. 2672–2680.

[85] E. McAlpine, P. Michelow, E. Liebenberg, T. Celik, Is it real or not? Toward artificial intelligence-based realistic synthetic cytology image generation to augment teaching and quality assurance in pathology, J. Am. Soc. Cytopathol. 11 (3) (2022) 123–132, http://dx.doi.org/10.1016/j.jasc.2022.02.001.

[86] J. Toutouh, S. Nalluru, E. Hemberg, U.-M. O'Reilly, Semi-supervised generative adversarial networks with spatial coevolution for enhanced image generation and classification, Appl. Soft Comput. 148 (2023) 110890, http://dx.doi.org/10.1016/j.asoc.2023.110890.

[87] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674–10685, http://dx.doi.org/10.1109/CVPR52688.2022.01042.

[88] A. Nichol, A. Ramesh, P. Mishkin, P. Dariwal, J. Jang, M. Chen, DALL·E 2 Pre-Training Mitigations, OpenAI, 2022, URL https://openai.com/research/dall-e-2-pre-training-mitigations.

[89] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, A. Ramesh, Improving Image Generation with Better Captions, OpenAI, 2023, URL https://cdn.openai.com/papers/dall-e-3.pdf.

[90] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving latent diffusion models for high-resolution image synthesis, 2023, arXiv:2307.01952.

[91] E. van Dis, J. Bollen, W. Zuidema, R. van Rooij, C. Bockting, ChatGPT: five priorities for research, Nature 614 (7947) (2023) 224–226, http://dx.doi.org/10.1038/d41586-023-00288-7.

[92] OpenAI, GPT-4 technical report, 2023, arXiv:2303.08774.

[93] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, arXiv:2302.13971.

[94] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J.E. Gonzalez, I. Stoica, E.P. Xing, Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, LMSYS, 2023, URL https://lmsys.org/blog/2023-03-30-vicuna/.

[95] H. Touvron, L. Martin, K. Stone, Llama 2: Open Foundation and Fine-Tuned Chat Models, Tech. Rep., Meta AI, 2023.

[96] A. Patrizio, Google Bard. TechTarget, TechTarget, 2023, URL https://www.techtarget.com/searchenterpriseai/definition/Google-Bard.

[97] A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, D. Song, The false promise of imitating proprietary LLMs, 2023, arXiv:2305.15717.

[98] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only, 2023, arXiv:2306.01116.

[99] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM, Databricks, 2023, URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[100] A. Moradi Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M.C. Desmarais, Z.M. Jiang, Github copilot AI pair programmer: Asset or liability? J. Syst. Softw. 203 (2023) 111734, http://dx.doi.org/10.1016/j.jss.2023.111734.

[101] T.J. Sejnowski, Large language models and the reverse turing test, Neural Comput. 35 (3) (2023) 309–342, http://dx.doi.org/10.1162/neco_a_01563.

[102] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation, Inf. Fusion 99 (2023) 101896, http://dx.doi.org/10.1016/j.inffus.2023.101896.

[103] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2242–2251, http://dx.doi.org/10.1109/CVPR.2017.241.

[104] K. Kazuhiro, R. Werner, F. Toriumi, M. Javadi, M. Pomper, L. Solnes, F. Verde, T. Higuchi, S. Rowe, Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images, Tomography 4 (4) (2018) 159–163, http://dx.doi.org/10.18383/j.tom.2018.00042.

[105] S. Guan, M. Loew, Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks, J. Med. Imaging 6 (3) (2019) http://dx.doi.org/10.1117/1.JMI.6.3.031411.

[106] A. Bissoto, E. Valle, S. Avila, GAN-based data augmentation and anonymization for skin-lesion analysis: A critical review, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2021, pp. 1847–1856, http://dx.doi.org/10.1109/CVPRW53098.2021.00204.

[107] V. Thambawita, J.L. Isaksen, S.A. Hicks, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, C. Ellervik, M.S. Olesen, T. Hansen, C. Graff, N.-H. Holstein-Rathlou, I. Strümke, H.L. Hammer, M.M. Maleckar, P. Halvorsen, M.A. Riegler, J.K. Kanters, DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine, Sci. Rep. 11 (1) (2021) 21896, http://dx.doi.org/10.1038/s41598-021-01295-2.

[108] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E.H. Chi, Q.V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, 2023, pp. 1–14.

[109] H. Yang, S. Yue, Y. He, Auto-GPT for online decision making: Benchmarks and additional opinions, 2023, arXiv:2306.02224.

[110] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face, 2023, arXiv:2303.17580.

[111] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S.K.S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, MetaGPT: Meta programming for multi-agent collaborative framework, 2023, arXiv:2308.00352.

[112] A.M. Bran, S. Cox, A.D. White, P. Schwaller, Chemcrow: Augmenting large-language models with chemistry tools, 2023, arXiv:2304.05376.

[113] D. Hall, J. Llinas, An introduction to multisensor data fusion, Proc. IEEE 85 (1) (1997) 6–23, http://dx.doi.org/10.1109/5.554205.

[114] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion (ISSN: 1566-2535) 57 (2020) 115–129, http://dx.doi.org/10.1016/j.inffus.2019.12.001.

[115] H. Hermessi, O. Mourali, E. Zagrouba, Multimodal medical image fusion review: Theoretical background and recent advances, Signal Process. 183 (2021) 108036, http://dx.doi.org/10.1016/j.sigpro.2021.108036.

[116] A.P. James, B. Dasarathy, A review of feature and data fusion with medical images, 2015, arXiv:1506.00097.

[117] W. Zhu, X. Wang, H. Li, Multi-modal deep analysis for multimedia, IEEE Trans. Circuits Syst. Video Technol. 30 (10) (2020) 3740–3764, http://dx.doi.org/10.1109/TCSVT.2019.2940647.

[118] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[119] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multimodal data fusion, Neural Comput. 32 (5) (2020) 829–864, http://dx.doi.org/10.1162/neco_a_01273.

[120] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2021) 4–24, http://dx.doi.org/10.1109/TNNLS.2020.2978386.

[121] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, http://dx.doi.org/10.1016/j.aiopen.2021.01.001.

[122] Z. Zhang, P. Cui, W. Zhu, Deep learning on graphs: A survey, IEEE Trans. Knowl. Data Eng. 34 (1) (2022) 249–270, http://dx.doi.org/10.1109/TKDE.2020.2981333.

[123] S. Felix, S. Kumar, A. Veeramuthu, A smart personal AI assistant for visually impaired people, in: Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018, 2018, pp. 1245–1250, http://dx.doi.org/10.1109/ICOEI.2018.8553750.

[124] A. König, L. Alčiauskaitė, T. Hatzakis, The impact of subjective technology adaptivity on the willingness of persons with disabilities to use emerging assistive technologies: A European perspective, in: Computers Helping People with Special Needs, 2022, pp. 207–214, http://dx.doi.org/10.1007/978-3-031-08648-9_24.

[125] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, K. Chen, MultiModal-GPT: A vision and language model for dialogue with humans, 2023, arXiv:2305.04790.

[126] H. Chen, J. Wang, M.Q.-H. Meng, Kinova gemini: Interactive robot grasping with visual reasoning and conversational AI, in: 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2022, pp. 129–134, http://dx.doi.org/10.1109/ROBIO55434.2022.10011896.

[127] R. Schwartz, J. Dodge, N.A. Smith, O. Etzioni, Green AI, Commun. ACM 63 (12) (2020) 54–63, http://dx.doi.org/10.1145/3381831.

[128] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650, http://dx.doi.org/10.18653/v1/P19-1355.

[129] R. Verdecchia, J. Sallou, L. Cruz, A systematic review of green AI, WIREs Data Min. Knowl. Discov. 13 (4) (2023) e1507, http://dx.doi.org/10.1002/widm.1507.

[130] R. Burnell, W. Schellaert, J. Burden, T.D. Ullman, F. Martinez-Plumed, J.B. Tenenbaum, D. Rutar, L.G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E.M. Voorhees, A.G. Cohn, J.Z. Leibo, J. Hernandez-Orallo, Rethink reporting of evaluation results in AI, Science 380 (6641) (2023) 136–138, http://dx.doi.org/10.1126/science.adf6369.

[131] J. Parmar, S. Chouhan, V. Raychoudhury, S. Rathore, Open-world machine learning: Applications, challenges, and opportunities, ACM Comput. Surv. 55 (10) (2023) 1–37, http://dx.doi.org/10.1145/3561381.

[132] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[133] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, http://dx.doi.org/10.1016/j.inffus.2023.101805.

[134] R. Bommasani, K. Klyman, D. Zhang, P. Liang, Do Foundation Model Providers Comply with the EU AI Act?, Center for Research on Foundation Models, 2023, URL https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

[135] I. Hupont, M. Micheli, B. Delipetrev, E. Gomez, J. Garrido, Documenting high-risk AI: A European regulatory perspective, Computer 56 (05) (2023) 18–27, http://dx.doi.org/10.1109/MC.2023.3235712.

[136] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, Y.N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A.G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahnemann, J. Brauner, S. Mindermann, Managing AI risks in an era of rapid progress, 2023, arXiv:2310.17688.

[137] N. Palladino, A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices, Telecommun. Policy 47 (5) (2023) 102479, http://dx.doi.org/10.1016/j.telpol.2022.102479.

[138] V. Almeida, L.S. Mendes, D. Doneda, On the development of AI governance frameworks, IEEE Internet Comput. 27 (1) (2023) 70–74, http://dx.doi.org/10.1109/MIC.2022.3186030.