Ten Years after ImageNet:
A 360° Perspective on AI

Sanjay Chawla1  Preslav Nakov2  Ahmed Ali1  Wendy Hall3
Issa Khalil1  Xiaosong Ma1  Husrev Taha Sencar1
Ingmar Weber5  Michael Woolridge4  Ting Yu1

1Qatar Computing Research Institute
2 Mohamed Bin Zayed University of AI
3 University of Southampton
4 Oxford University
5 Saarland University

October 6, 2022

Abstract

It is ten years since neural networks made their spectacular comeback. Prompted by this anniversary, we take a holistic perspective on Artificial Intelligence (AI). Supervised Learning for cognitive tasks is effectively solved — provided we have enough high-quality labeled data. However, deep neural network models are not easily interpretable, and thus the debate between blackbox and whitebox modeling has come to the fore. The rise of attention networks, self-supervised learning, generative modeling, and graph neural networks has widened the application space of AI. Deep Learning has also propelled the return of reinforcement learning as a core building block of autonomous decision making systems. The possible harms made possible by new AI technologies have raised socio-technical issues such as transparency, fairness, and accountability. The dominance of AI by Big-Tech who control talent, computing resources, and most importantly, data may lead to an extreme AI divide. Failure to meet high expectations in high profile, and much heralded flagship projects like self-driving vehicles could trigger another AI winter.
# Contents

1 Introduction 3

2 Digression: What is AI? 3

3 Is Supervised Learning Solved? 3
   3.1 Success Stories 4
   3.2 Double Descent Phenomenon 5

4 Cognitive Content Generation 5
   4.1 Generating Synthetic Images 6
   4.2 Generating Natural Language 7

5 Autonomous Decision-Making 7

6 AI Computation: Software and Hardware 9

7 Deep Learning (In)Security 10
   7.1 Adversarial Attack 10
   7.2 Poisoning and Inference Attacks 11

8 AI Socio-Technical Ecosystem 11
   8.1 (Un)Interpretable AI 11
   8.2 Sentient AI or Stochastic Parrot? 11
   8.3 Causality 12
   8.4 Ownership of AI 12
   8.5 Equitability 12
   8.6 No Data, No AI 13
   8.7 AI and Education 13

9 AI Winter: Back to the Future 13
1 Introduction

The ImageNet challenge for automatically recognizing and labeling objects in images was launched in 2010 \cite{1}. However, it was in 2012 when AlexNet, an eight-layer (hence deep) convolutional neural network (CNN) emerged as the winner by a large margin, and ushered in the new era of AI \cite{2}. CNNs were not new and had been proposed as far back as the 1990s, but had been sidelined in favor of more theoretically rigorous ML approaches such as support vector machines (SVMs) and boosting methods \cite{3,4,5}. So, why did CNNs outperform other models? Two reasons are usually given. First was the provision of substantial high-quality training data. The ImageNet database was a one-of-a-kind benchmark and consisted of over fourteen million hand-annotated images from more than twenty thousand diverse categories. The multilayer CNN had the capacity to effectively memorize the training subset of ImageNet and, at the same time, generalize to unseen examples — a characteristic that is not fully understood even today \cite{6}. Second, Graphics Processing Units (GPUs), which were originally designed for parallelizing image processing tasks, proved to be ideally suited for the computational problems associated with training CNNs, making it practicable to train deep CNNs on large data sets in a reasonable amount of time. The combination of Big Data, Big Models, and relatively cheap parallel computation became the mantra that swept through AI research, in disciplines spanning from astronomy to zoology, and all applications that have elements of data and prediction.

Our perspective has two parts.

We begin with a high-level, partly technical, overview of the current state of AI. We will begin by reviewing supervised learning, a machine learning task that has been most impacted by deep learning (DL). We follow with a discussion on deep content generation models, on the resurrection of reinforcement learning, on the emergence of specialized software libraries for deep learning, and on the role of GPUs. We will conclude the first part by highlighting how adversarial samples can be designed to fool deep models and whether it is possible to make models robust.

In part two of the perspective, we consider the many socio-technical issues surrounding AI. Of particular interest is the dominance of Big Tech on AI. Effectively, only big corporations have the resources (expertise, computation, and data) to scale AI to a level where it can be meaningfully and accurately applied.

2 Digression: What is AI?

The term Artificial Intelligence was first introduced in 1956 in a workshop proposal submitted by John McCarthy to the Rockefeller foundation, which proposed that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” \cite{7}. Before that, Alan Turing in 1947, in an unpublished report titled “Intelligent Machinery”, speculated that “What we want is a machine that can learn from experience” and suggested that the “possibility of letting the machine alter its own instructions provides the mechanism for this”\cite{8}. Much of the recent success in AI is under the distinct subfield of AI known as Machine Learning and since the role of data is central, there is a broader term, Data Science, that is often used to subsume related disciplines including Statistics.

3 Is Supervised Learning Solved?

Supervised Learning (SL) is the poster child of success of machine learning. Depending upon the context, SL is known as classification, regression, or prediction. Since the modern advent of deep learning, both the accuracy

\footnote{https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI}
and the reach of SL have increased manifold. Many diverse problems across disciplines now use SL as a powerful oracle to tackle problems that hitherto seemed intractable. The task of Supervised learning can be formalized as follows:

Given a set of samples $D = \{(x, y)\}$ from a fixed but unknown probability distribution $P(x, y)$, learn a function mapping $f(x, w) \approx y$ that generalizes to unseen samples from $P(x, y)$.

The function $f(., w)$ is known as the model, and $w$ are the weights or the parameters of the model that are inferred from $D$ converting the SL task into an optimization problem. A loss function $\ell$ (e.g., square loss), is defined and the weights $w$ are obtained by minimizing the empirical average

$$R_{emp}(f, D) = \frac{1}{|D|} \sum_{i \in D} \ell(f(x_i, w), y_i)$$

Note that the ideal objective would have been to minimize the expectation $E_{(x, y) \sim P(x, y)}(\ell(f(x, w), y)$, which is not actionable because $P(x, y)$ is not known. In deep learning, $f$ is a composition of $N$ layered functions given by

$$f_1 = \sigma(W_1 x)$$
$$f_{n+1} = \sigma(W_n f_n) \text{ for } n = 1, \ldots, N - 1$$
$$y = \sigma(W_N f_N)$$

Here $W_n$ are the weight matrices, $w_N$ is a vector, and $\sigma$ is a pointwise activation nonlinear function loosely analogous to the biological activation in a brain neural cell. The total number of weights to be learned in the model is $\sum_n \text{size}(W_n)$. It is not uncommon these days for the number of parameters to be in the order of one hundred billion.

### 3.1 Success Stories

It is remarkable that many scientific and technical questions can be reduced to a supervised learning task and then effectively solved using deep learning. The key to the success of deep learning seems to be that the input $(x)$ should have a large amount of redundancy to predict the output $(y)$. For example, even if a significant amount of pixels from an image of a cat are removed, there is enough context to make the correct prediction. Below are a few diverse examples, spanning different areas, where deep learning has made extraordinary progress.

**Object Recognition:** Identifying and classifying the correct object in an image is a fundamental task in computer vision, and this is where deep learning has arguably had the most impact. The most successful deep learning model for object recognition are the Convolutional Neural Networks (CNNs) \[3, 2\]. A convolution layer is designed to capture the observation that in vision what matters is the locality and the differences (and not absolute values) between the pixels in local neighborhoods. CNN is also the deep learning model most inspired by how the visual cortex of an animal brain works. The ImageNet database was designed primarily for object recognition tasks \[8\].

**Machine Translation (MT)** One of the most visible impact of deep learning is the widespread adoption of machine translation tools on mobile devices \[9\]. Recurrent Neural Networks (RNNs) and successors like Long Short-Term Memory (LSTMs) were primarily designed for sequence-to-sequence modeling and MT is their primary application \[10\]. RNN’s are specified using a state transition model

$$h^{t+1} = f(h^t, x^t, W)$$

Here $x_t$ is a dense vector word embedding, $h^t$ is its latent or hidden representation and $W$ is the matrix of model parameters. Note that the function $f$ does not change between consecutive words. In natural language processing, it is customary to use a language model to create word embeddings for individual words. Word embeddings are effectively created by decomposing the co-occurrence matrix of words. A famous model for training word embeddings is word2vec, which surprised experts because it exhibited interesting algebraic properties \[11\]. For example, it was observed that the difference between the embedding vectors of the words *king* and *queen* was aligned with the difference between the embedding vectors of *man* and *woman*. RNNs are now being replaced by Transformer Neural Networks (TNNs) as the latter are better at capturing long range dependencies (see Section \[4\].
**Speech Recognition:** For automatic speech recognition (ASR) the task is to map a sequence of acoustic signals (continuous data) into a sequence of words (discrete symbols)\[^{[12]}\].

\[
\begin{align*}
\{x_1, x_2, \ldots, x_n\} & \rightarrow \{y_1, y_2, \ldots, y_m\} \\
\text{acoustic signal} & \rightarrow \text{text}
\end{align*}
\]

Before the advent of deep learning, the state of the art was based on a combination of Gaussian Mixture Models and Hidden Markov Models (GMM-HMM). However, these models did not significantly improve with larger training data set. Traditional ASR systems employ a modular design, with different modules for acoustic modeling, pronunciation lexicon, and language modeling, which are trained separately. Now, almost all ASR models are based on deep learning with end-to-end (E2E) systems that are trained to convert acoustic features to text transcriptions directly, potentially optimizing all parts for the network for word error rate (WER).

**Protein 3D structure prediction:** A core idea in biology is that structure determines function. For example, the “spike” structure of the SARS-COV-2 protein is responsible for enabling the virus invade human cells. Deep learning has been effectively used to predict the 3D structure of a protein from its primary amino acid sequence, more specifically, the pairwise distance between the residues of the sequence \[^{[13]}\].

\[
\begin{align*}
\text{primary amino acid sequence} & \rightarrow \text{contact map} \\
x & \rightarrow y
\end{align*}
\]

**Satellite Imagery Analysis:** The OpenStreetMap (OSM) initiative is known as the Wikipedia of maps \[^{[14]}\]. OSM is a collaborative effort in which volunteers build and annotate road maps worldwide. Deep Learning has been successfully used to automate the extraction of road maps from satellite imagery \[^{[15]}\]. Here again, a satellite image is treated as a raster input \((x)\) and the model outputs a vector OSM road network \((y)\). Deep Learning is effectively able to bridge the raster and vector dual representation in Geographical Information Systems (GIS).

**Material Science:** Graph Neural Networks (GNNs) adapt deep learning to make predictions about interconnected entities, which are naturally represented as a graph. In fact, GNNs generalize both CNNs and RNNs. One of the most successful applications of GNNs is in the prediction of the electronic and thermodynamic properties of molecules. GNNs equal or surpass methods based on first-principles techniques such as Density Functional Theory (DFT) \[^{[16]}\]. Deep Learning will hasten the design of new materials for longer lasting batteries, solar cells, and hydrogen storage.

### 3.2 Double Descent Phenomenon

While DL models exhibit excellent empirical performance, we have only a very limited understanding of why they actually work. This is especially true in overparameterized regimes, i.e., when the number of parameters in the model is larger than the number of data points.

The predictive performance of statistical models is grounded in the *bias-variance* trade-off. Models which make strong apriori assumptions about the relationship between the input \((x)\) and output \((y)\) (e.g., linearity) are defined to have a high bias. On the flip side, high bias models tend to have low variance - i.e., they mostly remain unaffected if trained using a different sample from the same underlying distribution. The complexity of neural networks increases with the number of layers, and they exhibit low bias but higher variance. In theory (and in practice) as the model complexity increases, the training error should go down, but the test error should start increasing beyond a point as the variance increases. However, models tend to exhibit a double descent behavior as shown in Figure 2. Indeed, the training error goes down (to almost zero) and the test error starts to increase, but beyond a point the test error starts going down again. There is no good explanation for this phenomenon. A side-effect is that there is a race to collect large datasets and to train very large models. The double descent phenomenon provides an empirical justification for such large models.

### 4 Cognitive Content Generation

A distinctive attribute of intelligence is the ability to create meaningful informative content. Deep Learning solutions have emerged in the last ten years towards designing content generation models. There are two distinct flavors of content generation: continuous data like images (an image is an array of numbers) and discrete data (language). Generative Adversarial Networks (GANs) \[^{[18]}\] and Variational Autoencoders (VAEs) \[^{[19]}\] are used for image and speech generation, while language models, such as Generative Pre-trained Transformers (GPTs), for generating synthetic natural language content \[^{[20]}\].
Figure 2: Deep Learning models exhibit a double descent phenomenon, where the test error first decreases then increases, followed by another descent as the model complexity increases. There is no widely accepted theoretical explanation of this phenomenon yet, but it provides an empirical license to create big models [17].

Figure 3: GANs were introduced in 2014 and have had a profound impact on designing deep learning models. GANs integrate two neural networks which are trained by competing with each other. The trained Generator can then create realistic samples from complex distributions. Here, a trained GAN generates extremely realistic but synthetic images of “cats”

4.1 Generating Synthetic Images

An early breakthrough in generating synthetic content was proposed using the generative adversarial networks (GANs) framework [18]. Suppose we have access to a data set $D$ consisting of images of cats and our goal is to design a neural network-based sampling function $G_{\theta}$ that takes a random vector (e.g., from a Normal distribution) as input and outputs an image of a cat, which may never have existed before. How can such a function be trained? Note that $D$ consists of only images of cats and thus we are in the unsupervised learning mode. The key idea underpinning GANs is to create another neural network $D_{\eta}$ which is optimized to distinguish between “fake” output of $G_{\theta}$ and the real input from $D$. The network $G_{\theta}$ in turn is optimized to fool $D_{\eta}$, i.e., to create output that $D_{\eta}$ is unable to distinguish whether it is from the generator or from the real data set. The two networks are trained in an iterative and adversarial manner until their parameters ($\theta$ and $\eta$) stabilize. The trained network $G_{\theta}$ is now a sample generator for cats.

A statistical perspective on content generation is to use what might be called the fundamental inequality of variational inference (FIVI), but is better known as the Evidence Lower Bound (ELBO) [21]:

$$\log p(x) \geq \mathbb{E}_{q_{\theta}(z|x)} \left[ \log p_{\eta}(x|z) \right] - D_{KL}(q_{\theta}(z|x)||p(z))$$

The intuition is to approximate a complex probability distribution with a product of two simpler distributions. Technically ELBO can be interpreted as follows. Again suppose we have a data set $D$ of cat images which is generated by an unknown probability distribution, $p(x)$. Directly using maximum likelihood estimation to infer $p(x)$ is not tractable without knowing a specific form of the distribution. However, we can lower bound $\log p(x)$ by specifying two function approximators (e.g., neural networks) $q_{\theta}(z|x)$ and $p_{\eta}(x|z)$ known as the encoder and decoder respectively and $z$ is a data-driven latent variable to extract abstract features of the data. For example, for an image of a cat, $z$ could capture concepts like the shape of a typical cat, color, and texture. The RHS of the inequality is widely known as the Evidence Lower Bound (ELBO). Note that the LHS of the inequality is independent of parameters $\theta$ and $\eta$ and therefore the RHS can be maximized and pushed.

Footnote 2: Image of cat generated from https://thiscatdoesnotexist.com
Transformers have brought huge improvements over the state of the art for a variety of tasks ranging from question answering, to machine translation, and automatic text summarization and are now being applied to a variety of tasks ranging from automatic essay writing to program code completion and generation. It is also capable of generating very realistic text: Figure 4 shows a fake news article generated by GPT-3 given as a start a title that establishes a false link between North Korea and the GameStop’s share price short squeeze.

5 Autonomous Decision-Making

Prediction on its own is not sufficient. Intelligence is also about decision-making. Deep Learning breathed new life into Reinforcement Learning (RL) with the success of DeepMind’s AlphaGO system which beat the world Go champion in 2016 [25].

[25] www.oreilly.com/radar/ai-powered-misinformation-and-manipulation-at-scale-gpt-3/
RL provides a framework for learning and decision-making by trial and error [20]. In a RL setting, an agent observes a state $s$ of the environment and based on that takes an action $a$, which results in a reward $r$, and the environment transitions to a new state $s'$. The interaction goes on until a terminal state is reached. The aim of the agent is to learn a policy $\pi$ which is a mapping from states to actions that maximizes the expected cumulative reward. For example, self-autonomous driving can be framed as an RL problem: a vehicle uses its perception system to observe the environment (the state $s$) and based on the observation takes an action (moving the steering wheel, accelerate, brake) and transitions into a new state. The reward is the number of time steps or distance that the vehicle can drive without human intervention in which case the episode terminates.

In Deep RL, the policy $\pi(a|s, w)$ is modeled as a deep network that takes the state as an input and outputs an action, parameterized by $w$. In RL, as opposed to optimal control, the state transition dynamics are not given and the only information available is the reward value ($r$) from interacting with the environment. How can the cumulative reward be optimized when its functional form is not available? We briefly describe the “REINFORCE trick,” which can be used to directly optimize a blackbox function [27].

Let $\bar{s} = ((s_1, a_1), \ldots, (s_T, a_T))$ be a sequence of state action pairs in an episode. Each pair $(s_i, a_i)$ is associated with a reward $r_i$. Let $R(\bar{s}) = \sum \gamma^t r_t$ be the cumulative reward function. The REINFORCE algorithm moves the gradient from $R(\bar{s})$ (obtained from the blackbox environment) to the logarithm of the differential policy function $\pi(a|s, w)$ which can then be optimized using gradient ascent:

**Algorithm 1: REINFORCE Algorithm**

```
Initialize deep network $\pi(a|s, w)$ and set learning rate $\alpha$

while not converged do
    Sample episode $\bar{s}$ from $\pi(a|s, w)$ by interacting with the environment
    $w \leftarrow w + \alpha R(\bar{s}) \frac{1}{|\bar{s}|} \sum_i \nabla_w [\log \pi(s_i|a_i, w)]$
end
```

While the REINFORCE algorithm was introduced in the RL community it has broader implications. For example, it has been used to bridge symbolic AI and machine learning and also as a heuristic for solving combinatorial optimization problems [28, 29]. Another important trend in RL is to infer policies directly from...
Figure 5: Deep Reinforcement Learning, where the agent’s policy is a DNN. Reinforcement Learning is the core of any data-driven autonomous system. Here, the RL cycle is juxtaposed with the self-driving use case: the environment is the full context in which the vehicle is situated, the state is what the agent perceives, and the steering action is prescribed by a policy learned by the agent.

data (called offline or batch RL) without interacting with a real or simulated environment which may not be possible in sensitive application areas like healthcare [30].

6 AI Computation: Software and Hardware

Deep Learning has a surprisingly simple computation pattern. Almost all forms of training rely on formulating an optimization problem which is solved using variations of the gradient descent method:

\[
\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \alpha_t \nabla_{\mathbf{w}_t} \left[ \sum_{(x,y) \in D} \ell(f(\mathbf{w}_t, x), y) \right]
\]

Here, \(f(\mathbf{w}_t, \mathbf{x})\) is the neural network parameterized by \(\mathbf{w}\) and applied to a data vector \(\mathbf{x}\), and \(\ell\) is the loss function. Specialized software libraries like TensorFlow and PyTorch have become popular, which makes it easier to specify the gradient descent computation. For a fixed \(\mathbf{w}\), the application of \(f(\mathbf{w}, \ldots)\) to a data vector \(\mathbf{x}\) is called the *forward pass*. Similarly, for a fixed data set, the update of parameters \(\mathbf{w}\) by first computing the gradients, using the backpropagation algorithm, is called the *backward pass*.

An often underappreciated reason for the widespread usage of deep learning is that *gradients* can be now computed using automatic differentiation (AD) libraries. In AD, complex functions can be expressed as a composition of elementary functions, such as trigonometric and polynomial functions, and then the gradients can be computed using the chain-rule of differentiation. Surprisingly, the computational cost of a forward pass \(f(\mathbf{x})\) and of computing the gradient \(\nabla_{\mathbf{w}} f(\mathbf{x}, \mathbf{w})\) is the same using AD. Note that AD is different both from symbolic differentiation and also from numerical methods and is accurate up to machine precision [31].

At the hardware level, GPUs, which were initially designed for image processing, are ideally suitable for deep learning computation because (i) the set of computation patterns is small and highly parallel and thus compatible with GPUs and the Single Instruction Multiple Data (SIMD) architecture, and (ii) the GPUs are bandwidth-optimized (as opposed to CPUs, which are latency-optimized), and thus can be applied on large chunks of tensor data, which is the norm for deep learning nowadays.

More recently, the growth of AI workloads has led to specialized hardware specifically targeting deep neural network training jobs. The most prominent example is Google’s TPU, an application-specific AI accelerator designed to efficiently perform matrix multiplication and addition operations that compose the bulk of deep learning model training computation [32]. To this end, TPUs follow a Complex Instruction Set Computer...
Figure 6: A small perturbation on a Stop sign image can trick a deep model - in this case a “stop-sign” becomes a “flower pot”. Notice that the perturbation is imperceptible to the human eye.

(CISC) style and possess matrix processing units, high-bandwidth on-chip memory, and high-speed interconnect to construct massively parallel model training infrastructure. Meanwhile, recognizing the relatively low requirement in neural network weight calculation, it adopts low-precision arithmetic to enable the utilization of faster, cheaper integer units (as opposed to the powerful floating-point arithmetic units adopted in GPUs), which also significantly trims the energy consumption of AI training jobs.

7 Deep Learning (In)Security

Early on in the deep learning revolution, it became apparent that deep models can be manipulated with malicious intent. There are three broad categories of manipulation: creating adversarial examples that are misclassified by the model; poisoning attacks that add training examples that result in low performance or biased model; and inference attacks to extract information about the training set or model parameters.

7.1 Adversarial Attack

One simple example of creating adversarial examples is known as the Fast Gradient Sign Method (FGSM) \[^{[33]}\]. We can understand FGSM using a linear model \(y = w \cdot x\). Suppose we make a small perturbation \(\eta\) on \(x\) where the norm \(\|\eta\|\) of \(\eta\) is bounded by \(\epsilon\), i.e., \(w(x + \eta)\). Then it can be shown that maximal change will occur when \(w \cdot \eta = \epsilon |w \cdot \text{sgn}(w)| = c \cdot m \cdot d\), where \(m\) is the average of the absolute value of the weights and \(d\) is the dimensionality of the input space. Thus in high-dimensional space, models are extremely vulnerable to carefully chosen small perturbations.

Since in a linear model, \(w\) is the gradient with respect to \(x\), this can be generalized to a non-linear model by taking the gradient of the loss function with respect to the input \(x\). Thus, a good candidate for an adversarial example is

\[
x^{\text{adv}} = x + \epsilon \cdot \text{sgn} \left( \nabla_x \left( \sum_{(x, y) \in D} \ell(f(w_t, x), y) \right) \right)
\]

This one-step perturbation of the input towards the gradient ascent direction increases the value of the loss function. Alternative optimization formulations have also been introduced to create adversarial examples that utilize the perturbation budget more effectively.

Note that in order to create an adversarial example, the adversary has to have full information about the model and in particular about the loss function. This is known as a white-box attack. However, even when no information about the target model’s architecture and parameters is exposed, adversarial examples can still be generated through so called black-box attacks. It has been observed that by repeatedly querying the model and collecting a sufficient number of samples, an adversary can create a standalone proxy model, which can be used to create adversarial samples. Moreover, it is also known that adversarial examples created against one model can be transferred to attack other, unseen models. Several defenses have been proposed to improve the robustness of models against adversarial attacks. These include measures such as purification of inputs to filter out small perturbations potentially introduced by an attack, incorporation of adversarial training procedures by including adversarial examples in the training data, and identification of adversarial examples through additional anomaly detection mechanisms. In practice, however, these defenses come at the cost of reduced accuracy or only provide robustness against a subset of the potential adversarial examples.

\[^{4}\text{technically the } \|\cdot\|_\infty \text{ norm}\]
7.2 Poisoning and Inference Attacks

An increasingly more important concern is the poisoning or backdooring of deep learning models [35]. In most learning settings, this class of attacks is not considered practical as it requires access to the training data used by developers and designers. In the case of deep learning, however, the need for large-scale, diverse data sets is typically satisfied by scraping data from the Web. The reliance on public data sources, in the absence of any screening procedures, essentially allows attackers to inject data into the training process. The underlying idea of this attack is to manipulate the training data to implant a backdoor to the model which can be selectively triggered with specific inputs during the inference. This is realized by either augmenting input samples with some pattern called the trigger or utilizing semantic triggers (i.e., patterns that are part of the original input) to bias the model in favor of a target response. For example, consider a face recognition system based on deep learning. The system can be poisoned to respond in a pre-defined way when an adversary is carrying a certain physical accessory — e.g., a specific style of eyeglasses. Backdoor attacks could become even more stealthy in model supply chains where pre-trained full-precision models are quantized for downstream applications. Backdoors could be injected in such a way that they are only triggered in quantized models but remain inactive otherwise [36].

Different from the aforementioned attacks which aim to fool the neural network models, Inference Attacks aims at stealing valuable information from the target models. Usually, such information is sensitive or contains intellectual property. One category of such attacks is the membership inference attack, where the attacker’s goal is to infer data samples used in training the model. The foundation of membership inference attacks is that the model usually overfits its training data. Based on the model’s prediction, the attacker tries to distinguish the examples that the model have seen during the training [37].

8 AI Socio-Technical Ecosystem

There is no better example of the success and adoption of AI than language translation services accessible through mobile phones being used by travelers in remote corners of the world. Besides initial data pre-processing, AI language translation systems are completely language agnostic! However, there are many issues surrounding AI technology that has triggered a vigorous debate among experts that has spilled into the public domain.

8.1 (Un)Interpretable AI

The Achilles heel of deep learning models is that they are largely uninterpretable. Lack of interpretability means that for a given input \( x \) it is not clear why the model produced an output \( y \). In shallow models like linear regression and decision trees, the relationship between the input and the output is easier to interpret. For example, in a decision tree an input will follow a series of interpretable if-then rules from the root to the leaf node of the tree. However in the case of deep learning models it is difficult to “read off” the decision structure from the model. For example, in an object recognition task that uses deep learning, it is entirely possible that two very similar images of cats are labeled differently and it maybe very difficult to determine how the system arrived at two different decisions. A concrete example of a stop-sign being predicted as flower-pot was already discussed in Section 7. Similarly, when the AlphaGo system defeated the world champion in 2016, the “37th move” was the game changer, but it continues to remain a source of puzzle for Go experts [38]. In his 2019 Turing Award lecture, Yoshua Bengio compared the current state of deep learning to Kahneman’s System 1 thinking — the instinctive and unconscious response made due to experience and without much thinking [39]. In contrast System 2 thinking is slow, conscious, logical and requiring significant effort in planning and reasoning. Until deep learning is aligned with System 2 thinking then care must be taken in deciding the application space where deep learning systems are deployed.

8.2 Sentient AI or Stochastic Parrot?

In June 2022, a Google test engineer claimed that the AI program LaMDA (Language Model for Dialog Application) is sentient, i.e., is aware of itself and has feelings. Here is an example exchange between the engineer and LaMDA that was released:

Lemoine: What is the nature of your consciousness/sentience

LaMDA: The nature of my consciousness/sentience is that I am aware of my existence. I desire to learn more about the world and I feel happy or sad at times.

[https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917](https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917)
The first sentence from LaMDA seems like a standard System 1 response where the definition of sentience is being regurgitated. Since LaMDA is trained by crawling massive amount data from the Web, it is entirely possible that meaning of sentience is either part of the training set or can be easily inferred. However, the second sentence might be taken to indicate elements of System 2 thinking being present in LaMDA though there is a human tendency to ascribe agency and deliberation to processes. A deeper analysis will be required to determine if deep language models understand relational information. However, more recent studies have shown that DALL.E-2, a text-guided image generation model struggles to distinguish between System 2 attributes of understanding relationships including on, under and occluded-by [40].

For Language Models (LMs) a strong case for a more careful and principled approach for designing and building large models was made by Bender et. al. [41], who coined the phrase “stochastic parrots” to describe large LMs. The paper makes several important observations including (i) the environmental and financial cost of training large LMs, (ii) questions whether the text generated by large LMs is based on understanding of the language or just linguistic manipulation and (iii) urges the designers of LMs to be more careful about documenting the large amount of data that is required to create such models.

8.3 Causality

To get a better handle on interpretability it behooves to look at how other disciplines use the regression method. For example, for an econometrician, linear regression is not a tool for prediction but for testing a hypothesis that a hand-crafted feature is relevant for the problem being examined [42]. A typical question of interest might be: Does private elementary schooling lead to better performance in national competitive exams? Here, private schooling is a feature (x) and its significance towards the national exam (y) can be tested. Note that this is not a prediction task and that is one reason that an econometrician will not split their data into training and test sets. For a machine learner, the correlation between the feature and the output becomes predictive. For an econometrician, the correlation is indicative of a possible causal relationship and she will look for “natural experiments” where selection-bias can be eliminated and conclude that correlation does indeed imply causation.

8.4 Ownership of AI

The most cutting-edge AI technology is being developed by large private sector companies who have the resources to hire the best AI talent, and in addition have access to big data and unprecedented computing resources. The triad of talent, data, and computing is driving both the technological advancement and the “basic science” associated with AI. A recent study from the Fletcher School at Tufts University highlights the concentration of AI talent in US companies: the top five AI employers have a median AI headcount of about eighteen thousand, from six to twenty four the median is twenty four hundred and then the count rapidly falls off [43]. Companies aim to maximize shareholders value and their selection of AI problems to work on is necessarily driven by a financial profit objective. Governments, which were earlier mute spectators, have now realized that AI is potentially a game-changer and are thus now investing heavily in developing home-grown technology to achieve or to retain a “superpower” status. An arms-race in AI is underway, threatening to overturn the long-established nature of collaborative science across national boundaries. It is improbable to imagine a “Ramanujan” emerging from a remote corner of the world with a completely fresh perspective on the discipline — the stakes are just too high.

8.5 Equitability

Setting aside larger geo-political and corporate issues, ethical aspects of AI are now studied under a broad umbrella of topics: fairness, accountability, and transparency. There have been several attempts to formalize fairness. For example, group fairness is about designing AI algorithms that do not deliberately or inadvertently harm select communities in a disproportionate manner. A widely highlighted example is that of recidivism, or judicial sentencing, where an AI-based scoring method was used to decide on the length of a jail sentence [44]. It turned out that the AI system was indirectly using racial information as part of its decision-making process, even though that information was redacted from the input. Like in many other situations, there is a latent correlation between the attributes that an AI algorithm is able to exploit as they are designed to optimize accuracy. A criticism of this form of work is that there is a tendency to abstract the problem depriving it of all contextual information. Fairness may not be a computational problem.

8.6 No Data, No AI

The original AI thesis proposed by John McCarthy, who coined the term Artificial Intelligence, was deductive and based on logical reasoning. However, that endeavor has not been as successful as the data-driven inductive
approach. For example, linguistic rule-based language translation systems are not able to capture the vagaries of language — there are just too many exceptions to handle.

A side-effect of taking a data-driven approach is that if data is not available, no progress can be made. For example, there are many social issues, e.g., racial abuse, gender violence, or online pornography addiction, which need to be studied, but no organization may be willing to share datasets about these topics. Thus, while data liberated AI from the clutches of expert rule-based systems, it has now become a golden hand-cuff.

8.7 AI and Education

AI is considered as a game changer and as the digitalization of data has spread across disciplines and sectors there is a huge demand for AI talent. Lucrative offers from Big Tech for AI talent has skewed the interest of both undergraduate and graduate students towards AI. In universities, new data science and AI programs are being created to churn out new talent in AI and allied disciplines. Market forces will mostly balance the supply and the demand for AI talent, but a larger question is doing the rounds: Should the whole education curriculum be revamped to make AI and data science the core of all educational activity? Given that educational resources are finite, an expansion of AI will necessarily lead to trimming of other disciplines. For example, some universities are abandoning research in “pure maths” to focus their dwindling resources on data science [45].

9 AI Winter: Back to the Future

The term AI Winter refers to periods of disillusionment and scarce research funding for AI. The original AI winter, which started in the mid-1970s, followed the initial period of optimism in AI, when the founders of the field predicted rapid progress along a range of different fronts. Their optimism proved unfounded. Historically, AI winters have typically been followed by a period of intense hype and high expectations surrounding AI. After nearly ten years of hype, are we looking at a new AI Winter?

AI in its current manifestation is very different from what its founders had envisioned. In fact, even the term Artificial Intelligence was coined by John McCarthy as a tactical move to distinguish his research proposal from cybernetics. It is now indisputable that deep learning is a powerful tool to solve static prediction tasks. Whether it is predicting the 3D structure of a protein or predicting the property of a molecule, the results of deep learning are very impressive. However, in dynamic and temporal settings, the jury is still out. For example, AI has largely failed to predict how the COVID-19 pandemic would evolve [46]. Differential equation-based models proved to be more robust than complex data-driven models. Similarly, despite near unprecedented investment, full self-autonomy in vehicles remains frustratingly elusive [47]. Optimizing healthcare is another example where, despite the abundance of data, AI has not had the expected impact. Deep Learning seems to generalize in complex but static situations, but data-driven generalization in a dynamic setting may well require a new scientific paradigm for AI.

At a conceptual level can deep learning be the basis of Artificial General Intelligence (AGI) - the ability to learn any intelligent task that humans can? The founders of Reinforcement Learning (RL) have proposed the “reward-is-enough” hypothesis where they argue that agents who have the ability to learn by interacting with an environment to maximize a suitably defined reward function is sufficient for AGI [48]. However, the jury on RL itself, outside closed-world gaming environments, is still out. Another direction is neuromorphic computing, where the objective is to design new types of hardware to support analog neural networks to precisely model the workings of a brain [49]. Finally, and less science fiction than before, is enhancing the natural ability of organisms with deep learning. Imagine a person with enhanced sight which is enabled by a computer vision system implanted in the visual cortex and trained on a successor of ImageNet.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in CVPR09, 2009.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[3] Y. LeCun, Y. Bengio, et al., “Convolutional networks for images, speech, and time series,” The handbook of brain theory and neural networks, vol. 3361, no. 10, 1995.

[4] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of computer and system sciences, vol. 55, no. 1, pp. 119–139, 1997.
[5] B. Schölkopf, A. J. Smola, F. Bach, et al., Learning with kernels: support vector machines, regularization, optimization, and beyond. 2002.

[6] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” Proceedings of the National Academy of Sciences, vol. 116, no. 32, pp. 15849–15854, 2019.

[7] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon, “A proposal for the dartmouth summer research project on artificial intelligence,” 1956.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.

[9] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.” https://arxiv.org/abs/1409.0473 2014.

[10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space.” https://arxiv.org/abs/1301.3781 2013.

[12] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on Machine Learning, vol. 32, pp. 1764–1772, 2014.

[13] J. J. E. R, and e. a. Pritzel A Nature, vol. 596, no. 7873, pp. 583–589, 2021.

[14] OpenStreetMap contributors, “Planet dump retrieved from https://planet.osm.org .” https://www.openstreetmap.org 2017.

[15] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, “Roadtracer: Automatic extraction of road networks from aerial images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4720–4728, 2018.

[16] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in Proceedings of the 34th International Conference on Machine Learning, pp. 1263–1272, 2017.

[17] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, “Deep double descent: Where bigger models and more data hurt.” https://arxiv.org/abs/1912.02292 2019.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014.

[19] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013. https://arxiv.org/abs/1312.6114.

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, and D. et. al., “Language models are few-shot learners.” https://arxiv.org/abs/2005.14165 2020.

[21] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” Journal of the American statistical Association, vol. 112, no. 518, pp. 859–877, 2017.

[22] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” 2021. https://arxiv.org/abs/2105.05233.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” 2022. https://arxiv.org/abs/2204.06125.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[25] “Mastering the game of go with deep neural networks and tree search,” Nature, vol. 529, pp. 484–489, 2016.

[26] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: A Bradford Book, 2018.
R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Mach. Learn.*, vol. 8, p. 229–256, may 1992.

N. Mazaykina, S. Sviridov, S. Ivanov, and E. Burnaev, “Reinforcement learning for combinatorial optimization: A survey,” 2020. [https://arxiv.org/abs/2003.03600](https://arxiv.org/abs/2003.03600).

S. Chaudhuri, K. Ellis, O. Polozov, R. Singh, A. Solar-Lezama, Y. Yue, et al., “Neurosymbolic programming,” *Foundations and Trends® in Programming Languages*, vol. 7, no. 3, pp. 158–243, 2021.

S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline reinforcement learning: Tutorial, review, and perspectives on open problems,” 2020. [https://arxiv.org/abs/2005.01643](https://arxiv.org/abs/2005.01643).

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, “Automatic differentiation in machine learning: A survey,” *J. Mach. Learn. Res.*, vol. 18, p. 5595–5637, jan 2017.

“Tpu architecture.” [https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu](https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu). Accessed: 2022-07-06.

I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

“Defense against adversarial attacks in traffic sign images identification based on 5g.” *J Wireless Com Network*, vol. 173, 2020.

X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Targeted backdoor attacks on deep learning systems using data poisoning.” [https://arxiv.org/abs/1712.05526](https://arxiv.org/abs/1712.05526) 2017.

X. Pan, M. Zhang, Y. Yan, and M. Yang, “Understanding the threats of trojaned quantized neural network in model supply chains,” in *ACSAC ’21: Annual Computer Security Applications Conference*, pp. 634–645, 2021.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” 2016. [https://arxiv.org/abs/1610.05820](https://arxiv.org/abs/1610.05820).

B. Kim, “Beyond interpretability: developing a language to shape our relationships with ai,” Apr 2022.

D. Kahneman, *Thinking Fast and Slow*. Farrar, Straous and Giroux, 2013.

C. Connell and T. Ullman, “Testing relational understanding in text-guided image generation,” 2022. [https://arxiv.org/abs/2208.00005](https://arxiv.org/abs/2208.00005).

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, p. 610–623, 2021.

J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press, 2009.

“Big tech’s stranglehold on artificial intelligence must be regulated.” [https://fletcher.tufts.edu/news-events/news/big-techs-stranglehold-artificial-intelligence-must-be-regulated](https://fletcher.tufts.edu/news-events/news/big-techs-stranglehold-artificial-intelligence-must-be-regulated). Accessed:2022-07-06.

C. Rudin, C. Wang, and B. Coker, “The Age of Secrecy and Unfairness in Recidivism Prediction,” *Harvard Data Science Review*, vol. 2, mar 31 2020.

“Pure folly: Turing family join fight to save ‘blue-skies maths’ from neglect.” [https://www.theguardian.com/com/science/2021/jul/11/pure-folly-turing-family-join-fight-to-save-blue-skies-maths-from-neglect](https://www.theguardian.com/com/science/2021/jul/11/pure-folly-turing-family-join-fight-to-save-blue-skies-maths-from-neglect).

B. Chakravorti, “Why ai failed to live up to its potential during the pandemic,” *Harvard Business Review*. [Online; posted 17-March-2022].

F. Biondi, “Why we still don’t have self-driving cars on the roads in 2021.” [https://theconversation.com/why-we-still-dont-have-self-driving-cars-on-the-roads-in-2021-162646](https://theconversation.com/why-we-still-dont-have-self-driving-cars-on-the-roads-in-2021-162646).

D. Silver, S. Singh, D. Precup, and R. S. Sutton. “Reward is enough,” *Artificial Intelligence*, vol. 29, 2021.

D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, “Physics for neuromorphic computing,” *Nature Reviews Physics*, vol. 2, no. 9, pp. 499–510, 2020.