On the Trajectory of AI

Christopher Hong

December 9, 2024

1 Abstract

The goal of this public manuscript is four fold. First, educate about the historical development of **artificial intelligence** (**AI**). Second, give an objective analysis of the actual technical capabilities and limitations of the current state-of-the-art in AI research. The analysis may use academic jargon for clarity to the informed reader, but still overall be understandable to they layman. Third, carefully define the term AI for the sake of objective clarity. Fourth, create a biased but educated hypothesis about important challenges in the field and possible future developments.

An important note: whenever the term **contemporary** is used, it refers to the time period reflecting the overall state of academia or industry around the year 2024. The term is liberally used and is meant to be taken without preconceived judgement in any statement involving the term unless noted otherwise.

2 The History of AI

2.1 Birth

AI as a field is ill-defined and is multi-disciplinary by nature, and consequently its history is highly non-linear. Thus, many details will be skipped in favor of briefness, and only the most notable events will be covered.

Artificial intelligence as a serious academic consideration can be traced back to the invention of the computer, in which leading pioneers of computation, most notably Alan Turing, speculated the possibility of a thinking machine that could replicate the capabilities of a human mind. The actual birth of AI

as an academic field can be traced back to the invention of the Perceptron by McCulloch and Pitts, a computational model of an artificial neuron inspired by neuroscience. They showed that a network of these neurons could process any computable function [12]. The Perceptron model remains the fundamental building block of contemporary neural nets.

Marvin Minsky, soon to be a leading AI researcher, took inspiration from the Perceptron as a promising starting point to pursue AI research. He, along with John McCarthy, Allen Newell, Herbert Simon, and other well recognized researchers, organized a two month workshop on the topic of **artificial intelligence**, the term coined by John McCarthy. Afterwards the named researchers established leading AI research centers in MIT, CMU, and Stanford. IBM also was a prominent research center [21]. Their research goal was undoubtedly grand; it was to understand and replicate the capabilities of the human mind.

2.2 GOFAI (1952-1969)

Thus emerged the first wave of enthusiasm for AI in the academic community. Two general camps of AI research were founded: connectionism, which embraced neural nets, and symbolism, which embraced logic. Minsky published the book "Perceptrons" (1969) [14] analyzing the limitations of single layer neural nets, which is widely believed to have dissuaded many researchers from further pursuing connectionist research. Most researchers pursued symbolism, in particular structured logics, as the path to artificial intelligence. Some exemplary architectures include Newell and Simon's General Problem Solver [15] and Minsky's frames [13]. Research from this time period is now referred to as Good Old-Fashioned AI (GOFAI), mostly to contrast with the direction of contemporary research which is mostly focused on neural nets. In fact the term GOFAI could be perceived as luddite-ish and pejorative in AI research. The systems were first focused on being generic symbolic machines, but the scope of problems they could tackle was found to be in only very limited, highly idealized domains (simple games, block worlds, etc.) [21].

2.3 Expert systems (1969-1979)

The natural path forward to expand the scope of problems AI systems could tackle was to embed these generic symbolic systems with a wealth of domain knowledge encoded by knowledge experts. Thus the systems developed during this era were called **expert systems**. Expert systems also became popular in industry, with investors pouring in billions of dollars. One notable system

from this era is Cyc (1984) [11], which was particularly invested in the idea that scaling expert knowledge in a symbolic system would yield increased intelligence. The enthusiasm spread to industry and private financial capital, but about a decade after these expert systems also proved to be brittle, and soon the enthusiasm for not just expert systems but AI in general waned in both industry and academia. The following period is now known as the AI winter, in which the pursuit of AI was generally shunned.

2.4 Resurgence of neural networks, emergence of probability theory in AI, modern robotics (1986-present)

Following the collapse of symbolic research in AI, Rumelhart and McClelland published "Parallel Distributed Procesing" in 1986 [19], which ushered in a resurgence of interest in neural networks and more generally connectionist approaches to AI. Around the same time, the most important work concerning connectionist research was published in Nature by Rumelhart, Hinton, and Williams (1986) [20] detailing a technique called back-propagation. Shortly after in 1988 Judea Pearl published "Probabilistic Reasoning in Intelligent Systems" [17], which brought in probability theory into the mainstream of AI research. This was an important development in AI research, as purely symbolic AI was unable to handle uncertainty and ambiguity present in situtations that intelligent systems should handle. Around this time, robotics research flourished as well, although fragmented in domains such as control theory, path planning, computer vision, and reinforcement learning (RL). In 2005, Sebastian Thrun published "Probabilistic Robotics" [24], incorporating methods from control theory and probability into robotics research. Using the methods laid out in "Probabilistic Robotics", Thrun and a team of Stanford researchers won an autonomous vehicle race called the DARPA Grand challenge in 2006, beating competitors by a large margin [25]. Probabilistic methods and ideas still remain highly influential in contemporary applications of autonomous driving, such as conducted by Waymo and Tesla, as well as contemporary AI in general.

2.5 Machine learning and Big Data (1995-present)

After the invention of the computer and in particular after the inventions of semiconductors and VLSI, computation and data storage gradually became more capable and cheaper on an exponential scale. This progress naturally led researchers to explore the possibility of harnessing large datasets to solve problems. Yarowsky (1995) [30] showed a way to disambiguate the use of the

world "plant" between fauna and factory using a large corpus of unlabeled text. Banko and Brill (2001) [2] showed that such unsupervised learning techniques improved as the corpus enlarged from millions to billions of words [21]. Such results, which continue to be reinforced by contemporary research, led to the repudation of expert systems of the past, as large corpuses of data prove to be a more reliable source of knowledge than expertly encoded knowledge or features when combined with probabilistic methods that incorporate uncertainty. This trend of collecting large corpuses of data to facilitate model learning is known as Big Data. Around this time the academic field known as **machine learning** was already taking off. Machine learning can be seen as the application of probability theory and data to generate predictions of new data, and is used as the dominant theoretical basis of contemporary neural net work.

2.6 Big Computation (2010-present)

Around 2009 a group of researchers at Stanford published ImageNet [7], an unprecedentedly massive dataset of 14 million labeled images, continuing the trend of Big Data.

In 2010, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2010) was won by Krizhevsky, Sutskever, and Hinton using a convolutional neural net with a large number of parameters [10]. The goal of the challenge was to classify correctly as many images as possible to the correct category. They achieved an error rate of 15.3%, far exceeding the runner-up of the competition with an error rate of 26.2%, attracting the interest of the computer vision and wider machine learning communities where neural nets are still the dominant model in contemporary research.

The computation required was so large at the time such that a different form of computing was necessary for it to be practical. The researchers chose to implement the convolutional neural net using Nvidia GPUs with the proprietary CUDA programming language suited for general-purpose GPU (GPGPU) computing. GPU computing is highly adept at running programs in which a large majority of its computation is parallelizeable (see Amdahl's Law [1]), which was perfect for neural nets. Thus we can mark 2010 as the beginning of Big Computation accompanying Big Data in machine learning necessary to run stochastic gradient descent (SGD, the modern, refined version of back-propagation) on large datasets. The action of applying SGD to a neural net model on a dataset is known as **training** the model, or if the subject and object are flipped, when training the model using SGD, the model **learns** from the data. Training is the algorithm that allows a model to recog-

nize features in the data that can then allow it to perform the desired objective (classifying, segmenting, recommending, optimal decision making, etc.).

The general construction of neural nets from this point is arranged in a great number of layers, leading them to be relatively "tall" versus "wide". Thus neural net researchers now generally attribute their work as **deep learning**, so from this point I will refer to neural net work as deep learning.

2.7 Bigger Models, Bigger Data with Self-Supervised Learning (2017-present)

In 2017, Vaswani et al. at Google published a paper introducing the Transformer neural net architecture [28], which in 2024 remains the dominant model for AI in academia and industry across all practical domains. The Transformer model was notable for being particularly well suited to parallel computation, enabling researchers to run neural net models for sequential datasets with even larger parameter counts; at the time, the dominant neural net model for sequential data was Recurrent Neural Nets (RNNs), which suffered from lower training efficiency relative to computation and model size as well as back-propagation issues.

In 2019, Devlin et al. at Google released BERT (Bidirectional Encoder Representations for Transformers) [8], a training paradigm on natural language corpuses that required no human labelling. The general training framework is referred to as "pretraining" or sometimes "self-supervised learning" (although the exact concepts referred to these two terms differ, they are for this historical discussion practically similar). BERT led to large performance gains in natural-language processing (NLP) benchmarks, and its success led the deep learning community to evaluate whether pretraining could be applied to modalities apart from natural language. The size as well as the amount of computation and data expended to produce these NLP models led researchers to label them as **Large Language Models** (**LLM**s).

In 2021 OpenAI released CLIP (Contrastive Language-Image Pretraining) [18]. CLIP represents the adoption of pretraining in a multimodal setting, scraping images and their captions from the Internet and using their pairings to train a large language-image model. This departs from the tradition popularized by ImageNet of manually labelling images for the purpose of training neural net models.

In 2022 OpenAI released ChatGPT, a LLM similar in architecture and training

paradigm to BERT. ChatGPT was notable for marketing and releasing an LLM to the general public and quickly capturing its attention.

The emergence of very large neural net models such as BERT, ChatGPT, and CLIP led mainstream AI researchers to refer to them as foundation models [3]. Researchers found that scaling both neural net model size and amount of data led to gains in performance on standard NLP and computer vision benchmarks, echoing the trend observed in 2001. The emergence of foundation models have academics and the public again pining for the original promise of AI, now branded as **AGI** (Artificial General Intelligence).

3 Analysis of contemporary AI

3.1 Broad analysis

Having recounted a brief history of AI as a field of research with some industry impacts, it is worth taking a step back to analyze the current state and progress in AI.

The analysis of the positive mainstream capabilities will be quick in the interest of briefness and the expectation that the reader will be at least somewhat familiar with its capabilities after large deep learning models were released by private industry to the public. The briefness will likely at least somewhat undersell the positive impact of current deep learning models.

Deep learning models are extraordinarily adept at extracting low level features from high dimensional data, which can be applied as useful information in any machine learning task. This was observed with stark clarity in AlexNet, which beat out human-crafted computer vision algorithms used by competitors. Subsequent analysis of the lower levels of convolutional neural nets trained on images show a surprising activation similarity with low level visual neurons in the brain [26]. Deep learning models are also great at learning higher level features that provide information (in the information Theoretic sense) for various downstream tasks such as image classification and segmentation. This emphasis on learning from data is a fundamental shift in the historical development of AI, since many systems in the past (both symbolic and probabilistic) were designed often with the development of a specialized algorithm by a domain expert. The learning systems proved to be more performant and robust given enough appropriate data despite the system itself not being developed with any domain expertise [22]. The domain expert still has a role in appropriately defining a working definition of a problem and identifying the appropriate

data, but no longer has a significant role in creating the algorithm itself. This allows a faster iteration of AI systems to adapt to a particular problem, since the developers of the AI system no longer have to design and maintain the underlying algorithm, although at the cost of computation and data collection / organization.

As can be gleaned from the recounting of AI history, the dominant paradigm of AI in research and industry as of 2024 is deep learning, in particular foundation models. Thus most of the following analysis will focus on deep learning and the theoretical foundations it currently relies on. The symbolic AI that was dominant at the birth of AI as a research field has been found to be successful in only very narrow domains, and was brittle when applied to real world problems. This is mostly due to symbolic AI's theoretical inability to cope with the uncertainty, ambiguity, and high-dimensional nature of real world problems. Thus, all mainstream researchers, if even aware or knowledgeable of symbolic research, now reject a purely symbolic approach to AI.

Many AI researchers note the historical death of the pursuit of AGI in the AI winter. To recall, the seminal goal of AI was to replicate the capabilities of the human mind. In particular, these founding researchers were fascinated by the multi-domain aspect and adaptivity of humans in solving practical problems. Researchers (including myself) are still fascinated with these human attributes, and the term AGI emphasizes the general nature of human intelligence, contrasting with the capabilities of past and current systems developed under the AI umbrella.

It is important to hypothesize why the general nature of current machine learning and reinforcement learning techniques (used in deep learning, the dominant AI paradigm) may not necessarily lead to AGI. First it is worth noting that creating general machine / reinforcement learning techniques is in itself an achievement. These techniques allow a practitioner to apply any appropriate off-the-shelf algorithm to a particular problem, given that they have adequate data to support the algorithm. This data may even not be manually labeled as long as it is gathered in sufficient quantity, as exemplified by BERT, ChatGPT, CLIP, etc. However, what is noted by both mainstream and AGI researchers is that given any algorithm applied to a particular dataset for a particular problem leads the resulting model to be highly specialized to that problem and dataset. Thus mainstream AI and AGI researchers refer to this contemporary phenomemon as **narrow AI**, which can be perceived as itself a limitation despite the general nature of the algorithms used.

One could reasonably argue that in fact foundation models like ChatGPT and

CLIP have solved this problem and are leading to AGI. The idea of foundation models is to create one large model that can be applied to many downstream tasks, so in that sense these models have general adaptability. I would agree that the specialization problem of contemporary machine learning has been side-stepped to a degree. However, given some empirical observations and theoretical considerations, the promise of foundation models of being highly adaptable may be misleading. For example, Zhang and Bottou (2023) [31] observe that the training dynamics of neural net models leads to the loss of rich representations, particularly when adding regularization. This loss is not consequential in the stationary distribution setting, but in a shifting distribution setting the loss of some representations leads to a degredation in performance. To compensate they demonstrate that an ensemble of nearly identical models is more robust under shifting distribution, concluding that foundation models may not be as adaptable as expected.

In addition, one could counter that the pretraining methods of LLMs may seem to lead to general intelligence but are in fact very specific; they have a simple, unitary goal of statistically replicating text corpuses.

3.2 Known flaws of deep learning

Deep learning is highly successful on a plethora of standard machine learning benchmarks (CIFAR, ImageNet, GLUE, etc.). In fact, new benchmarks are consistently created by the academic community because old benchmarks quickly have been made inadequate and obsolete. However, deep learning has also been noted to have unhuman-like behaviors since its popularization. One notable example is the possibility of generating adversarial examples, where deep learning models make mistakes where humans would not. This was demonstrated by Szegedy et al. (2014) [23] by adding deliberate but tiny noise to images that caused deep learning models to misclassify images. This suggests that deep learning models do not learn disentangled or semantically meaningful features, at least the same high level features that humans recognize. The repeated success of deep learning models and probing individual perceptrons suggest otherwise, but the fact remains that the representations learned by deep learning models are somehow different and/or perhaps not as rich compared to the representations recognized by humans in some fundamental sense.

Deep learning models may also not be robust to shifting distributions, the detailed meaning of which will be explained in the next section. For now the reader can interpret distribution as "character of the dataset". One em-

pirical observation supporting this claim is the "catastrophic forgetting" phenomenon observed since the inception of connectionist research. Generally this means that if the character of the data changes during training of a neural net, then the neural net will fail to recall aspects of the former character of the data. Research related to a specialized field of probability called causality also demonstrate the brittleness of mainstream deep learning models under shifting distributions.

Deep learning also has a problem of **sample inefficiency**. This means that deep learning generally requires large amounts of data to be trained effectively. More importantly, deep learning requires a certain level of diversity in the dataset in order for the resulting model to be effective. Another view of this problem is that if a deep learning model has not been trained on a data point sufficiently similar to the data point received when deployed, the deep learning model will likely fail to recognize and characterize that data point. This is similar to the problem of brittleness under shifting distribution.

So the current flaws of deep learning relate to its brittleness under real world circumstances, the very problem that caused the demise of symbolic AI methods. Deep learning methods are certainly more robust, but the brittleness problem has not been completely solved, nor is it near being completely solved.

I have used the term "robust" and the opposite term "brittle" liberally without explaining them too much under the vague concept of "shifting distributions", but hopefully the following sections will elucidate their meanings. At this point the reader can note that they are related to the adaptability people expect of an intelligent system.

3.3 Deep analysis of deep learning

In order to understand the flaws of current deep learning models, it is important to have at least a cursory understanding of the theoretical foundations of these models. Deep learning models trained with SGD rely on the assumptions of machine learning, which itself is derived from probability theory. Machine learning, at least in its current form, generally assumes a **statistically stationary** data model. What this means is that the likelihood (in the probabilistic sense) of any datapoint to appear in a dataset is always constant. However, people know intuitively that this is not true; the likelihood (in the intuitive sense) of an event is conditioned on past events, and history rhymes but does not repeat. Probability theory is in theory adequate as the foundation of machine learning under uncertainty, but real world problems often cannot

practically fulfill the requirements of probability theory. Using a stationary probability world model requires that all relevant variables are incorporated in the world model. In real world problems however, an agent never has close to all of the relevant variables required to predict future datapoints under probability theory.

Machine learning and reinforcement learning practitioners may argue that current techniques already account for this problem; so-called **hidden models** (ex. Hidden Markov Models) assume that an agent cannot sense all of the relevant variables required to predict future events. I would counter that the practice of using hidden models involves creating a hypothesis of the hidden states, in other words predict what it cannot see. However, the hidden model that the agent discovers must itself be stationary while limited in capacity with respect to the number of variables it keeps track of. It is not self evident that a practical model can exist for all real world problems, particularly open ended ones. On the physical extreme, the hidden model could represent every relevant atom in the world, and the transition model to be discovered can be well known physics. While the transition model from one state to the next may be relatively simple and obviously stationary, the number of atoms necessary to account for a model with reasonable fidelity would be so large that the whole model would become computationally intractable, not least from combinatorial explosion.

4 On defining Artificial Intelligence

With the preceding criticism of contemporary AI research and wisdom, it is worth discussing the definition of AI. What should the term mean (what is its normative usage)? Often the term is thrown around without much care to its meaning, which is normal for words but uniquely dangerous with AI.

As far as industry (in 2024) is concerned, AI generally refers to methodology when meaningfully used, usually foundation models, machine learning, or contemporary autonomous vehicle technology. This is a shame because AI in its founding as a research field and etymology refers to a goal and not a particular method, so goal and methodology are often conflated. This conflation propagates the hegemonic belief that contemporary AI techniques will result in the goal of AI or AGI, a popular belief but not self-evident. The term AI when used as a methodology is then semantically poor when a more specific term could be used. Often "AI" in industry is abused to refer to methods or technologies wholly unrelated to one of the three fields initially mentioned to

garner positive attribution with an hype filled term and attract private capital. Vice-versa foundation models and machine learning are often used inappropriately without knowledge or regard to their limitations and referred to as "AI" to attract hype and financial capital (and possibly from general excitement irrespective of financial motive).

Shifting our attention towards public discourse, AI in 2024 is often synonymous with the surge of interest in foundation models such as ChatGPT. This is due to the large amount of private capital invested in the marketing and dissemination of foundation models to the public. Again, we see a conflation of technique and goal in contemporary public discourse of AI, and so the limitations of foundation models become apparent when the wider public expects these foundation models to behave "intelligently", usually meaning "like a human of average or adequate intelligence for the task at hand".

Now I reflect on the historical and normative usage of the term AI in academic discourse, which contributes to the confusion and lack of direction in defining the term at the public level. Most people do not have trouble interpreting the "Artificial" in AI, which clearly means human-made. The confusion arises from the term "Intelligence". What is intelligence? At the beginning of the manuscript, the implicit assumption made was that AI is the goal of replicating human capabilities in machines, which begs the question: Which human capabilities represent intelligence? After all, computers calculate addition infinitely faster than humans, but we do not regard computers as intelligent, or at least more intelligent than humans [29].

Some perspectives from founding members of AI research:

- By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity" (Newell and Simon, 1976)
- Intelligence usually means "the ability to solve hard problems" (Minsky 1985a)
- AI is concerned with methods of achieving goals in situations in which the information available has a certain complex character. The methods that have to be used are related to the problem presented by the situation and are similar whether the problem solver is human, a Martian, or a computer program (McCarthy, 1988)

The latter two perspectives by Minsky and McCarthy suggest that AI is related to solving hard problems. With the benefit of hindsight, researchers now recognize that one of the biggest missteps (or discovery) in early AI research was the unquestioned assumption that the researchers could define what a hard problem was, or even that a problem's difficulty can be evaluated on an objective linear scale. For example, in 1966 Minsky famously gave a graduate student the task of solving computer vision over a single summer ("connect a camera to a computer and get the machine to describe what it sees" [6]), and later found the task to be much more difficult than expected. One can speculate that this optimism was founded on the observation that most humans find visually recognizing objects and animals to be a trivial task. It was not until much later with the application of convolutional neural nets with sufficient computing power that visual object recognition started to become feasible in computers. Often earlier AI researchers tackled what they considered a hard problem, such as chess playing, only to realize that their solutions far outpaced human ability. However, other tasks that humans found trivial such as driving a car, doing household chores, etc. could not be easily replicated artificially. Thus the common mantra was born that what's easy for humans is difficult for computers and vice versa. While the mantra is a given for contemporary AI researchers, Rodney Brooks is notable in identifying the trend in a popular work "Elephants don't play chess" [4]. The achievements and struggles in historical AI development suggest that humans are highly specialized in certain abilities, particularly relating to the physical world, and their strength is not in performing any one particular task but their ability to learn and adapt to multiple tasks. Speculatively, evolutionary pressures drove these strengths in humans who otherwise would uselessly expend great amounts of valuable energy on cognition.

So defining intelligence as the ability to solve hard problems is useless and worse highly misleading in the context of AI research, although this is the common interpretation of public usage of the term. The common interpretation is useful in the context of human discourse because "intelligence" is usually referred to when comparing humans to other humans who develop similar cognitive architectures and have similar sensorimotor experiences. However, the common interpretation is useless in artificial intelligence because objective, linear comparisons of difficulty cannot respect the vast possibility of differences in cognitive architectures and sensorimotor experiences between humans and artificial machines [29]. Similar difficulties can be self observed when attempting to compare intelligences between different animals.

Newell and Simon's definition of intelligence seems more realistic when seri-

ously scrutinized. They suggest that, given a fixed amount of computation and time, an intelligent system should be able to adapt to the demands of the environment. The keyword here is *adapt*. The definition reinforces the problems observed in contemporary AI approaches, which struggle to adapt under changing circumstances. Pei Wang develops this idea much further in his paper "On the definition of Artificial Intelligence", in which he argues that a normative definition of intelligence should be the ability to adapt under insufficient knowledge and resources (AIKR) [29]. Much of the current discussion on defining AI is inspired by the paper, and I take the position that his analysis is largely valid. Intelligence stems not from the ability to solve problems, but the ability to adapt to solve new problems. This adaptation can be seen as a *derivative* of ability (in the calculus sense). Perhaps intelligence has not been perceived as a derivative of fluid ability for a while because derivatives are more subtle and more difficult to define or measure. Recent mainstream research trends suggest a growing interest in this concept, and the next section evaluates what I consider to be possibly productive avenues of future AI research and development justified by the the concept.

One more idea worth discussing in light of recent trends is the idea of the Turing Test in measuring the progress of AI. To briefly recap, the Turing Test is an idea spawned by Alan Turing that a thinking machine can be evaluated by a human's ability to distinguish its output from another human's. As a thought experiment, Alan Turing briefly fiddled with the idea of using the Turing Test as an evaluation metric for artificial intelligence [27]. Generally, Turing's idea is interpreted in the context of textual chat, or chatbots in modern parlance. Resurgence in the idea of the Turing Test as an appropriate measure for Artificial Intelligence has gained considerable ground in recent years due to the introduction of large language models, which are trained to statistically mimic corpuses of human text and commonly used in modern chatbots.

Ironically, Turing himself rejected the idea that the Turing Test is a necessary condition for intelligence, though it may be sufficient. As previously argued and explicated by Wang, an artificial intelligence with different sensorimotor experiences should not be compared to human intelligence, so the Turing Test should not be taken as a serious evaluation metric for AI [29].

5 Revised expectations with a fresh perspective

5.1 Promising contemporary AI research

I now present a consequent analysis and outline a few directions of contemporary AI research that seem more promising than others in light of the conclusion that AI should focus on adaptive ability rather than ability itself, and in particular the adaptive ability of an agent given a fixed amount of computation and time under the concept of adaptation under insufficient knowledge and resources.

First, the dedication to using standard machine learning benchmarks (CIFAR, GLUE, etc) has largely driven progress in the deep learning community, leading to amazingly practical results. While benchmarks are still important, the preceding analysis suggests that the particular benchmarks are not as important for artificial intelligence as defining a rigorous standard for measuring the *rate* of progress on a benchmark given a standardized baseline. The success of a particular AI method should be measured as the rate of adaptation against the computation and time expended (ex. parameter size for deep learning models) as well as against the preceding amount and quality of data. While the enthusiasm for transfer learning using foundation models is justified for practical purposes, foundation models receive too much pretraining to meaningfully evaluate a model's adaptive ability given previous experience, and thus are ill suited in exploring the science of artificial intelligence. Some of the conclusions reached here are echoed by François Chollet [5].

A natural area of contemporary research to look at given the preceding analysis is metalearning, which directly addresses the need for deep learning models to adapt to changing (statistical) distributions. A representative method called MAML [9] directly optmizes for the derivative of adaptation with respect to datasets of disparate distributions. Metalearning is the obvious contemporary answer to the need to account for distributional shifts. However, it leaves the implicit assumption that "probability theory is the best way to encode uncertainty" untouched, so its success relies on probability theory's ability to not just account for uncertainty in a statistically stationary environment but also uncertainty in a statistically dynamic environmnent. At the present time this ability is not self evident. In addition, current metalearning relies on the curation of large datasets with at least some human intervention. It is then costly and confusing to measure the technique's actual ability to react to changing

environments since the objective also becomes the optimization metric, and "a change in distribution" is almost by definition unexpected (see discussion of hidden models in "Deep analysis of deep learning" section). Directly optimizing for the objective has always been a double edged sword in machine learning.

Another area of contemporary research addressing the problem of adaptability is the field of probabilistic causality, popularized by Judea Pearl in his book "Causality" [16]. Pearl builds on the idea of Bayesian networks to extend probability theory to include the idea of **interventions**, or deliberate changes to statistical function model. This more directly addresses probability theory's inadequacy of accounting for inevitable changes in distribution. However, instead of wholly rejecting probability theory Pearl adds a mechanical model on top of probability to account for distributional shifts. It remains to be seen whether Pearl's Bayesian-mechanical model is enough to handle distributional shifts according to probability theory or whether it is simply an inadequate band-aid on an actually severe wound. One point against Pearl's model is that interventions are discrete and pre-defined, so it is difficult to back-propagate or optimize on the concept of interventions according to causality theory.

A fringe research program led by Pei Wang explores the possiblity of using "non-axiomatic logic" to model a system that can adapt under insufficient resources and knowledge. It is worth mentioning here since many ideas contained in this manuscript are inspired by him. It remains to be seen whether a system based on non-axiomatic logic can succeed in creating an intelligent system that is practical in the physical world.

Lastly I briefly note that the current trend of relying on extended computation with chain-of-thought reasoning in LLMs is not likely to lead to progress in AI. It may lead to an increase of the capabilities of LLMs, but the preceding analysis shows that not absolute capability but capability with respect to the amount of computation and time expended is more important in evaluating the intelligence of a system.

5.2 A sobering perspective on the likely progress of AI

In contrast to prognosticators of imminent AGI or even an explosion of superintelligence (a hopeless idea devoid of meaning), this manuscript advocates for a sober perspective on the likely progress of AGI.

The preceding analysis of the definition of AI suggests that much of contemporary AI research may actually not lead to AI, since we have established

that the *science* of AI is inadequate. Currently both research and industry value immediate practical impact and could be classified as *engineering* and *computer science* endeavors. This is natural because if AI is to be a science, AI systems must eventually be proven in some kind of computational environment. However the danger is that computer science and algorithmic studies become conflated with AI studies [29]. Their work is not to be denigrated, but their classification and resulting significance should be more thoroughly questioned.

Many artificial intelligence researchers are also roboticists, not simply interested in adaptation under an arbitrary environment, but capability and adaptation under the physical universe we live in. A large part of human intelligence intuitively is human's ability to understand and adapt within the physical world, and many people would like intelligent agents to adopt drudgerous but necessary physical tasks. Unfortunately, in contrast with the hopes of reinforcement learning researchers, it could be the case that while humans have a special adaptive and learning capabilities apart from other animals, humans along with many animal's cognition are already highly adapted by evolution to the physical world in a manner difficult to replicate in the functional sense. Neuroscience has not established a consensus on whether animal perception and planning are innate, tabula-rasa learning, or some combination (and if so the characterization of the split). Thus there is no existence proof of purely tabula-rasa methods for training physically intelligent and capable agents within a biologically plausible lifetime that justifies using deep learning methods for robots. One could reasonably question to what degree human capability in the physical world is in fact innate and how much is actually intelligent according to the definition advocated in this manuscript.

One may also speculate that current deep learning methods struggle with practical physical reasoning because they do not have an adequate sensorimotor system to evaluate a real physical environment. This can be described as the **embodiment hypothesis**, where an adequate embodiment is needed for an intelligent agent to adapt within a particular kind of environment. The idea of physical grounding is related, although the embodiment hypothesis makes a more specific claim. I focus on human (or animal) sensorimotor systems in explaining physical intelligence because human interactions in modes such as natural language rely on a shared understanding of the physical and social world, which current large language models seem to not have acquired.

The embodiment hypothesis is also supported by ideas from causality theory; although causal models could be inferred from a purely data driven method

and causality researchers tend to explore purely data driven methods, the theory makes clear that physical intervention is the most direct and clear way to learn causal models. The well established scientific idea of a critical experiment is a related concept. Past research in robotics and reinforcement learning suggests the value in relying on non-realtime data (known as offline RL) because realtime data is costly, but perhaps what is really needed is a method to aggregate non-realtime data to suggest appropriate interventions, and a way to effectively propagate reasoning entailments from these interventions to the entire cognitive system when they are actually performed. SGD as a candidate for strong propagation of interventional information is questionable. First, would SGD be effective in handling very little interventional data? More importantly, what should be propagated since there is not necessarily an error involved, but rather competing hypotheses? Answering these questions may lead to productive research.

The actual role of intelligence as defined in this manuscript in explaining human capabilities is brought into question, which is a radical departure from the Western notion that human intelligence leads to human capability. If the goal really is to adapt within settings practical to humans, wouldn't it be necessary to develop appropriate sensorimotor capabilities first? This has been a separate challenge since the dawn of robotics. How important is the cultural environment of humans in explaining current human capabilities? Are we too obsessed with the mind according to the Cartesian split of mind and body that we completely forgot about the body and environment? The embodiment hypothesis posits that an agent becomes less effective at reasoning about an environment the more detached its sensorimotor system is from it.

It is worth noting that the embodiment hypothesis is an underexplored area and has not been accepted nor rejected in AI research, although mainstream AI researchers are warming to the idea.

Ultimately this manuscript reaches two conclusions regarding future development of artificial intelligence. First, the theoretical foundations of contemporary AI research are inadequate, leading to a lack of structure and direction. Second, even if abstract research on AI progresses rapidly, it is not obvious that this will directly lead to systems that can effectively reason about the physical world at the human scale. This is not a complete repudiation of industry or academic progress labelled as artificial intelligence. For example, the analysis leaves open the possibility that deep learning, machine learning, and even proper AI research can lead to advances in medical, quantum, and other specialized fields. It in fact advocates for the likeliness of AI agents that

effectively understand and adapt to human interactions in the digital, Internet world. However, the analysis also cautions against the natural tendency of people to anthropomorphize AI, despite the original goal of AI research being the replication of human capabilities. We should use the term AI less liberally and evolve our understanding of the original goal in light of recent evidence.

References

[1] *Amdahl's Law*. In: *Wikipedia*. Dec. 7, 2024. URL: https://en.wikipedia.org/w/index.php?title=Amdahl%27s_law&oldid=1261657056.

[2] Michele Banko and Eric Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. The 39th Annual Meeting. Toulouse, France: Association for Computational Linguistics, 2001, pp. 26–33. DOI: 10.3115/1073012.1073017. URL: http://portal.acm.org/citation.cfm?doid=1073012.1073017.

[3] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. July 12, 2022. DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258 [cs]. URL: http://arxiv.org/abs/2108.07258. Pre-published.

[4] Rodney A Brooks. "Elephants Don't Play Chess". In: (), p. 12.

[5] François Chollet. *On the Measure of Intelligence*. Nov. 25, 2019. DOI: 10.48550/arXiv.1911.01547. arXiv: 1911.01547 [cs]. URL: http://arxiv.org/abs/1911.01547. Pre-published.

[6] Daniel Crevier. *AI : The Tumultuous History of the Search for Artificial Intelligence*. New York : BasicBooks, 1993. 406 pp. ISBN: 978-0-465-00104-0. URL: http://archive.org/details/aitumultuoushist00crev.

[7] Jia Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. URL: https://ieeexplore.ieee.org/abstract/document/5206848.

[8] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805. Pre-published.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. July 18, 2017. DOI: 10.48550/arXiv.1703.03400. arXiv: 1703.03400 [cs]. URL: http://arxiv.org/abs/1703.03400. Pre-published.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[11] Douglas Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. 1st ed. USA: Addison-Wesley Longman Publishing Co., Inc., Nov. 1989. 372 pp. ISBN: 978-0-201-51752-1.

[12] Warren McCulloch and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[13] Marvin Minsky. *A Framework for Representing Knowledge*. MIT, 1974.

[14] Marvin Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, Sept. 22, 2017. ISBN: 978-0-262-34393-0. DOI: 10.7551/mitpress/11301.001.0001. URL: https://direct.mit.edu/books/monograph/3132/PerceptronsAn-Introduction-to-Computational.

[15] Alan Newell and Herbert Simon. "Report on a General Problem Solving Program". In: *IFIP Congress* 256 (1959).

[16] Judea Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009.

[17] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc., 1988.

[18] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. Feb. 26, 2021. DOI: 10.48550/arXiv.2103.00020. arXiv: 2103.00020 [cs]. URL: http://arxiv.org/abs/2103.00020. Pre-published.

[19] David Rumelhart and James McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, 1986. ISBN: 978-0-262-29140-8.

[20] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0.

[21] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Pearson Education, Inc., 2010.

[22] Rich Sutton. *The Bitter Lesson*. Mar. 13, 2019. URL: http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[23] Christian Szegedy et al. *Intriguing Properties of Neural Networks*. Feb. 19, 2014. DOI: 10.48550/arXiv.1312.6199. arXiv: 1312.6199 [cs]. URL: http://arxiv.org/abs/1312.6199. Pre-published.

[24] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2006.

[25] Sebastian Thrun et al. "Stanley: The Robot That Won the DARPA Grand Challenge". In: *Journal of Field Robotics* 23.9 (Sept. 2006), pp. 661–692. ISSN: 1556-4959, 1556-4967. DOI: 10.1002/rob.20147. URL: https://onlinelibrary.wiley.com/doi/10.1002/rob.20147.

[26] Tom Dupre la Tour et al. "A Finer Mapping of Convolutional Neural Network Layers to the Visual Cortex". In: NeurIPS. Oct. 12, 2021. URL: https://openreview.net/forum?id=EcoKpq43Ul8.

[27] Alan Turing. "Computing Machinery and Intelligence". In: *Mind* (1950).

[28] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 2, 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762. Pre-published.

[29] Pei Wang. "On Defining Artificial Intelligence". In: *Journal of Artificial General Intelligence* 10.2 (Jan. 1, 2019), pp. 1–37. ISSN: 1946-0163. DOI: 10.2478/jagi-2019-0002. URL: https://www.sciendo.com/article/10.2478/jagi-2019-0002.

[30] David Yarowsky. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods". In: *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* -. The 33rd Annual Meeting. Cambridge, Massachusetts: Association for Computational Linguistics, 1995, pp. 189–196. DOI: 10.3115/981658.981684. URL: http://portal.acm.org/citation.cfm?doid=981658.981684.

[31] Jianyu Zhang and Léon Bottou. "Learning Useful Representations for Shifting Tasks and Distributions". In: *International Conference on Machine Learning* (2023). URL: https://arxiv.org/abs/2212.07346.