Next big challenges in core AI technology

Dengel, A.; Etzioni, O.; DeCario, N.; Hoos, H.H.; Li, F.F.; Tsujii, J.; ... ; Ghallab, M.

Next Big Challenges in Core AI Technology

Andreas Dengel[1], Oren Etzioni[2], Nicole DeCario[2], Holger Hoos[3], Fei-Fei Li[4], Junichi Tsujii[5], and Paolo Traverso[6(✉)]

[1] DFKI, Kaiserslautern, Germany
[2] Allen Institute for AI, Seattle, USA
[3] Universiteit Leiden, Leiden, The Netherlands
[4] Stanford University, Stanford, USA
[5] AIRC, Tokyo, Japan
[6] FBK, Trento, Italy
traverso@fbk.eu

Abstract. The field of AI is rich in scientific and technical challenges. Progress needs to be made in machine learning paradigms to make them more efficient and less data intensive. Bridges between data-based and model-based AI are needed in order to benefit from the best of both approaches. Many real-life situations cannot yet be addressed by current robots, demanding progress in perception, scene interpretation or group coordination. This chapter addresses some of the major scientific and technological challenges in core AI technology.

1 The Need to Address Scientific and Technological Challenges for an AI for Humanity

AI for Humanity is not only a matter of regulations, normative frameworks, legal, ethical, political, and social issues. AI for humanity needs also to address key scientific open problems. In spite of several AI success stories in the past, even going back to the 90's and before (see e.g., [1]), there is no doubt that the current impact and high expectations raised by AI is due, to a large extent, to recent successes in data intensive (supervised) machine learning, and especially to deep learning. Deep learning has led to impressive gains on most key areas of AI, such as computer vision, natural language understanding, speech recognition, and game playing. Considering the field for instance of computer vision, in the last ten years, deep learning techniques have achieved incredible results, moving the capability of machines to recognize thousands of everyday objects, sometimes better than humans (see, e.g., [2]). It is well known that certain important tasks in health care, like screening for diabetic retinopathy, are better performed nowadays by deep learning AI techniques than by doctors [3].

In spite of this significant progress, we still need a lot of work in research and a paradigm shift in AI to develop a real AI for humanity - a human centric AI. We need research to build AI systems that are able to augment and enhance people rather than replacing them, and to help humans by interacting with them and collaborating with them. Some key open research challenges are the following (see chapters 2 and 3).

B. Braunschweig and M. Ghallab (Eds.): Reflections on Artificial Intelligence for Humanity, LNAI 12600, pp. 90–115, 2021.
https://doi.org/10.1007/978-3-030-69128-8_7

- **Less data-intensive AI.** Most of the current deep learning techniques require a huge amount of training data. However, in certain applications, high volumes of data are not available, and in most cases, training deep neural networks is time consuming and requires a lot of effort. We need less data intensive approaches, e.g., along the lines of representation learning, such as unsupervised learning, unsupervised pre-training, domain adaptation, transfer learning, one-shot learning, zero-shot learning (see [4] for an overview).
- **Explainable AI.** One of the major problems of deep neural networks is their opaqueness, i.e., the lack of explainability of the results, the lack of explanation of how they work and why they lead an AI system to take some decisions. Most often, deep neural networks are essentially "black boxes". In several cases however, human-centric AI must be explainable. It must be, as much as possible, a "white box". A major research challenge is to develop techniques that provide the ability to understand deep neural networks such that humans can debug, interpret, control, and reason about them. AI systems should be able to explain the assumptions and criteria under which they take some decisions or provide some results. AI systems should be "auditable", i.e., they should be able to answer questions asked by humans and interact with them in an understandable way for humans. Moreover, if AI techniques are not understandable by humans, it is very difficult to build systems that interact with humans. As a result, it is difficult to keep humans in the loop and to give them true control over AI systems.
- **Trustworthy and verifiable AI.** One of the major potential outcomes enabled by AI techniques is the ability to build autonomous systems, such as self-driving cars. More generally, AI can be a key technology for the new generation of intelligent robots, drones, automated plants for Industry 4.0, transportation systems, medical systems for diagnosis and health care, etc. Most often, AI technology is part of safety-critical systems, where errors can have a tremendous impact on human life and/or the environment. The complexity and opacity of some AI techniques (e.g., deep learning) do not help. Research should provide trustworthy and verifiable AI techniques that guaranty safety for humans and environmental preservation. There is a need for interdisciplinary research joining competences in AI with competences in formal methods and software engineering, such as techniques for theorem proving, model checking, testing, and simulation.
- **AI for security and privacy.** Simple but very effective adversarial examples and attacks with even small imperceptible perturbations can compromise the results of deep learning systems, e.g., in image understanding. AI systems should be secure and resilient to such attacks. Moreover, most AI systems rely on personal data. Human-centric AI should guarantee confidentiality and privacy. AI systems collecting personal data can also give rise to societal and political problems. For example, personal profiling can lead to threats to democracy, as in the well-known case of Cambridge Analytica. Security and privacy should, however, be balanced with the need for sharing data for individual and social good. As an important example, take the case of personal health care data. We should guarantee the privacy of sensitive data for individuals but, at the same time, we should open up the way to science and progress in medicine by analyzing health data. We should not get to the point that we will fail to discover cures for chronic disease because of

privacy! Finding the right balance between privacy and the need for sharing data for social good is not only a matter of regulations and laws. A good example is the idea of the "Web of Clinical Data" (see [5]), where privacy and equality are guaranteed, but a huge amount of data about our health is available to researchers (even from private companies), who can use it for principled experimentation with new AI techniques for improved health care. There is a compelling need for an interdisciplinary research involving computer scientists, lawyers, and sociologists to address this issue.

- **Integrative AI.** Most of AI applications for individual and social good require integrating different kinds of AI technology. They require the computational modeling/mechanization of a diverse range of cognitive tasks, the scientific and theoretical/formal integration of different representations and reasoning techniques, e.g., symbolic (knowledge based and semantic representation) and sub-symbolic (numeric and probabilistic) representations, as well as data-driven learning and model-based (e.g., deductive) reasoning[1]. Human-centric AI systems should be able to combine data from different, highly heterogeneous sources (video, audio, social networks, crowd-sourced data, IoT, remote sensing, natural language source, non-structured and structured data) and to reason from these disparate data sources, using a variety of approaches (e.g., machine learning, deduction and knowledge reasoning).
- **The integration of perception, action, and human interaction.** Current AI techniques have been very successful in recognizing images, analyzing natural language text and speech, and playing games. The "AI superiority" over human champions in the difficult game of Go has been clearly demonstrated. However, most games have a relatively small set of precise rules, and take place in a well-defined, strictly limited setting, even though they may permit a huge number states or moves between states. "Teaching an AI system to play a game" is much easier than teaching a machine to "develop intelligence step by step from the learning by interaction with humans and the natural environment", where perceiving, acting, and interacting with humans are tasks that cannot be devised in isolation, but that deeply influence each other. There is a big step and a change in paradigm to move from games, images, and text to AI systems that can interact with humans and the world. Acting in the world and interacting with the environment influences perception, and vice versa. The integration between perception, action, and human interaction deserves novel research, and perception and action/interaction require tight integration. Models for planning, acting, and interacting depend on perception capabilities, and perception tasks should be informed by actions and interactions.
- **Reducing the barriers in designing, delivering, and maintaining AI systems.** AI systems are very challenging to build, deploy, monitor, and maintain. Most of the problems mentioned above - such as AI systems that are not safe, secure,

[1] The Integrative AI research challenge is beyond and not only a matter of software engineering, i.e., of putting together different components based on different AI representations and techniques. Notice that we do not mean that software engineering is a minor issue for the development of AI systems, especially from the point of view of democratization. An interesting question is what new fundamental research questions in software engineering are motivated by AI systems.

trustworthy, difficult to verify, and difficult to understand - are in part due to the intrinsic difficulty in building AI systems. Building "good" AI systems requires high expertise, but there is a need to "democratize" the use of AI in a way that AI can be developed by more and more people, including those that do not have the high level of expertise required today for building high-quality AI systems. This gives rise to an important research challenge: to devise techniques and tools that could help humans in designing, delivering and maintaining AI systems.

All these challenges are interconnected. For instance, the challenge of trustworthy AI has clear overlaps with the challenge of security, privacy, explainable, and integrative AI. The integration of perception, action, and interaction with humans is closely related to integrative AI, and to less data-intensive, trustworthy, and explainable AI. Only long-term, integrative, and interdisciplinary research can address the highly interconnected and interdisciplinary scientific challenges for AI for humanity. Unfortunately, current research evaluation methods and academic criteria tend to favor vertical, short-term, narrow, highly focused, community- and discipline-dependent research. It is the responsibility of all scientists in the academic world to foster a methodological shift that facilitates (or at least does not penalize) long-term, horizontal, interdisciplinary, and very ambitious research.

In the remainder of this chapter, we propose a more in-depth discussion of some of the research challenges mentioned previously, and some ideas of possible approaches to address these challenges and open a way towards AI for humanity.

In Sect. 2, we will deal with the requirement to understand how deep neural networks can debug, interpret, control and reason about their results. A possible approach is to measure the influence of the inputs and the relevance of the filters of a deep neural network, and their importance in providing results and possible decisions of an AI system. A major challenge here is to generate narratives (e.g., through text generation techniques) that can explain the network and can be easily understood by humans. Generating explanations and narratives can open up the possibility to build systems that interact with humans, such that humans are in control of the learning and reasoning process. This provides the basis for meaningful human control of AI systems.

Section 3 deals with the problem of building trustworthy AI systems. It provides some interesting examples that show how current AI systems for computer vision and natural language understanding based on deep learning are not trustworthy. The major issue is the "lack of context" of such techniques. The research challenge is to build robust AI systems that are resilient to errors, explainable, transparent, and safe by integrating learning techniques with background and common-sense knowledge, including knowledge about common facts, intuitive physics and intuitive psychology. An intermediate goal is to build "auditing AI programs", i.e., AI systems that are required to answer questions about some specific cases.

Finally, Sect. 4 addresses the problem of reducing the barriers in designing, implementing, delivering, and maintaining AI systems. This will help to address the pressing problem of the "talent bottleneck" in AI, i.e., the lack of highly skilled experts in building AI systems. The research challenge is that of "AutoAI" - Automated Artificial Intelligence, i.e., the automated design of AI systems, based on advanced statistics, optimization, and machine learning, a significant extension of the concept of

Automated Machine Learning (AutoML), since it considers methods and techniques across the entire spectrum of artificial intelligence.

2 Endowing Deep Neural Networks to Show and Explain Behavior and Decision Making[2]

Deep Neural Networks (DNN) have become ubiquitous. They have been successfully applied in a wide range of sectors including automotive, government, wearable, dairy, home appliances, security and surveillance, health, and many more, mainly for regression, classification, and anomaly detection problems. The neural network's capability of automatically discovering features to solve any task at hand makes them particularly easy to adapt to new problems and scenarios. Since the initial successes, the development of innovative deep learning approaches has accelerated rapidly. Deep learning approaches are becoming more complex, with new forms and architectures, learning more parameters and becoming increasingly better. Consequently, it is not easy to understand which architecture would best fit to which input and task. In order to be able to see through the forest of alternative architectures, network types, components and tools available to support individual tasks, Subsect. 2.1 introduces a TagTool, based on a faceted browsing approach which gives an orientation for users to select the right approach for a given problem.

Although many of these systems provide high accuracy, all those models reveal a black-box nature, i.e. they are lacking of transparency/intelligibility of their decisions. The applicability of DNN has also been compromised due to the lack of understanding the network decision processes well as the deficiency of explaining the decision [6]. This is specifically true for domains like business, finance, natural disaster management, health-care, self-driving cars, industry 4.0, and counter-terrorism where reasons for reaching a particular decision are equally important as the prediction itself. In this respect we may distinguish between two areas:

- Interpretability refers to the observation and representation of cause and effect within a system, without necessarily knowing why something happens
- Explainability, on the other hand, concerns the ability to explain the inner function of a system in human terms (e.g. by means of a given example).

In many cases full transparency may not be always possible or even required. In general, AI systems are designed to optimize behavior, i.e. to maximize accuracy with respect to a given goal. But they depend on the data, which might have a bias, e.g. when the data is not objective, complete, and balanced. At least, we should be able to understand the decision processes and identify the data responsible for the decision. One step towards the interpretability of DNN is addressed in Subsect. 2.2 Specifically, we describe a method to quantify the amount of information that CNNs extract from their input by investigating different best practice architectures for image classification.

[2] Sheraz Ahmed, Joachim Folz, and Sebastian Palacio contributed to this section.

However, if we are considering non-visual input, such as time series, it is even more difficult to decode and understand intermediate states in a deep network because of the automated feature engineering. In other words, features, which are extracted by these models, are hard to interpret and understand for humans, especially in cases with high-dimensional data. In Subsect. 2.3, we introduce a method that measures and visualizes the influence of the input data on the output or decision of the network. Furthermore, we extract and visualize the patterns which are present in most of the influential filters to finally generate a textual explanation easy to be understood by the user.

2.1 AI Landscape and Architecture Search

Advances in neural network accuracy have been driven by improvements to architectures [7–9] and training methods [10–12], but also availability of compute power [13, 14] where the number of parameters increased from millions to hundreds of millions and operations per sample exceed several billions within the last ten years. This increase in model size and complexity is even more evident in recent models for natural language processing. Comparing representative models from 2018 to 2020 shows a more than 10-fold increase in the number of parameters per year:

- 355M - BERT-Large [15]
- 1.5B - GPT-2 [16]
- 11B - T5–11 [17]
- 175B - GPT-3 [18].

GPT-3 shows accuracy on NLP tasks increases with the power law in terms of parameters. While performance is impressive, especially on unseen tasks without fine-tuning, this growth of model size is not sustainable as it outpaces the growth in available memory more than 100-fold. Eight GPUs with the largest currently available memory capacity (Nvidia Quadro RTX 8000 48 GB; assuming 2 bytes per parameter half-precision is used) are required to hold the parameter set, which makes just inferencing with this model challenging. Training takes hundreds of GPU-years and several million dollars of cloud budget to complete within a reasonable timeframe. Hence, more specialized architectures are still required for most use cases.

All GPT models are trained as next-word predictors: during training, the model output is compared word for word against large text corpora. No additional metadata is required and, most importantly, labor-intensive manual labelling is not necessary. Similarly, recent work on self-supervised learning on images, where training does also not require additional data, shows that several times as many parameters are required compared to supervised training [19–22]. These models are trained with so-called contrastive losses, where one or more model should output similar values for inputs that are known to be similar, and conversely dissimilar outputs for dissimilar inputs. For images, this is achieved by manipulating them in various ways, such as spatial and color transformation. Best results are currently achieved with very large batch sizes of several thousand images to ensure that sufficiently dissimilar images can be found. Hence, it can be argued that there is a tradeoff between dataset quality/cost and model

size/training effort with self-supervised learning, though the state-of-the-art at least with respect to required parameters is improving quickly.

Finding appropriate network architectures that strike a good balance for a given dataset and task has traditionally been a manual process of trial and error, involving highly skilled researchers and engineers adapting and/or extending existing known good examples. Given the number of meta-parameters (number and type of neurons, graph connectivity, transfer functions, etc.) an exhaustive search may never be feasible for what would be considered reasonably sized models and datasets at the time. While the idea of systematically creating the architecture and optimizing a neural network from scratch is not new [23], it has only recently been demonstrated that state-of-the-art accuracy on large-scale datasets with millions of samples can be achieved [24–26]. These first examples of neural architecture search employ methods borrowed from reinforcement learning. A generator produces model candidates that are trained on a target dataset and the achieved accuracy is transformed into a reward for the generator. This approach is computationally intensive, requiring hundreds of GPUs for relatively simple datasets. Further improvements, such as predicting the accuracy of a model [27, 28] or reusing parameters of identical blocks that had already been trained previously [29, 30], made it feasible on single GPUs. More recently, approaches foregoing reinforcement learning entirely have been proposed [31]. The problem is reformulated as a continuous search problem and can thus be optimized by standard gradient descent methods, providing further efficiency gains.

Making these rapid advances in deep learning techniques available to practitioners is another core issue. New forms of organizing and sharing knowledge are necessary to keep up with the rapidly growing body of work surround this topic, but it is also necessary to reduce the effort required to evaluate the efficacy of an existing model towards a new problem. Two systems illustrate possible.

The TagTool allows to create, interlink, and share several tag clouds at once. One such instance[3] of this tool is configured to collect and show six simultaneous views about deep learning models:

- Signal Types (image, text, time series, etc.)
- Network Types (CNN, RNN, GAN, etc.)
- Tasks (classification, detection, forecasting, etc.)
- Network Architectures (ResNet, ReNet, Siamese Networks, etc.)
- Components (convolution, activation, normalization, etc.)
- Links to external resources (papers, reference implementations, tools, etc.).

Tags can be linked within or between clouds and selecting an element shows what it is linked to. This can be used to, for example, find out what kind of networks are useful for image segmentation and what operations they a comprised of. There are also two advanced selection modes: AND and OR, where multiple tags can be selected to show either the intersection or the union of linked tags. For example, we can look for image segmentation networks that use an encoder-decoder type architecture. The

[3] tag cloud for our project DeFuseNN: https://defusenn.letstag.it/.

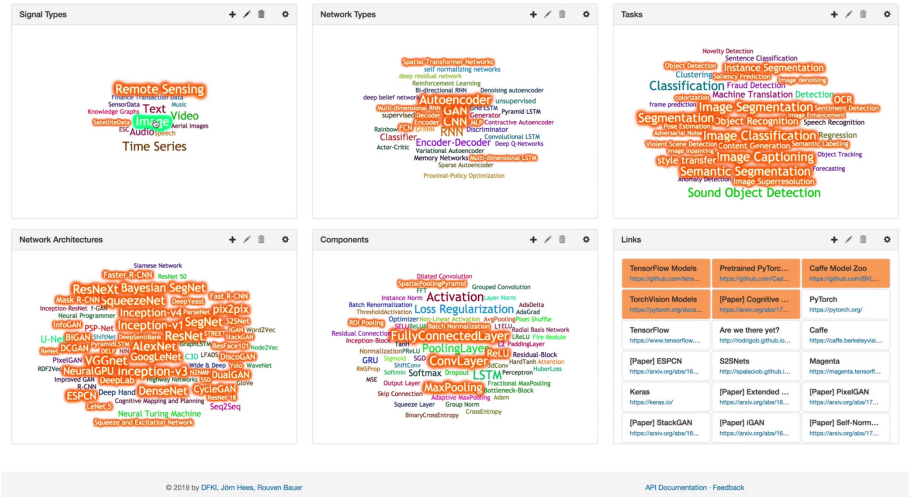

Fig. 1. The TagTool provides the opportunity to interact with the landscape of deep learning via faceted browsing and narrows down the solution space by combining different facets.

external links view provides resources for further reading, such as papers and reference implementations. Figure 1 shows an activation example of the TagTool.

The second tool, the Deep Learning Sandbox, complements the TagTool by making it easy and fast for exploring existing models and datasets and testing fully trained networks on new data. It allows to interact with a variety of models via a Web interface. Capabilities and performance metrics are displayed, allowing the user to make a pre-selection of interesting models that may be applicable to a new use case. Images, audio, and text input modalities are currently supported and can be uploaded to a Web interface for testing purposes. Each model specifies that it requires one or more samples of each modality to operate. The sandbox matches available inputs to applicable models and runs those selected by the user.

The approach is complemented by an intelligent scheduler, which reduces latency during inferencing process. Low volume requests may be handled faster by CPU-only operation, since initialization of a GPU-accelerated model can take longer than processing on the CPU. This implies that models are moved to a GPU if there are enough requests and one is available. Results are displayed next to each model and can be compared to each other (see Fig. 2).

2.2 Interpreting Deep Neural Networks

Current methods for interpreting modern ML pipelines have focused on a variety of narrow properties at play. Said properties can be broadly categorized as model-based explanations and data-based explanations. For the former, a common strategy in the image domain consists of reverse-engineering a neural network in order to find an input which elicits a high response from a particular neuron or layer [32, 33]. Having an

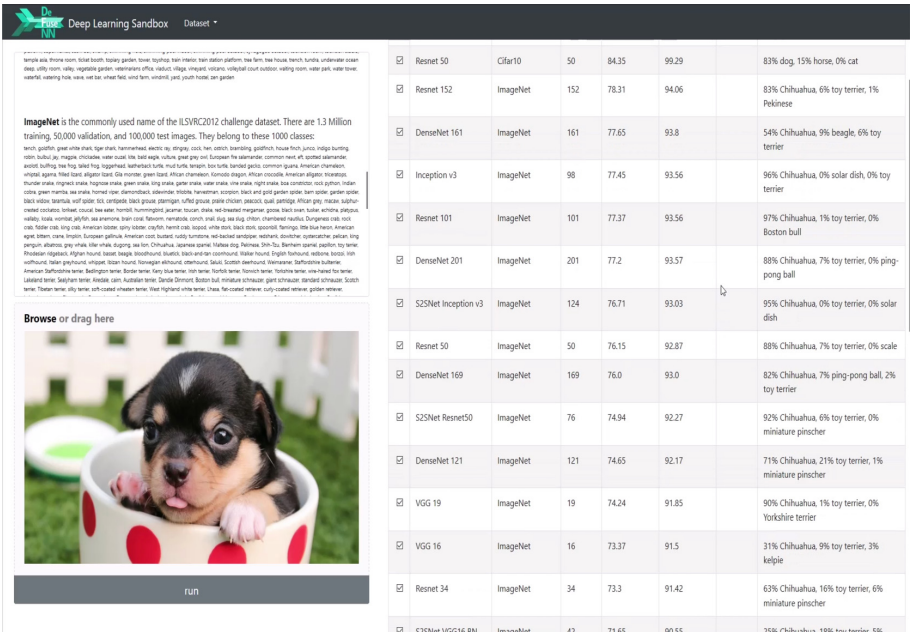

Fig. 2. The DL Sandbox offers a right set of pretrained models, which may be individually applied to uploaded data samples or may be compared respecting their accuracy.

image pattern expressed in the input domain makes said pattern more amenable for humans to infer what the neural network is looking for or reacting to.

An orthogonal approach consists on analyzing valid, existing samples individually and recording high activation patterns as they traverse the neural network. These activations can be traced back to the original input and visualized as relevance scores for that particular sample [34–36].

However, patterns affecting the entire model (not just a single layer or neuron) remain undetected under these interpretability strategies, since they influence all input samples equally.

In order to unveil these kinds of global patterns, we wish to capture properties of the input space that are relevant not only to individual samples but also to the entire dataset. Once these properties are conveyed, the most relevant ones can be selected for further analysis. Parametrization of the input space can be done via Autoencoders [37] where a neural network learns a parametrized approximation of the respecting identity function.

In order to achieve a low reconstruction error for the input space of arbitrary natural images (and therefore, a better approximation of the input distribution), a large autoencoder known as SegNet [38] is used. Preventing overfitting for such a large network usually requires the use of extensive and careful regularization techniques. Alternatively, the unsupervised optimization objective for autoencoders allows more relaxed constraints at the expense of using a larger training set. The YFCC100m [39] is a weakly supervised image dataset that provides the scale needed to train the SegNet

autoencoder with low reconstruction error, requiring only one pass (one epoch) before having fully converged.

Once the input space has been parametrized by the differentiable autoencoder, a pre-trained image classifier is evaluated with the reconstructions of the autoencoder, i.e., the parametrized version of the input space. This ensemble yields a composite function, where the identity Function is used as input for the classifier.

Intuitively, it is expected that the pre-trained classifier is selectively processing information contained in each input sample (e.g., ignoring the background and identifying salient parts of the image). More generally, any ML model will selectively use information in the input, depending on its task. To unveil exactly what information is being used by the classifier, one adapts the autoencoder described above. Thanks to the parametrized (and end-to-end differentiable) version of the input space, a further optimization of the autoencoder allows the reconstructed samples to match the information that the pre-trained classifier expects. Concretely, decoding layers of the autoencoder are fine-tuned with gradients from the classifier according to its classification objective. The resulting fine-tuned autoencoder is referred to as a structure-to-signal network (S2SNet) [63]. Once an S2SNet has been obtained, we can verify that a distinct artifact is introduced when reconstructing original samples with it. This artifact *is constant for all samples in the dataset* and indicates that information conveyed by values where the artifact is now present, do not carry information that is useful for the classifier. To quantify the constancy of said artifacts, the normalized mutual information (nMI) [40] is computed between the original samples and corresponding S2SNet reconstructions. This is referred to as the intra-class nMI and measures the information that has been dropped w.r.t. the original input. Furthermore, the nMI is computed between S2SNet reconstructions of random samples, with high values indicating the degree of constancy that comes from the reconstruction process.

Through these two nMI metrics it is possible to establish the amount of information used (i.e., "useful information") by high-performance image classifiers like Alexnet [41], Resnet50 [8], VGG16 [42] and Inception v3 [43]. Based on this notion of "useful information" we see (cf. Fig. 3) how Alexnet takes in the least amount of information, followed by Resnet50, Inception v3 and VGG. The constancy of reconstruction artifacts (according to nMI measurements) does not directly correlate with accuracy, network depth, normalization or pooling operators, and has links to the informal notion of "model capacity": a term often used in the literature to convey the ability of a neural network to approximate a richer set of functions.

For instance, heatmap reconstructions based on Deep Taylor Decomposition [44] exhibit higher resolution when computed on Inception networks compared to results based on Alexnet. From the standpoint of useful information this behavior is expected, as the latter model produces more tenuous reconstruction artifacts, and therefore, more useful information from the input is projected back into the heatmap. Similarly, the high amount of information used by VGG justifies its use for building convolutional autoencoders or networks for image segmentation; a common practice seen, for example, in the architecture of SegNet itself.

One additional property of image reconstructions using S2SNets is that the constant artifacts (i.e., reconstructed pixels with a constant value, regardless of the values in the original samples) represent a projection of the original input into a lower dimensional

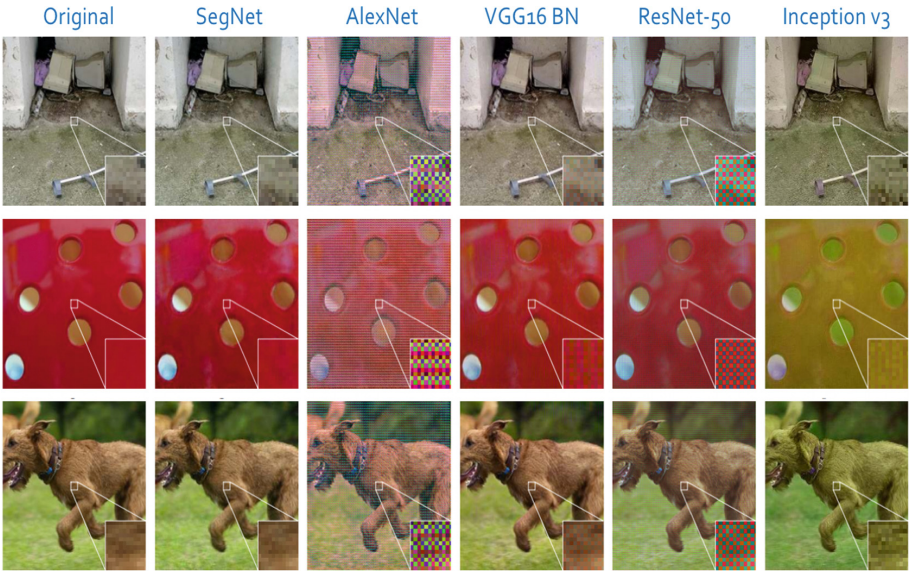

Fig. 3. Randomly selected image samples and results on different classifier architectures.

space. This is especially valuable when the model has to cope with adversarial attacks [45]: maliciously perturbed samples intended to cause an ML model to fail. Adversarial perturbations have shown to be less effective when created and evaluated on an S2SNet compared to their performance against a classifier alone [46]. A comprehensive evaluation for different gradient-based adversarial attacks like FGSM [47], BIM [48] and CW [49] provides compelling empirical evidence that S2SNets mitigate the malicious effects of these attacks [64]. This robustness is achieved without the need for additional assumptions regarding the attacks and preserves the accuracy of the original classifier when clean images are evaluated; a compromise that is often made by some alternative defense mechanisms.

In short, global interpretability measures can be extracted by parametrizing the input space and enhancing its properties with respect to a given task (e.g., image classification). Instead of focusing on individual samples or a particular module of an ML model, S2SNets make use of fine-tuned autoencoders to filter the amount of information that a classifier effectively uses from the input. Characterizing the amount of "useful information" elucidates on multiple reports of the otherwise informal notion of "model capacity" often found in the literature and serves as a robust alternative to mitigate the effects of adversarial attacks.

2.3 Explainable AI

In the domains where human lives are directly or indirectly linked to a machine's decision or high-stakes decisions are based on them, the trustworthiness of the decision-making system is more important than accuracy. This trustworthiness can be achieved by enabling a system to answer the *"HOW"* and the *"WHY"* of a decision.

The *HOW* part can be addressed when a system is capable of showing how it has taken a particular decision. In this process, the system must highlight the major observables to show how they are behaving and changing. The *WHY* part can be addressed when a system provides an explanation of a decision. It is important to provide reasons for a particular decision taken by a system. The attached facts to an explanation make an explanation more transparent which eventually makes the whole system trustworthy.

There have been significant attempts to uncover the black-box nature of deep learning-based models [33, 50–54], where visualization of the model has been the most common strategy. Almost all of the proposed visualization systems are image-centric where visualizing the image is directly interpretable for humans (natural association to similar looking objects like eyes, faces, dogs, cars etc.). These visualizations help humans understand the thinking process of an Artificial Neural Network (ANN). Most of these visualization and interpretability ideas are equally applicable to time-series, but the unintuitive nature of the time-series data makes it difficult to directly transfer these ideas to aid human understanding. To demystify a deep model for time-series analysis, Siddiqui et al. [55] proposed a framework – *TSViz*. This framework introduces an influence tracing algorithm to compute the input saliency map, which enables an understanding of the regions of the input that were responsible for a particular prediction. In addition to that, an approach to compute the filter's influence using the proposed influence tracing algorithm is also introduced in this framework. Filter importance is computed based on its influence on the final output. This information provides an idea to the user regarding the filters of the network that were important for a particular prediction. These visualizations enable a system to answer the *HOW* part.

Though a picture is worth thousand words, still it provides an overview, not a detailed explanation. To understand the details, it is necessary to have a logical description of the picture. It has been well established in the prior literature that an explanation of the decision made by a DNN is essential to fully exploit the potential of these networks [56, 57]. With the rise in demand for these deep models, there is an increasing need to have the ability to explain their decisions. For instance, big industrial machines cannot be powered down just because a DNN predicted a high anomaly score. It is important to understand the reason for reaching a particular decision, i.e. why the DNN computed such an anomaly score. Adequate reasoning of the decision taken increases the user's confidence in the system. To address this *WHY* part, *TSXplain* is introduced by Munir et al. [58]. This framework is inspired by the human psychology of logical reasoning for a particular decision. It contributes to the WHY part by generating natural language explanations of the decisions made by a DNN. Powerful statistical features are aligned with the most influential data points to generate textual explanations as they are exemplarily shown in Fig. 4. The two-level explanation provides ample description of the decision made by the network to aid an expert as well as a novice user alike.

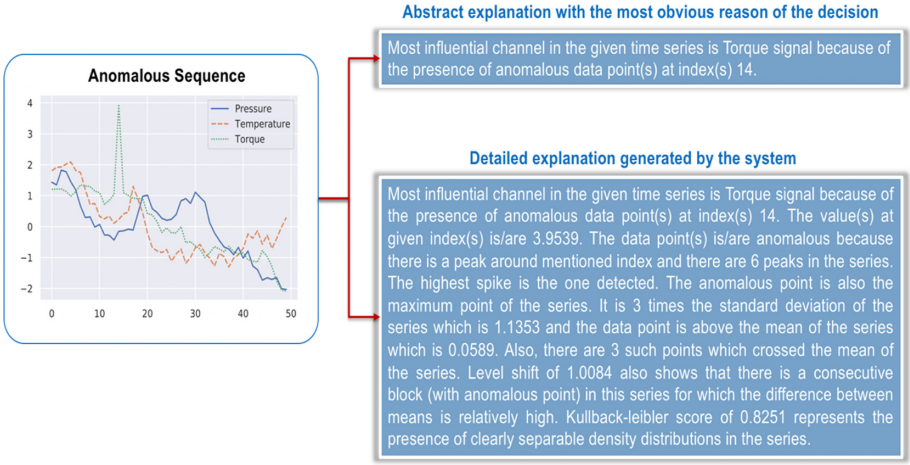

Fig. 4. Two levels of explanation are generated by the system depending on what an end-user-desires: abstract or detailed explanation.

3 The Challenge of Trustworthy AI

The recent excitement about GPT-3, the latest autocomplete language tool from OpenAI, is a stark reminder of the need for trustworthy AI. It's been heralded as "astonishingly powerful". GPT-3 is, indeed, surprisingly powerful and fluent but it is also utterly untrustworthy. In one experiment run by Summers-Stay, Marcus, and Davis, GPT-3 presented this [59]: (Prompt) You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to… (GPT-3 generated text) remove the door. You have a table saw, so you cut the door in half and remove the top half. One can imagine how quickly the comical text turns to terrifying if such an error appeared in a legal document or medical chart.

In fact, if you type "trustworthy AI" into Google, you are met with over 14 million results. Articles, books, blogs, and even entire websites are dedicated to defining trustworthy AI and offering solutions and frameworks for building it.

AI systems today are entrusted with making decisions that deeply impact our lives, such as who gets a mortgage or determining medical diagnoses. Yet, how and why these decisions are made remains a mystery, even to the creators of the technology. So, it's no surprise that when these systems make egregious errors, trust in them quickly erodes.

When courts across the country began using a risk assessment tool to determine who should receive parole, ProPublica uncovered a disturbing trend: black defendants received higher risk scores than white defendants with similar profiles. Because the tool's creator would not divulge information about the proprietary algorithm, we may never know why Gregory Lugo, who crashed his car into another one while drunk, was rated 1 (low risk) despite the fact that it was at least his fourth DUI, but Mallory Williams was rated 6 after one DUI.

Since algorithms "learn" based on the data they are fed, any bias in that data can become amplified. For example, when Amazon created an automated tool to review resumes, they soon realized the hiring algorithm taught itself something unexpected: it was excluding women from technical jobs because their resumes included words like "women's" and downgrading graduates of certain all-women's colleges. The tool was ultimately abandoned, but it shined a spotlight on the perils of using automated systems for important decisions in the absence of transparency.

In response to challenges, like these, there has been a push for explainable AI, ensuring decisions can be understood by a human. For instance, your loan application would be accompanied by specific reasons for a rejection. On the surface, this approach sounds like a no-brainer, but in practice it is a formidable mandate for three key reasons [60]. The most common retort is that explainable AI can reveal proprietary data and trade secrets. A bigger challenge is the inherent difficulty of explaining the behavior of nonlinear neural network models trained over massive data sets. It is virtually impossible to explain decisions made in this way- not in a linear, logical, feeling, human way, but conclusions derived from a weighted, nonlinear combination of thousands of inputs, each contributing a microscopic percentage point toward the overall judgement. For example, if you've watched Netflix, you've likely noticed the "Because you watched" category which recommends other shows to watch based on your viewing history. These seemingly simple recommendations are actually built on complex algorithms factoring in multiple inputs. While Netflix viewing recommendations are a harmless oversimplification of the process, such generalizations can prove dangerous in more high-stakes settings.

Finally, AI models are vulnerable to a common phenomenon known as Simpson's paradox, which occurs when trends in groups of data reverse as that data is combined. Perhaps the most well-known example of Simpson's paradox involves graduate school admission data from UC Berkeley from 1973. When the data was viewed in aggregate, it appeared that men were admitted at a significantly higher rate than women. When that same data was viewed differently, focusing on individual departments, it showed a small but statistically significant bias in favor of women. This example is not unique, and additional data is not the solution.

Given these challenges with explainable AI, perhaps a better, more transparent approach is auditable AI: AI systems that are queried externally with hypothetical cases. These hypothetical cases can be real or imagined and allow for instant, automated monitoring. This is an especially useful way to screen for bias. For example, loan applications can be run through models that change gender or neighborhood to see if approval changes with each tweak.

Auditable AI has several advantages. Primarily, auditable AI is investigated by a neutral third-party immune from any bias or control of the algorithm's creator. It also eliminates the concern that explaining AI systems exposes trade secrets and proprietary data since the audit would not reveal this. Audited AI is a welcome counterbalance to explainable AI; auditing can help investigate, endorse, or even invalidate AI explanations. For example, if Pandora recommends Elton John because I listened to Billy Joel, will it also recommend other classic rock musicians? Does it recommend Elton John to everyone who's listened to Billy Joel?

Auditable AI is already gaining traction. The Best Student Paper Award at the 2019 AAAI AI Ethics and Society Conference was focused on an audit of software like Amazon's Rekognition tool which was nearly twice as likely to misidentify people of color. While still short of perfect, the audit paved the way for a reduction in error rates and deeper awareness of flaws in these systems which are becoming more broadly used. The private sector is also moving toward creating and using these key capabilities. WhyLabs is a startup out of the Allen Institute for AI Incubator creating products for tracking and auditing model performance post launch to verify they are performing as expected. Yet, auditable AI is not a bullet-proof solution. There are, indeed, high-stakes decisions, like medical diagnostics, that warrant an accurate and understandable explanation, not just an audit. While these use cases and paths to explanation undergo the essential research they should, auditable AI can increase transparency and combat bias.

Ultimately, to make AI trustworthy, we must create robust, intelligible AI systems where it is clear what factors caused the system's action and users can predict the system's behavior with input changes. The degree to which an explanation is available or provided with AI decisions will vary based on use case. Psychologists have studied explanation for decades, and those learnings can shape how we build interactive systems to ensure a data scientist or developer debugging a system and a loan seeker can glean the different details important to them from the same system [61].

As the field of AI rapidly develops, oversight must also adapt. In the future, we can envision a comprehensive auditing ecosystem providing deeper insights into AI and "AI guardians" that address challenges and respond to the potential risks associated with increasingly autonomous AI systems. These systems are not meant to be overly strict or rigid, but to ensure AI systems remain aligned with the guidelines of their programmers. AI systems are learning systems, and like us, learning humans, latitude for trial and error is required. However, clear boundaries and understanding of risk so AI systems adhere to laws and ethical norms are crucial.

4 Addressing the AI Talent Bottleneck by Automating Artificial Intelligence

Roughly since 2011, there has been a marked increase in research activities, applications and public interest in artificial intelligence, accompanied by ample speculation about future capabilities and uses of AI technology, as well as of the benefits and risks they may bring. This development is triggered, to a large extent, by impressive progress in a specific area of AI, namely that of machine learning, and focused around the concept of learning with deep neural networks. It is important, however, to realise that this is not the first wave of enthusiasm for AI, and that the reasons underlying this latest surge in interest run far deeper than deep learning. In the following, we will outline these reasons and their consequences, discuss a serious threat associated with the current and ongoing boom in AI applications, and explain how this threat can be mitigated by judiciously automating the design, deployment and maintenance of future AI systems, following an approach dubbed *AutoAI*, with an emphasis on the

technological challenges arising in this new and exciting area. We will conclude with some thoughts on the future of AI technologies and their applications.

4.1 Causes and Consequences of the Current Boom in AI

Interest in AI has peaked before and then waned. Common wisdom has it that the main factor causing past downturns in AI was the inevitable disillusionment following wildly exaggerated expectations [65]. This, of course, suggests that the current boom in AI, or "AI summer", may be similarly destined to be followed by a bust, or "AI winter", a marked decrease in public interest (see, e.g., [66]). While a detailed discussion of the history of AI, and specifically, the causes of previous "AI summers" and "winters", is beyond the scope of this chapter, it is illuminating to discuss the causes of the latest, marked increase in interest. In our view, these include advances in computing hardware, advances in AI techniques and algorithms, a dramatic increase in the availability of useful data, and a high degree of "AI readiness" across industry and society.

The first of these factors, impressive and sustained progress in hardware, is well known, so we refrain from covering it in detail; it is instructive, however, to note that computations that would have taken 10 h in 1991 could be performed in less than 3 min by 2007, thanks alone to sustained progress in computer hardware (see, e.g., [67]). What is less widely known is the fact that advances in algorithms (i.e., in software) are even more dramatic - especially when it comes to solving the kinds of problems that fall into the area of AI, problems that when solved by humans require significant intellectual effort, often in combination with substantial amounts of experience.

A well-known example comes from the area of solving an optimisation problem known as mixed-integer linear programming (MIP), which has a broad range of real-world applications in industry and academia (see, e.g., [68]). Progress in MIP algorithms in the widely used commercial MIP solver CPLEX was shown to have achieved a more than 28000-fold speed-up between 1991 and 2007 when solving the same benchmark instances on the same hardware, while the speed-up due to improvements in hardware over the same 16-year period corresponds to a factor of 218 [69]. By combining the hardware- and software-related speedups in this example, an astonishing 6.2-million-fold speed-up was achieved over a period of only 16 years. A recent study on hardware- *vs* software-related improvements in solving the propositional satisfiability problem (SAT) - one of the most intensely studied AI problems, which plays a key role in verifying the correctness of computer hard- and software - yielded qualitatively similar results, indicating the dramatic effects of algorithmic improvements over a period of about 20 years [62].

The third factor, an increase in the availability of useful data, is certainly of key importance in the area of supervised machine learning, where the amount and quality of training data is known to play a crucial role for the performance obtained from state-of-the-art techniques. There are several reasons for the increased availability of data; firstly, more data is being produced, as a result of advances in the design of sensors and their increasingly broad deployment, but also as a direct consequence of the transition to digital media and storage formats, including the global rise of social media; secondly, an enormous amount of data has been collected for several decades now,

facilitated by cheap storage and easy transmission of large volumes of data; and finally, much of this data is now broadly and efficiently available via the internet. Interestingly, the dramatically increased availability of data benefits many areas of AI beyond machine learning, since the development of new algorithms often depends on performance assessments on large sets of benchmark instances (this is the case, for example, in the previously mentioned areas of MIP and SAT solving).

The fourth factor, "AI readiness", is a consequence of the broad use of computation across all sectors of industry and many aspects of our daily lives. Modern production environments, aircraft, ships, medical equipment and administrative processes (to name but a few examples) are now run by algorithms and operate routinely on large amounts of digital data. As a result, in many cases, a transition to AI techniques requires merely a change in software rather than dramatically more costly and disruptive changes in specialised hardware. Furthermore, in cases where AI techniques require substantially higher computational resources than currently available, upgrades or virtualisation of general-purpose hardware components are far easier and cheaper to achieve than the earlier transition to algorithmic data processing and control. This means that there is now an increasingly low barrier to the first-time adoption of AI techniques, and an even lower barrier to subsequent transitions to more advanced techniques.

While most AI experts would agree that these four factors played an important role in the large increase in broad interest in AI, there are two further, perhaps more contentious factors at play. The first of these is directly related to the fact that many regard the present AI boom as mostly caused by fundamental advances in the area of multi-layer neural networks. While advances in neural networks - enabled by readily available, high-performance hardware (notably, GPUs), innovation in algorithms (both in terms of the neural network models themselves, as well as in the algorithms for training them) and large amounts of training data - have doubtlessly played a key role, the impact of these advances has been amplified by the fact that for at least two decades prior to 2011, neural networks were marginalised, and on many occasions outright dismissed, by large parts of the mainstream AI community. This led to a situation where relatively few researchers seriously worked on and with neural networks, a set of versatile AI techniques with a history dating back to the 1940s. As a result, progress in this area was likely artificially slowed, but poised to accelerate rapidly as soon as it became a major focus of attention. This brings us to the second additional factor at play, which is a combination of the inherent interest, especially among young researchers and the broader public, in biologically inspired techniques, such as neural networks, which are far more relatable than other, more abstract AI approaches, and the enormous publicity generated by companies that chose to invest into this "new wave of AI".

With this analysis of causes for the present AI boom in mind, we will now argue that this boom is different from previous peaks in interest in AI, in that it will likely have far broader and more lasting consequences. The last two factors - relatability and marketing of a specific set of techniques, and the penned-up impact and innovation potential of these techniques - fail to provide a compelling basis for this argument, and in fact could be seen as evidence to the contrary. The combination of the remaining four factors - advances in hardware, AI algorithms (broadly defined), markedly increased availability of data and general AI readiness in real-world application

contexts - does, however, suggest that AI techniques have now reached a critical level of usefulness at which they can and do provide substantial value to industry and society, at relatively moderate cost and effort - in other words, in a rapidly increasing range of applications, they enable new and valuable products, services and experiences.

This now rapidly occurring, broad and accelerating valorisation of AI technology is what distinguishes the current boom from previous waves of interest in AI. The use of AI brings tangible competitive advantage in many sectors of industry; this advantage increases further with the power of the AI techniques that are being deployed successfully, which creates a powerful incentive for industry and society to invest into research and innovation in AI. It is for this reason that, while the enthusiasm for particular AI techniques or approaches, such as deep neural networks, will continue to wax and wane, the overall high level of interest in AI is here to stay. Because of their broad applicability, across all sectors of industry and society, and in light of their emergence as key enablers of scientific and technological progress, AI systems and techniques are poised to fundamentally transform the way we live and work (see, e.g., [70]).

4.2 The Biggest Risk Associated with AI

In much of the main-stream fictional depiction of AI and some of the contemporary debate on the topic, the focus is firmly on broad-spectrum, super-human AI turning antagonistic and causing harm - a scenario we may dub “strong AI going bad”. While, in our opinion, this is a concern that deserves being taken seriously (for reasons beyond the scope of this section), it is by far not the most pressing risk associated with the development and use of AI technology. The main reason for this is that we are still quite far from being able to realise broad-spectrum, human-level or super-human AI.

Another commonly emphasised risk is that of a massive loss of jobs due to AI systems outright replacing human workers (see Chapter 4). This is doubtlessly a more pressing risk, since the increase in automation afforded by broad use of AI brings a large potential for eliminating, or at least much reducing, the need for human labour across an increasingly broad spectrum of occupations. However, it is possible that new kinds of occupations will in part make up for these effects, and that mechanisms for the fair distributions of the benefits derived from this kind of automation can further mitigate the inequities that may otherwise be caused by broad use of AI. Still, job loss caused by sharply accelerated, AI-enabled automation is a serious issue that needs to be addressed in the near future.

However, by far the biggest risk associated with the pervasive use of AI is of a very different nature, and requires no assumptions on further progress in AI technology: the risk of well-intentioned, yet incompetent use of weak AI - of the kind of AI systems and techniques available right now. This risk necessarily arises from the combination of three facts: one, that AI technology is complex and difficult to develop, deploy and maintain; two, that the highly specialised expertise required for effectively and responsibly developing, deploying and maintaining AI systems and techniques is relatively rare and difficult to acquire; and three, that the demand for this expertise far exceeds the supply. The last of these is what we call the talent bottleneck, since not only the number of competent AI developers and users is low compared to the demand

for them, but also the number of those that with a moderate amount of additional training can reach the required level of expertise. The high demand for AI expertise is directly caused by the usefulness and rapidly increasingly scope for successful valorisation of the technology and can be expected to further increase, quite rapidly, for the foreseeable future.

The consequences of this bottleneck in talent and expertise are obvious: Increasingly, AI systems will be developed, deployed and maintained by people who are lacking the proper knowledge and experience. As a result, these systems and their use will be prone to malfunction and unintended side effects; they will cause problems which will often be difficult to detect before significant damage has occurred. This is of particular concern in situations that involve the use of complex machine learning techniques and large amounts of data in a black-box fashion, as is the case in most deep learning approaches. The degree to which even moderately complex software (and hardware) is difficult to design in a correct and robust fashion is evident from well-known examples of costly, and sometimes deadly, malfunctions, such as the MCAS system that caused the loss of two Boing 737 Max aircraft in late 2018 and early 2019 [71], and this difficulty is much more pronounced when dealing with even more complex AI systems (see, e.g., [72]). To make matters worse, the highly undesirable consequences of well-intentioned, underqualified use of weak AI will be particularly pronounced in areas where it is difficult or impossible to successfully compete for properly trained AI experts - notably, in the public sector and in non-profit organisations.

The most obvious way to address this talent bottleneck is to step up AI education. Currently, competent development and deployment of AI systems requires post-graduate, and in many cases PhD-level training specifically in AI, typically on the basis of a bachelor-level degree in computer science. There are much-needed efforts underway to expand these programmes, and to start suitably chosen components of AI education earlier, but the available and interested talent still forms a serious bottleneck. To address this, it is crucial to further develop the effectively accessible talent pool - first and foremost by taking measures to increase the participation of women, and secondly by tapping further into the enormous potential present in developing economies.

Clearly, stepping up education, in terms of improved and broadened educational offerings, an earlier start, and the development of AI-related professional occupations (e.g., related to the deployment, monitoring and maintenance of AI systems) is crucial in terms of addressing the talent bottleneck, but it will not close the gap between supply and demand of AI expertise, since current AI technology is simply too difficult to develop and use responsibly. Therefore, it is of crucial importance to lower the level of expertise required for effectively and responsibly working on and with AI systems, which brings us to the technical challenges that are at the core of this section.

4.3 Automating Artificial Intelligence

Within the last decade, there have been two revolutions in machine learning (ML), one of the most prominent areas of AI even prior to these developments. One of these, the (re-)emergence of neural networks as a dominant paradigm, has played out with great

fanfare and substantial resonance far beyond the field of AI. The other has been quiet in comparison, and largely hidden from the eyes of the broader public, but is nonetheless at least as relevant: the birth and rise of the concept of *automated machine learning (AutoML)*.

AutoML is an approach that aims to automate a set of task associated with making effective use of ML methods and tools, including the choice of ML techniques and the settings of the hyperparameters that determine their performance in particular use cases (see, e.g., [73]). The concept arose, under that name, around 2013, and has rapidly gained traction in the machine learning community and beyond. From the very beginning, work in the area of AutoML has sought to not only help ML experts to obtain better performance from existing ML algorithms, but also to lower the threshold for the effective use of a broad range of ML techniques [74].

Interestingly, while programming can be understood as the principled automation of well-structured tasks, machine learning fundamentally concerns the automation of programming for tasks such as classification, regression and interaction with complex environments, and hence corresponds to the automation of automation. This explains in part why the rise of broadly applicable and successful machine learning techniques can be legitimately seen as a technological revolution. Under this view, AutoML takes automation to the next level, enabling an even higher degree of substitution of broadly and readily available computation for scarce and expensive human expert knowledge.

AutoAI is based on the same idea, applied to *all* of AI rather than just machine learning. This is extremely relevant, since firstly, contrary to widely held beliefs, there are other areas of AI that are remarkably successful in terms of real-world impact, including automated reasoning (which forms the basis for the design of all modern hardware, and is increasingly used for ensuring software correctness), optimisation (with a broad range of applications across industry and academia), and multi-agent systems (which play an increasingly crucial role in the automation of decision making in situations involving multiple actors or agents with possibly conflicting goals and preferences). Most AI experts are convinced that next-generation AI systems need to combine learning, reasoning and other techniques, in order to achieve robust performance and effective interaction with human users and stakeholders.

Concretely, AutoAI aims to automate critical aspects of the development, deployment and responsible operation of AI systems. This includes task such as selection of AI techniques and algorithms that are suitable in a given use context, optimisation of the performance of these algorithms for the data characteristic of that use context, and monitoring of the behaviour of AI systems after deployment, with the goal of detecting, and clearly signalling, when the operational conditions deviate far enough from those considered at the time of development and deployment to cause problems.

This gives rise to several technical challenges. Firstly, fully or partially automated selection of AI techniques and algorithms for a given use case is a daunting task, considering that many real-world problems do not easily map to a small set of well-know AI problems, and require non-trivial combinations of techniques to be tackled effectively. Furthermore, where mappings to existing problems (such as MIP) exist, these are often not unique, but rather admit a potentially very large range of encodings, the choice of which can have dramatic impact on the performance of standard

algorithms for those problems. Secondly, while automatic performance optimisation techniques exist (see, e.g., [75]), these are far too limited to be applied broadly to AI systems with many design choices and parameters that can potentially impact performance. In particular, with very few exceptions (see, e.g., [76]), these general-purpose automated algorithm configurators are restricted to optimising a single performance objective, such as solution quality or running time, while in realistic scenarios, there is often a need to find good trade-offs between multiple, competing performance criteria, such as solution quality and resource consumption. Thirdly, broadly applicable techniques for monitoring the operation and performance of AI systems in relation to changes in the environment they operate in, and for signalling when these systems get "out of their depth", are largely unexplored; we refer to the automated combination of AI systems with such monitoring capabilities as *self-monitoring AI.*

At the same time, recent progress in AI techniques for algorithm selection, configuration and performance modelling provide a solid basis for work towards meeting these challenges, and hence for effective AutoAI methods and tools. It is important to realise that the goal of AutoAI, as we see it, is not full automation of the previously mentioned tasks, but rather effective support for the humans that tackle them, at various levels of expertise, ultimately substituting substantial amounts of costly and scarce human expertise with large amounts of readily available computation. At the same time, by automating key aspects of building, deploying and maintaining AI systems, AutoAI makes explicit the assumptions, practices and insights brought by human experts to these tasks, and thus not only renders these accessible to a broader range of developers and users, but also facilitates their critical assessment and improvement. Finally, by making it substantially easier to realise the performance potential inherent in AI algorithms and components in a broader range of specific application situations, AutoAI can be expected to make it possible to decrease the complexity of the systems and methods that need to be brought to bear to achieve desirable performance in many use cases.

4.4 The Way of the Future

The idea of machine intelligence has fascinated humankind for centuries; it is inextricably linked with the development of computing technology that, since the 1980s, has become the main driver for technological progress and innovation. The advanced computational methods developed in AI represent the next major step on this path. While interest in AI has shown several distinct peaks and troughs since the inception of the field in the 1950s, as we have argued in Sect. 4.1, there are good reasons to believe that the latest boom is of a different nature, as AI technologies have begun to rapidly change the way we design and use computation across all sectors of industry and society, and will thus bring about a lasting transformation in the way we live and work.

There is a rather high sensitivity to the risks associated with the development and use of AI technologies; unfortunately, as we have explained in Sect. 4.2, the most serious risk in the near and medium term is rarely recognised: the well-intentioned, yet incompetent use of weak AI systems, such as the ones we presently build and deploy, that is inevitably going to occur increasingly and on a large scale, especially in the public sector and non-profit organisations, as a result of the dearth of properly trained

and qualified AI experts, in combination with the inherent complexity of current AI technology. This gives rise to the formidable challenge of enabling the effective and responsible design, deployment and maintenance of AI systems at significantly lower levels of expertise.

The technical direction for addressing this challenge we have outlined in Sect. 4.3 known as *AutoAI (automated AI)*, is based on the idea of harnessing AI techniques for the effective and responsible design, deployment and maintenance of AI systems. We have outlined several challenges for AI technology in this area, including the concept of self-monitoring AI, which permits the automated construction of AI systems that can detect and signal when they are no longer operating in a safe and effective fashion.

AutoAI can bring a broad range of benefits beyond alleviating the talent bottleneck; these include markedly increased performance and robustness of AI systems; substantial savings in the energy required for building and operating AI systems along with the associated costs of these systems; broader effective applicability and easier customisation of AI systems; reduced requirements for data; and broader access to AI technology (e.g., in the context of citizen science). AutoAI thus aims to facilitate work on and with AI systems across many levels of experience and expertise, from highly skilled specialists to technically adept laypersons.

Naturally, the concept of AutoAI brings its own challenges, which need to be addressed by research on this topic as well as in the way AutoAI technologies and tools are used. This includes the potential for creating even more complex AI systems that perform better, but end up being more opaque, less reliable and more difficult to use responsibly, as well as the potential acceleration of research and developments aimed at artificial general intelligence.

We are deeply convinced that AutoAI is the next logical step in the development of AI technology, with the potential to fundamentally transform the way we design, deploy and maintain AI systems. Of course, as is the case with present-day AI techniques and many other powerful technologies, AutoAI can be used in ways we find problematic, troubling or outright objectionable - in particular, for constructing AI systems whose use undermines human rights, freedom or dignity, or the fair and responsible use of critical resources. In our view, such objectionable uses include the development of AI that aims to replace, rather than augment, human intelligence. Therefore, it is of the utmost importance to complement work on the technological challenges associated with AutoAI with work on mechanisms, including regulation, that ensure responsible use. This requires skills beyond those required for the technical work on AutoAI (and AI in general), as well as political determination.

The way we develop and use AI will doubtlessly shape our future. The transformative power of AI technology can be readily glimpsed from recent applications, and will become more evident in the near future. AutoAI will further amplify this power, but developed and used judiciously, it will also allow us to better harness it not only for the benefit of relatively narrow segments of society, but also for the collective welfare of humankind, while avoiding many of the risks associated with the careless development and use of AI technology. It will thus play an important role in paving our way into the future - a future that much depends on our values, choices and determination.

References

1. Muscettola, N., Nayak, P.P., Pell, B., Williams, B.C.: Remote agent: to boldly go where no AI system has gone before. Artif. Intell. **103**(1–2), 5–47 (1998)
2. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
3. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**, 2402–2410 (2016)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
5. Gori, M., Campiani, G., Rossi, A., Setacci, C.: The web of clinical data. J. Cardiovasc. Surg. **23**, 717–718 (2014)
6. Melis, D.A., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems, pp. 7775–7784 (2018)
7. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
8. He, K., et al.: Deep residual learning for image recognition. CoRR abs/1512.03385, pp. 646–661 (2015)
9. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7), 2121–2159 (2011)
11. Alex, G.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678 (2016)
14. Bianco, S., et al.: Benchmark analysis of representative deep neural network architectures. IEEE Access **6**, 64270–64277 (2018)
15. Devlin, J., et al.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Radford, A., et al.: Improving language understanding by generative pre-training, vol. 12 (2018)
17. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
18. Brown, T.B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
19. He, K., et al.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
20. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
21. Chen, T., et al.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
22. Caron, M., et al.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
23. Tenorio, M.F., Wei-Tsih, L.: Self organizing neural networks for the identification problem. In: Advances in Neural Information Processing Systems (1989)

24. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)
25. Baker, B., et al.: Designing neural network architectures using reinforcement learning. arXiv preprint arXiv:1611.02167 (2016)
26. Zoph, B., et al.: Learning transferable architectures for scalable image recognition. CoRR abs/1707.07012. arXiv preprint arXiv:1707.07012 (2017)
27. Brock, A., et al.: Smash: one-shot model architecture search through hypernetworks. arXiv preprint arXiv:1708.05344 (2017)
28. Baker, B., et al.: Accelerating neural architecture search using performance prediction. arXiv preprint arXiv:1705.10823 (2017)
29. Elsken, T., Jan-Hendrik, M., Frank, H.: Simple and efficient architecture search for convolutional neural networks. arXiv preprint arXiv:1711.04528 (2017)
30. Pham, H., et al.: Efficient neural architecture search via parameter sharing. arXiv preprint arXiv:1802.03268 (2018)
31. Liu, H., Karen, S., Yiming, Y.: Darts: differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
32. Erhan, D., et al.: Visualizing higher-layer features of a deep network. Univ. Montr. **1341**(3), 1 (2009)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol. 8689. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
34. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (2018)
35. Mahendran, A., Andrea, V.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
36. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
37. Ballard, D.H.: Modular learning in neural networks. In: AAAI (1987)
38. Badrinarayanan, V., Handa, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
39. Thomee, B., et al.: The new data and new challenges in multimedia research. CoRR abs/1503.01817 (2015)
40. Strehl, A., Joydeep, G.: Cluster ensembles–-a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
41. Krizhevsky, A., Ilya, S., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
42. Simonyan, K., Andrea, V., Andrew, Z.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)
43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR, abs/1512.00567 (2015)
44. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Muller, K.-R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. **65**, 211–222 (2017)
45. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)

46. Folz, J., et al.: Adversarial defense based on structure-to-signal autoencoders. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2020)
47. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
48. Kurakin, A., Ian, G., Samy, B.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
49. Carlini, N., David, W.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE (2017)
50. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
51. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
52. Kumar, D., Taylor, G.W., Wong, A.: Opening the black box of financial ai with clear-trade: a class-enhanced attentive response approach for explaining and visualizing deep learning-driven stock market prediction. arXiv preprint arXiv:1709.01574 (2017)
53. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: 2015 IEEE Information Theory Workshop (ITW), pp. 1–5. IEEE (April 2015)
54. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
55. Siddiqui, S.A., Mercier, D., Munir, M., Dengel, A., Ahmed, S.: Tsviz: demystification of deep learning models for time-series analysis. IEEE Access **7**, 67027–67040 (2019)
56. Saad, E.W., Wunsch II, D.C.: Neural network explanation using inversion. Neural Netw. **20** (1), 78–93 (2007)
57. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl.-Based Syst.**8**(6), 373–389 (1995)
58. Munir, M., Siddiqui, S.A., Küsters, F., Mercier, D., Dengel, A., Ahmed, S.: TSXplain: demystification of DNN decisions for time-series using natural language and statistical features. In: Tetko, I., Kůrková, V., Karpov, P., Theis, F. (eds.) Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science, vol. 11731. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30493-5_43
59. Davis, E., Marcus, G.: GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review (2020)
60. Etzioni, O., Li, M.: High-stakes AI decisions need to be automatically audited. WIRED (2019)
61. Weld, D., Bansal, G.: The challenge of crafting intelligible intelligence. Commun. ACM**62** (6), 70–79 (2019)
62. Fichte, J.K., Hecher, M., Szeider, S.: A time leap challenge for SAT-solving. In: Simonis, H. (ed.) Principles and Practice of Constraint Programming. CP 2020. Lecture Notes in Computer Science, vol. 12333. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58475-7_16
63. Palacio, S., Folz, J., Dengel, A., Hees, J., Raue, F.: What do deep learning networks like to see?. In: Proceedings CVPR 2018 International Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA (June 2018). https://arxiv.org/abs/1803.08337
64. Folz, J., Palacio, S., Hees, J., Dengel, A.: Adversarial defense based on structure-to-signal autoencoders. In: Proceedings WACV 2020, IEEE Winter Conference on Applications of Computer Vision, Aspen, Co, USA (March 2020). https://arxiv.org/abs/1803.07994
65. Floridi, L.: AI and its new winter: from myths to realities. Philos. Technol. **33,** 1–3 (2020). https://doi.org/10.1007/s13347-020-00396-6

66. Shead, S.: Researchers: are we on the cusp of an 'AI winter'?BBC News (2020). https://www.bbc.com/news/technology-51064369. Accessed 27 Nov 2020
67. McKenney, P.E. (ed.): Is parallel programming hard, and, if so, what can you do about it? (2017). https://www.kernel.org/pub/linux/kernel/people/paulmck/perfbook/perfbook.2017.01.02a.pdf
68. Gleixner, A., et al.: MIPLIB 2017: data-driven compilation of the 6th mixed-integer programming library. Mathematical Programming Computation (2020). (accepted for publication)
69. Bixby, R.: A brief history of linear and mixed-integer programming computation. Documenta Mathematica, Extra Volume: Optimization Stories, pp. 107–121 (2012)
70. Daugherty, P.R., Wilson, H.J.: Human+Machine: Reimagining Work in the Age of AI. Harvard Business Press, Boston (2018)
71. Travis, G.: How the Boeing 737 Max Disaster looks to a Software Developer. IEEE Spectrum, Piscataway (2019)
72. Hand, D.J., Khan, S.: Validating and verifying AI systems. Patterns **1**(3), 100037 (2020)
73. Hutter, F, Kotthoff, L, Vanschoren, J. (eds.): Automated Machine Learning: Methods, Systems, Challenges. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-030-05318-5
74. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 847–855 (2013)
75. Hoos, H.H.: Automated algorithm configuration and parameter tuning. In: Hamadi, Y., Monfroy, E., Saubion, F. (eds.) Autonomous Search. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21434-9_3
76. Blot, A., Hoos, H.H., Jourdan, L., Kessaci-Marmion, M.É., Trautmann, H.: MO-ParamILS: a multi-objective automatic algorithm configuration framework. In: Festa, P., Sellmann, M., Vanschoren, J. (eds.) Learning and Intelligent Optimization. LION 2016. Lecture Notes in Computer Science, vol. 10079. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50349-3_3