A White Paper on the Future of Artificial Intelligence

Research Proposal · July 2019

DOI: 10.13140/RG.2.2.32564.19844

2 authors:

Helmut Linde
Covestro Deutschland AG
18 PUBLICATIONS **210** CITATIONS
SEE PROFILE

Immanuel Schweizer
Merck
47 PUBLICATIONS **399** CITATIONS
SEE PROFILE

A White Paper on the Future of Artificial Intelligence

Helmut Linde*
Merck KGaA, Darmstadt, Germany
helmut.linde@merckgroup.com

Immanuel Schweizer
Merck KGaA, Darmstadt, Germany
immanuel.schweizer@merckgroup.com

July 5, 2019

Abstract

In the present white paper we discuss the current state of Artificial Intelligence (AI) research and its future opportunities. We argue that solving the problem of *invariant representations* is the key to overcoming the limitations inherent in today's neural networks and to making progress towards Strong AI. Based on this premise, we describe a research strategy towards the next generation of machine learning algorithms beyond the currently dominant deep learning paradigm.

Following the example of biological brains, we propose an unsupervised learning approach to solve the problem of invariant representations. A focused interdisciplinary research effort is required to establish an abstract mathematical theory of invariant representations and to apply it in the development of functional software algorithms, while both applying and enhancing our conceptual understanding of the (human) brain.

***K*eywords** Artificial Intelligence · Invariant Representations · Neuroscience · Strategy

1 Introduction

In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton published [15] their results on the ImageNet LSVRC-2010 contest, a computer vision challenge to automatically classify 1.2 million high-resolution images into 1,000 different classes. Their use of *deep neural networks* yielded a substantial improvement in error rate and marks the beginning of the recent wave of interest in machine learning and artificial intelligence. In the subsequent years, deep learning has been applied to a considerable number of other problems and used productively in applications such as voice recognition for digital assistants, translation software, and self-driving vehicles.

But despite all these impressive success stories, deep learning still suffers from severe limitations. For one thing, enormous amounts of labeled data are required to train the networks. Where a human child might learn to recognize an animal species or a class of objects by seeing only a few examples, a deep neural network typically needs tens of thousands of images to achieve similar accuracy. For another thing, today's algorithms clearly are far away from grasping the essence of an *entity* or a class in the way humans do. Many examples show how even the most modern neural networks fail spectacularly in cases that seem trivial to humans [22].

While neural networks are quite fashionable nowadays, their conceptual foundations are actually rather old; they were already being intensely studied in the 1950s and 1960s, inspired by the brain's anatomy – according to the understanding at that time. Today's deep neural networks are essentially the same as those classical networks except for their higher number of layers. They owe their success in recent years largely to an increase in computing power and the availability of huge amounts of training data.

Our central hypothesis, which drives our research strategy, is that the current limitations in AI can only be overcome by a new generation of algorithms. These algorithms will be inspired by today's neurosciences and – to some extent – by advances in our understanding of the brain which are yet to come. Our envisioned path forward is an interdisciplinary

*Learn more at https://www.merckgroup.com/en/research/ai-research.html

research effort at the intersection of mathematics, computer science and neuroscience, enabling us to tackle this challenge from three different angles:

1. A mathematical model of invariant representations and how to learn them in an unsupervised way,
2. Prototypical implementations of unsupervised representation learning algorithms, and
3. Inspiration and validation of our algorithms by comparison to biological brains.

The remainder of the paper is organized as follows. In the next section, we will discuss Artificial Intelligence with a focus on the feasibility of Strong AI as a long-term goal and different aspects of value generation along the way. Afterwards, we will shed some light on the role that neurosciences can play today to inspire AI research. In the last section, we will discuss our research strategy in more detail and propose some specific and tangible ideas on where to start.

2 The future of Artificial Intelligence

2.1 Definitions

Intelligence, while a broad and comprehensive concept, is also a notoriously elusive one. In their comprehensive survey of available definitions of intelligence, Legg and Hutter [18] list and review more than 70 different notions. Extracting the most common features, they define intelligence as follows:

> Intelligence measures an agent's ability to achieve goals in a wide range of environments.

The difficulty in grasping what intelligence actually is directly carries over to the attempts of emulating it in machines. The term *Artificial Intelligence* has been around for many decades and, depending on technological progress at the respective time, it has carried quite different connotations. The fact that the marketing departments of large software companies have captured the term AI for their own purposes has not helped to improve clarity either. If we apply Legg and Hutter's definition strictly, then no Artificial Intelligence exists today. No existing computer program today is able to achieve goals in a range of environments comparable to the one in which humans operate. On the other hand, algorithms can perform certain well defined cognitive tasks - like playing a computer game or recognizing the race of a dog in an image - on (super-)human level. Hence, the AI community distinguishes between **Narrow AI** and **Strong AI**.

Similar to the definition above, we can define **Narrow AI** as an agent's ability to achieve goals in a (very) limited range of environments. Often it is assumed that the performance of this Narrow AI will be as good or better than the human performance. From self-driving cars to AlphaGo's defeat of Go champion Lee Sedol [25], all the current AI breakthroughs are examples of Narrow AI.

A **Strong AI** – also called General AI or Artifical General Intelligence (AGI) – would be a system which fits Legg and Hutter's definition above, i.e., it would achieve goals in a wide range of environments. Again, it is often assumed that in doing so this agent would be at least as successful as an average human being. In particular, such a Strong AI would pass the famous Turing test [29] by tricking a human conversation partner in an online chat (and taking into account recent advances in voice synthesis, certainly also on the phone) into believing that it is actually another human rather than a machine.

2.2 Consequences of Strong AI

The current wave of AI is driven by Narrow AI and the impact is already significant. The realization of Strong AI, however, would likely lead to the most dramatic changes to society and economy in human history. The precise nature of these changes cannot be foreseen as of today and it is beyond the scope of this paper to present any of the many different conceivable scenarios. The reader is referred to the following books providing several interesting perspectives to this fascinating question: Brynjolfsson and McAfee [3] (especially on the near- and mid-term economic consequences), Bostrom [2] (on the dangers of Strong AI), Kurzweil [17] (with mostly very optimistic long-term scenarios), and Tegmark [27] (with a more balanced view on different long-term scenarios).

2.3 Feasibility of Strong AI

With the stakes being that high, it is a question of paramount interest if Strong AI is technically feasible in principle and by when it can be realized.

The optimistic school of thought, as represented by people such as Ray Kurzweil [17], predicts the arrival of Strong AI for the 2030s or 2040s, based mostly on extrapolating available computing power according to Moore's law and

comparing it to estimates of the computational complexity of a human brain. As serious objection, one can point out that computing power alone (as measured in additions or multiplications per second) is not enough to replicate human intelligence, but that it is rather a question of having the right algorithm. As of today, the "algorithms" running in the brain that lead to intelligent behavior are only poorly understood, and it is much harder to predict the speed of progress on the algorithmic side than to extrapolate future hardware improvements. Nevertheless, it seems to have become a mainstream opinion that Strong AI will be possible during this century [2, 27]. To support this point of view, we will examine typical arguments that are brought forward *against* the possibility of developing Strong AI:

The complexity argument: "The human brain is so incredibly complex that we will never (or at least not soon) be able to replicate or simulate it."

It is true that the human brain is mysterious and incredibly sophisticated to the point that it is by far the most complex structure in the known universe [28]. Yet there are at least two reasons why this might not hinder us from successfully building a Strong AI:

Firstly, as Jeff Hawkins points out [11], a lot of the complexity of the brain is due to being a biological system that needs to fulfill many requirements irrelevant for a computer: it needs to be completely self-assembling in each phenotype, it likely carries a lot of unnecessary complexity for historical evolutionary reasons (keep in mind that nature is a tinkerer, not an engineer!), it needs to take care of a very complex mix of "hardwired" behavior (like controlling body temperature, heart rate, hunger, and sexual desires) and flexibly learned behavior (like playing chess), and it needs to do so reliably during all stages of its own development process from before birth to death. It also needs to get along with what little energy a biological body can produce, and it is limited to variations of the "biotechnology" (i.e. nerve cells) that evolution had happened to come up with for less intelligent animals. For our purposes of developing Strong AI, on the other hand, all these requirements can be dropped. We do not need to simulate a full human being - it is completely sufficient to replicate the core algorithms that provide its intelligence.

Think of the following analogy to make this point clearer: before the invention of aerodynamics people studied and admired birds for their ability to fly. Anatomical analysis would show that a bird - or even only its wing - is an incredibly complex "machine" made out of bones, muscles, arteries, feathers, and so on, and that there is no hope of replicating a bird's wing down to the finest details of each feather's structure. Yet once you understand the basic principles of aerodynamics, you see that only a few quite simple factors are important to make a bird fly: the geometrical shape of its body (especially the wings), its weight, and its relative speed to the surrounding air. Once the basic concept is understood, it is not that difficult to build an airplane, and it turns out that most of the bird-specific complexities (like flapping wings or delicately structured feathers) can happily be dropped.

Then there is a second reason to believe that intelligence emerges in a way that is relatively simple compared to the complexity of the brain as such: the neocortex is the brain region that performs cognitive tasks like recognizing images, understanding spoken language, controlling body movement, or generally thinking. It has been proposed that there is essentially only one "algorithm" executed everywhere in the neocortex that completes all these different and difficult tasks in a highly parallelized way. This idea of one neocortical algorithm goes back to Mountcastle [21], and it has been adopted and popularized by Jeff Hawkins [11], Ray Kurzweil [16], and others. It's an encouraging take home message: if you understand what a tiny piece of neocortex tissue does, you may have understood the essence of intelligence.

Two important arguments in favor of Mountcastle's idea are the uniformity of the microscopical tissue structure across the neocortex and its plasticity - i.e. the fact that a piece of neocortex tissue can take over very different tasks depending on the type of input signals that it receives. For example, Marina Bedny et al. showed [1] that the visual cortex of congenitally blind people can take over tasks in language processing. A review written by Daniel Feldmann [7] gives an overview on the phenomenon of neocortex plasticity.

The computational argument: "We simply don't have enough computing power to build Strong AI."

It is hard to estimate when we will have sufficient computational power to build a Strong AI as long as we don't understand how the algorithm works. The bird analogy above shows that it might be possible to achieve Strong AI with an algorithm that does much less than simulating the whole brain in its full complexity. Possibly even today's computing power would be sufficient. But even if not, under the assumption that Moore's law continues to hold, one can estimate that the available computing power will come close to that of a brain within the next decades [17].

The C-word argument: "You cannot build a self-conscious machine, because X," where X is an arbitrary philosophical or religious statement.

All arguments along the lines that only humans can be "truly intelligent" because only humans are endowed with a consciousness or a soul are flawed for the same reason: they fail to clearly separate the subjective metaphyiscal level from the objectively observable behavior of the system. Even though humans happen to have a consciousness *and* intelligence at the same time, there is no reason to believe that the two need to be causally connected. You have probably

already made many intelligent decisions completely sub-consciously and, conversely, you are conscious of yourself even when you lie in bed without doing or thinking anything intelligent.

When we talk about Strong AI, we do not imply that such a system will become self-conscious (though we would not rule it out either), and as far as technical feasibility is concerned, this question is completely irrelevant.

The creativity argument: "Only humans have the creativity to invent new things, do research, or perform art."

This is actually more an observation about the present than an argument about what may or may not be possible in the future. In fact, even as an observation about the current state of affairs this statement is not fully accurate anymore, as the example of AlphaGo Zero - which taught itself to play Go, inventing new moves and strategies on the way, including some previously unknown to humans - shows.

For the reasons above, we support the view that Strong AI is feasible in principle and that we are getting close to having sufficient computational power at our disposal to realize it (or maybe we even have it already). It seems that finding the right algorithms - inspired by the human brain but not replicating it too closely - is the key to achieving this goal.

2.4 Value Creation on the Way Towards Strong AI

Artificial intelligence provides considerable value today in different applications such as Alexa or Google Assistant, AI-generated music, modern picture tools with automatic search and sorting based on content, or modern smartphone cameras. And while there are many uncertainties around if and when Strong AI will be feasible, it seems likely that any step in that direction will yield by-products of significant intellectual and economic value.

Here are a few examples of potential intermediate results on the way towards Strong AI:

- Computer vision with minimal training data: while deep neural networks achieve human-level accuracy in certain image recognition tasks, the requirement of large training data sets still renders many use cases technically or economically unfeasible. In particular, this is the case when the system needs to learn a large number of edge cases (e.g. self-driving cars) or when the availability of training data is limited (e.g. in certain medical diagnosis scenarios).
- Natural language processing: it is still hard for computers to form an understanding of unstructured textual data. Recently there have been impressive steps in this direction such as Google's BERT [5], which uses unsupervised representation learning with bidirectional networks for transfer learning in natural language processing.
- Accurate action recognition: it is still quite challenging to train algorithms in recognizing actions performed in a video sequence [30]. Advances in action recognition might unlock economically relevant use cases, e.g. in security surveillance or in human-machine interaction.
- Motion control: despite impressive advances in robotics and motion control, even simple animals are in command of a motion control system which is more flexible, adaptive, robust and versatile than modern robots. For example, programming a robot to pick up a large variety of different objects (with different size, weight, shape and stability) is still a very difficult task today. Finding algorithms which mimic biological motion control could dramatically expand the range of use cases for robotics (e.g. in industrial, logistics or household applications) while reducing their cost of development.
- Content generation: today's Narrow AI is already capable of composing music (commercialized, for example, by the start-up company AIVA) or generating images of human faces [14]. But most content generation algorithms today operate in the space they are trained on. We are seeing the first approaches using reinforcement learning to be creative beyond the scope of the training data, and this will only improve on the way to strong AI. It seems possible that at some time most content - from music, text, and video to new chemical structures and medical active ingredients - can be created by AI.

In the following section we will explain why we believe that brain research can be an important source of inspiration for the next steps on the long way towards Strong AI.

3 Neurosciences as an Inspiration for AI Research

As stated in the introduction, today's deep learning is loosely based on neuroscientific concepts from roughly 60 years ago. Looking at the current explosion in algorithms and applications, it is safe to say that this inspiration has been a success. In the meantime, scientists have generated vast amounts of additional information about the brain. Will this knowledge help us to design a next generation of AI algorithms? A major difficulty is our still limited understanding of

the brain's fundamental working principles. How could they be revealed? Is it possible to advance neuroscience and AI research hand in hand? To give some intuition for these questions, it is quite instructive to start with an analogy from a related area where we do have a full understanding of the system in question: computers.

3.1 Can You Find Out How a Microchip Works?

Imagine that we had to find out how a computer works. We are given a large number of identical computers (so that we don't have to refrain from destructive experimental methods) and all the lab equipment we want, but without any prior knowledge about the principles of information technology. How would we proceed?

There are two obvious approaches to tackle the problem:

In the "high level approach", we would try to identify the purpose of the different components of the computer. For example, we would find out what the graphics adapter or the hard drive do simply by removing them and observing the effect that this has on the functioning of the computer. But, of course, it would still remain a mystery how these components work.

In the "low level approach", we would try to find out how any one of the components works internally. We might manage to open a processor without destroying it and make a series of experiments with it. We would probably be struck by the enormous complexity of the artifact: under a microscope we would see structures that look like a road network with millions of streets and crossings. We would note that the "streets" are electric conductors and the "crossings" have the peculiar property that electrical current can flow between two of their connections only if there is a voltage applied to the third one. We might call these crossings "transistors", and we would wonder how the computer can produce such complex behavior if it essentially consists only of such simple components.

In our ignorance of the working principles of a microchip, we would probably make more experiments and gather additional data. Jonas and Kording give a beautiful account of the type of information that you would find by applying the neuroscientist's arsenal of experimental techniques to a real-life microchip [13]. They also argue that none of the resulting data is likely to help us understand how a microchip really works.

Ultimately, would we be able to build a computer on our own? Certainly not. Even if we managed to reproduce tiny transistors and connections between them, we would have to replicate the enormously complex structure of the computer chip. Without an understanding of the underlying principles, we would have to build the chip blindly, without being able to test its components. Even the tiniest error in the wiring might make the chip useless.

Obviously, what we are missing is a theory of the basic operating principles of a computer. Such a theory would combine a few key concepts, in particular:

- Numbers can expressed in the binary system,
- Binary numbers can be stored as on/off states in an electrical circuit,
- Transistors can be connected in such a way that they perform basic mathematical operations on binary numbers, and
- Depending on the context, a binary number can be interpreted as a data point (e.g. the color of a pixel in an image), a command to the processing unit (e.g. to add or multiply some other binary numbers), or as an index pointing to some location in memory (e.g. to store the result of the computation there).

Once these basic concepts are clear, we would probably make rapid progress in understanding all the details of how the processor works. We might identify clusters of transistors which perform additions, other clusters which perform multiplications, or which move a binary number from one register to another and so on. The millions of seemingly chaotically connected transistors would soon turn out to be a much more manageable number of functional blocks, each with a clear meaning and purpose. Finally we would, at least in theory, be able to build our own computers.

We will call this theory that would enable us to understand the computer the "middle layer". The obvious reason is that it lies conceptually between the "low layer", i.e., all the physical details of how a transistor works, and the "upper layer", i.e. the over-all architecture of the computer with its graphics adapter, hard disk, and so on. It is important that this middle layer can really exist separately from the other two; you could replace the low layer by a completely different technology (e.g. vacuum tubes or electrical relays instead of transistors) and the middle layer would still remain valid. Similarly, the upper layer might assume forms as different as a digital wrist watch, a laptop computer, or an industrial robot, without having to change anything in the middle layer.

So how could we have discovered the ideas and concepts of the middle layer? Collecting more and more details about the lower and the upper layer doesn't seem to suffice. Granted, it will not hurt to have more information at hand, and

some of it might be useful to get some inspiration or sanity checks for hypotheses about the middle layer. But ultimately, there does not seem to be a direct way of deducing the middle layer from experiments and observations. Rather, it seems that in order to understand how a computer works, we have to *invent* the computer - guided and inspired by observations on the upper and lower layer, but ultimately from first principles!

3.2 The Brain's Middle Layer

As of today, the vast majority of available knowledge about the brain either belongs to the upper or lower layer, where we define the layers as follows.

The lower layer consists of the individual neurons that play their role as the "transistors of the brain". There is an enormous amount of information available on all the different types of neurons, their biochemistry, the ways how they connect to each other, and how they fire. Many sophisticated experimental methods have been developed in order to manipulate and observe the behavior of individual neurons or small groups of them. A large variety of theoretical models have been proposed to explain their behavior, e.g., from a chemical, physical, or information-theoretical perspective. For an introductory overview of such models, see Dayan and Abbott's book [4].

The upper layer of the brain consists of the different brain regions and the distribution of work between them. Historically, some of the earliest knowledge was obtained from observing the effects that brain lesions had on victims of war or accidents. For example, railway worker Phineas Gage became famous in 1848 for surviving an explosion that shot an iron rod through his skull, destroying much of his brain's left frontal lobe. The injury resulted in a dramatic change in Gage's personality, drawing a lot of scientific attention to the brain's role in determining personality. In the 1980s, several clinical case studies of brain patients have become known to a broader public through Sacks' book *The Man Who Mistook His Wife for a Hat* [24]. Brain lesions have also been afflicted intentionally in many animal experiments in order to study their effects. In addition, task-dependent activity in different brain regions can be studied directly through visualization methods like positron emission tomography (PET) or functional magnetic resonance imaging (fMRI). Consequently, there is a large amount of information about which part of the brain is involved in which task, such as image or voice recognition, abstract vs. emotional thinking, etc.

Yet despite the efforts of countless brain scientists, we still do not really understand the working principles of the brain - i.e., its middle layer. For example, it is far from clear how the brain learns to recognize objects regardless of the details of their appearance, how this information is encoded in memory, or how experiences can be combined and re-used in such a flexible way that common sense emerges.

It seems that our situation is similar to the microchip thought experiment: we know a lot about the brain's top and bottom layer, but this doesn't suffice to solve the mystery of the middle layer. And our best bet to understand the brain seems to be to invent one. However, inventing a theory for the whole brain is a monumental task. In order to make it more tractable, we will start with a fundamental property of the brain's perception: the capability to form invariant representations in an unsupervised fashion. We will discuss invariant representations as the focus of our research strategy in the next section.

3.3 Invariant Representations

It is a central function of the brain to analyze input streams, e.g., sensory impressions (like images or sounds), and to identify "entities" (like a cat or a song or a story) in this stream on incoming information. Memories of these entities must be encoded - or "represented" - somehow in the brain's neural structure. A fascinating feature of these representations is their invariance under entity-preserving transformations. For example, there is an astronomical number of ways a cat image could look like, given all the cat's possible positions, postures, lighting conditions, fur colors, etc. Nevertheless, the brain creates one *invariant representation* of what a cat is, and after having seen only a few examples every child is able to recognize any cat in an image. Similarly, a song can be played with different instruments, in a different key, or in different acoustic environments - but to our brain it remains the same song. A story can be told with different words, in a different language, in a book or in a movie - but it remains the same story.

These invariant representations seem to be the building blocks that form the basis of all higher cognitive functions; common sense largely rests on the ability to extrapolate ideas and concepts in a flexible way. As an example, assume that we find a bird of a species previously unknown to us. Our invariant representation of the concept "bird" firstly enables us to recognize this unknown animal as a bird at all. Secondly, it immediately leads us to the assumption that this animal can fly because most other birds can.

Abstract thinking might also be based largely on invariant representations. Even when thinking about something as abstract as a mathematical formula, we tend to have visual or auditory impressions in our minds - be it of how the formula looks like when written in a text book, the sound of pronouncing it, or some visualization of a concept that this

formula describes. Try thinking about the Pythagorean theorem without "hearing" $a^2 + b^2 = c^2$ or "seeing" the written formula or a sketch of a triangle in your mind!

Sensorimotor functions also seem to be tied to invariant representations: we have an idea of a certain type of action - e.g., "drink a cup of coffee" - and we can perform this action independent of the many details that might differ, like the size and the weight of the cup or the position of the handle.

In summary, invariant representations appear to be fundamental to human cognition and therefore we consider them to be of utmost importance on the path towards Strong AI. As of now, it is unknown how invariant representations are created in biological brains or how they could be created in abstract algorithms (a good overview of recent ideas and literature can be found in the book by Poggio and Anselmi [23]). Making progress towards the solution of this mystery is at the core of our research strategy.

3.3.1 The Cortical Algorithm

To find an angle from which we can approach the problem of invariant representations, let's think back to Mountcastle's conjecture of the one cortical algorithm. Is it possible that a single algorithm is powerful and flexible enough to learn how to recognize both a cat and a Beatles song? If you are skeptical, think about the enormous flexibility the human brain exhibits. It manages to create invariant representations based on very different kinds of input data; no matter if humans are blind, deaf, paralyzed, one-eyed, or short-sighted, they manage to create a mental representation of the world that is coherent and "correct" in the sense that it allows them to successfully navigate and interact with their surroundings.

If we assume that the *cortical algorithm* exists, then we must conclude that it can only make minimal assumptions about the input data it receives. We know that it must work with visual data captured on a two-dimensional retina as well as audio data registered as a frequency spectrum in the cochlea. On what assumptions can the algorithm be based? Or, in other words, what are the commonalities between entities as different as a cat and a Beatles song?

We hypothesize that the essential assumptions are as follows:

1. Entities are composed hierarchically: a 3D object consists of simpler objects, surfaces or edges and a sound is a mix of frequencies.
2. Entities may be (hierarchical) sequences in time: an action ('drink coffee') is a sequence of "sub-actions" ("take cup", "drink", "put cup on table"). A song is a sequence of notes, each of which is a time evolution of a mix of frequencies when played on a certain instrument. Note that in general entities are sequences in time, but they don't have to be. A motionless 3D object or a spoken vowel would be special cases of entities with a trivial time sequence.
3. Entities are stable in time: the way how we perceive entities changes at a much faster rate than the entities themselves. For example, when we watch a car driving down the street, we see it from many different angles before it is out of sight. Between appearing and disappearing, we will perceive one and the same entity.

Mountcastle's hypothesis makes us believe that no specific assumptions about visual images, sounds, or other specific types of sensory perceptions are hard-coded in the cortical algorithm. Traditionally, when developing a computer vision or a voice detection system, we would put as many reasonable assumptions as possible about the respective problem domain into the algorithm. In our cortical algorithm approach, this cannot be the case - our hypothesis is that the three assumptions above must suffice.

Furthermore, we can assume that the cortical algorithm creates invariant representations in a mostly unsupervised way. Of course, a child usually learns from his or her parents what a "cat" is. But even without external supervision, a human would learn to recognize cats as a class of animals and maybe invent a spoken word as a label for this class on his or her own.

3.3.2 Unsupervised Learning of Invariant Representations

We have seen above that there should be some unsupervised machine learning algorithm capable of detecting entities in a stream of input data and creating invariant representations of them under fairly weak assumptions. How might such an algorithm look like?

In order to identify something as an entity, the algorithm needs *repeated similar* observations. The repetition is necessary, since otherwise it might just pick up some noise for an entity. The problem is the concept of similarity (or rather the lack thereof): in order to decide whether two data points are similar, the algorithm needs a metric on the input data, i.e. a mathematical function that determines a distance measure between any two possible input data

points. Initially it might be tempting to make a naive choice like an Euclidean metric on the input data space. But in practice this attempt does not get us very far; for example, in the case of image recognition a simple translation results in two pictures which are practically identical (to the human eye) but are separated by a fairly large distance in the pixel-by-pixel Euclidean metric. A useful metric would have to take into account that two pictures can be entirely different on a pixel-by-pixel basis but still show the same object (seen from two different perspectives, for example).

Thus, if we want to limit ourselves to unsupervised learning based on only the three assumptions from the previous section, then we need an algorithm that can *learn* the right metric in an unsupervised way from the input data. The third assumption (stability in time) makes that possible in principle: the metric distance between two subsequent data points should usually be small, since two subsequent data points usually are manifestations of the same entity (or set of entities). Our algorithm should construct and progressively improve such a metric based on the input data it receives.

There is a problem: the space of possible metrics is astronomically large and we have only a comparatively small number of subsequently observed data points. How can the algorithm extrapolate from the observations to the full space? Here is a rough outline of how such an algorithm might work:

1. Recognize simple entities: at the beginning of our learning process, it is clearly impossible to recognize complex entities (like images of a cat or a full song), since usually there won't be repeated identical observations of the same entity and we do not yet have a useful concept of similarity. It is possible though to recognize certain elementary entities based on statistics; for example, in images the algorithm would find edges as repeating patterns. Similarly, in the case of music it might identify certain combinations of frequencies (e.g. tones and overtones) as elementary patterns.
2. Learn transformations: by observing the time dependency of the detected elementary entities, basic transformations could be learned. For example, in the case of video sequences the algorithm would learn that edges tend to move in certain ways that are highly correlated.
3. Refine the metric: based on the learned transformations, certain sets of elementary entities could be grouped together in equivalency classes. For example, the algorithm might decide that a set of edges that appear in certain positions relative to each other are always "the same thing", independent of where this set of edges is located in the image. (Putting combinations of entities into equivalency classes essentially refines the metric, i.e. the concept of similarity, on the input space: objects in the same class must be very close to each other in that metric sense.)
4. Recognize more complex entities: Thanks to the refined metric, new clusters of data points emerge and can be identified via unsupervised learning. Data points that appeared to be far apart in our initial naive metric might actually turn out to be part of the same equivalency class. Consequently, the algorithm can learn entities that are more complex than those found in the first step.
5. Iterate: more complex entities will exhibit more complex transformations, which in turn allow for larger equivalency classes, which then enable the algorithm to detect even higher level entities.
6. Learn sequences: the unsupervised entity detection process described above needs to be intertwined with the learning of repeated sequences on the different levels of abstraction. For example, the sound of a single piano key is a sequence of frequency patterns. A whole song, while also a sequence, needs to be represented on the level of notes and intervals in order to be independent from the instrument on which it is played.

Clearly the concept above still needs a fair amount of clarification and detailing. A simpler version of it has been proposed and implemented by Jeff Hawkins' company Numenta under the name Hierarchical Temporal Memory (HTM) [10]. The main addition of our approach compared to Numenta's HTM is the idea to use the temporal evolution of the input signal to learn transformations, create equivalency-classes from that, and thus unlock higher-level entities for unsupervised learning.

3.4 Summary of our AI Research Strategy

The primary objective of our research is to create new algorithms that go beyond deep learning to perform cognitive tasks in a way that is similar to how biological brains think.

Taking all the ideas and considerations from this white paper into account, we have defined the following research strategy to pursue this goal:

- **Focus on invariant representations:** the key to Strong AI is a theory of the brain – the brain's middle layer. The key to the brain's middle layer are invariant representations.

- **Unsupervised and assumption-poor:** in the brain, invariant representations seem to be formed in an unsupervised way by a single cortical algorithm. It is therefore our priority is to find algorithms that create invariant representations without supervision and with minimal assumptions on the input data.
- **Rethink the basics:** we work to understand the foundations of intelligence and to find radically new ways of creating AI. It is not our primary objective to improve classification accuracy in some machine learning challenge by yet another few percentage points. This means that we will sometimes choose levels of abstraction that might limit the practical usability in the short term but that might enable us to better understand intelligence.
- **Interdisciplinary approach:** we aim to solve the problem of invariant representations from three angles:
  1. The abstract approach: define mathematical objects that describe general entities and build a "theory of invariant representations" on top,
  2. The software approach: build prototype algorithms that implement aspects of the theory, and
  3. The neuroscience approach: identify neural circuits that implement aspects of the theory of invariant representations to validate and adjust the theory and to improve our understanding of the brain. We believe that to understand the brain, you have to invent it. Rather than trying to explain what the brain *does*, we focus on what a brain *should* do to obtain the capabilities that it is known to have.

The task that we have chosen is clearly not an easy one and support from or collaboration with others are most welcome. We aim to reflect the interdisciplinary nature of our work in the background of our research group's permanent staff - covering expertise in neuroscience, computer science, mathematics, and theoretical physics. In addition, we are interested in building up collaborations with academic research groups from the mentioned fields. Given our ambition to rethink the basics of AI, we also feel that important contributions might come from brilliant junior researchers with limited experience (and bias). We are therefore actively engaging with students and young researchers via internships, university competitions, and similar formats, trying to generate as many fresh ideas as possible for this fascinating and important field of science.

References

[1] M. Bedny et al.: "Language Processing in the Occipital Cortex of Congenitally Blind Adults" (Proceedings of the National Academy of Sciences 108, no. 11 (March 15, 2011): 4429-34)

[2] N. Bostrom: "Superintelligence - Paths, Dangers, Strategies" (Oxford University Press, 2014)

[3] E. Brynjolfsson, A. McAfee: "The second machine age" (Norton paperback, 2016)

[4] P. Dayan, L.F.Abbott: "Theoretical Neuroscience" (Massachusetts Institute of Technology, 2001)

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[6] J. J. DiCarlo, D. Zoccolan, and N. C. Rust: "How does the brain solve visual object recognition?", Neuron . 2012 February 9; 73(3): 415–434. doi:10.1016/j.neuron.2012.01.010.

[7] D. E. Feldmann: "Synaptic Mechanisms for Plasticity in Neocortex" (Annu. Rev. Neurosci. 2009. 32:33–55)

[8] B. Fritzke: "A growing neural gas network learns topologies" (Tesauro, G.; Touretzky, D. S. and Leen, T. K. (Eds.), Advances in Neural Information Processing Systems 7, MIT Press, 1995, 625–632)

[9] Future of Life Institute: "An Open Letter - Research Priorities for Robust and Beneficial Artificial Intelligence" (https://futureoflife.org/ai-open-letter/?cn-reloaded=1)

[10] J. Hawkins, S. Ahmad: "Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex" (Frontiers in Neural Circuits, 10 (2016) 1–13)

[11] J. Hawkins, S. Blakeslee: "On Intelligence" (Holt Paperbacks, 2005)

[12] J. Hawkins, M. Lewis, S. Purdy, M. Klukas, and S. Ahmad: "A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex" (Frontiers in Neural Circuits 12, p. 121, 2018)

[13] E. Jonas, K.P. Kording: "Could a Neuroscientist Understand a Microprocessor?" (PLoS Comput Biol 13(1): e1005268. https://doi.org/10.1371/journal.pcbi.1005268, 2017)

[14] T. Karras, S. Laine, T. Aila: "A Style-Based Generator Architecture for Generative Adversarial Networks" (arXiv:1812.04948 [cs.NE])

[15] A. Krizhevsky, I. Sutskever, G. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks" (NIPS, 2012)

[16] R. Kurzweil: “How to Create a Mind” (Penguin Books, 2013)

[17] R. Kurzweil: “The Singularity Is Near” (Penguin Books, 2006)

[18] Legg, Shane, and Marcus Hutter. “A collection of definitions of intelligence.” Frontiers in Artificial Intelligence and applications 157 (2007): 17.

[19] A. McAfee, E. Brynjolfsson: “Machine - Platform - Crowd” (Norton paperback, 2018)

[20] R. Mok, B.C. Love: “A non-spatial account of place and grid cells based on clustering models of concept learning” (bioRxiv. https://doi.org/10.1101/42184, 2018)

[21] V. B. Mountcastle: ”An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System” in G. M. Edelman and V. B. Mountcastle, eds., “The Mindful Brain” (Cambridge, Mass.: MIT Press, 1978)

[22] A. Nguyen, J. Yosinski, J. Clune: “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images” (Computer Vision and Pattern Recognition, IEEE, 2015)

[23] T. Poggio, F. Anselmi: “Visual Cortex and Deep Networks: Learning Invariant Representations” (MIT Press, 2016)

[24] O. Sacks: “The Man Who Mistook His Wife for a Hat” (Gerald Duckworth & Co Ltd., 1985)

[25] D. Silver et al.: “Mastering the game of Go with deep neural networks and tree search” (Nature 529, p. 484–489, 2016)

[26] D. Silver et al.: “Mastering the game of Go without human knowledge”. (Nature. 550, p. 354–359, 2017)

[27] M. Tegmark: “Life 3.0: Being Human in the Age of Artificial Intelligence” (Knopf, 2017)

[28] R. F. Thompson: “Brain: Introduction to Neuroscience” (W.H.Freeman & Co Ltd, 1985)

[29] A.M. Turing: “Computing machinery and intelligence” (Mind (journal) 236, p.433, 1950)

[30] M. Vrigkas, Ch. Nikou1, I.A. Kakadiaris: “A Review of Human Activity Recognition Methods” (Frontiers in Robotics and AI, 2015)