Minds and Machines (2020) 30:325–347
https://doi.org/10.1007/s11023-020-09538-3

GENERAL ARTICLE

The State Space of Artificial Intelligence

Holger Lyre[1]

Received: 11 October 2019 / Accepted: 19 August 2020 / Published online: 7 September 2020

Abstract
The goal of the paper is to develop and propose a general model of the state space of AI. Given the breathtaking progress in AI research and technologies in recent years, such conceptual work is of substantial theoretical interest. The present AI hype is mainly driven by the triumph of deep learning neural networks. As the distinguishing feature of such networks is the ability to self-learn, self-learning is identified as one important dimension of the AI state space. Another dimension is recognized as generalization, the possibility to go over from specific to more general types of problems. A third dimension is semantic grounding. Our overall analysis connects to a number of known foundational issues in the philosophy of mind and cognition: the blockhead objection, the Turing test, the symbol grounding problem, the Chinese room argument, and use theories of meaning. It shall finally be argued that the dimension of grounding decomposes into three sub-dimensions. And the dimension of self-learning turns out as only one of a whole range of "self-x-capacities" (based on ideas of organic computing) that span the self-x-subspace of the full AI state space.

Keywords Artificial intelligence · Deep learning · Self-learning · Semantic grounding · State space of AI · Self-x-property · Self-x-capacity

1 Introduction

There is much to suggest that 15 March 2016 should be regarded as a historical date. On this day Lee Sedol, one of the strongest Go players in the world, lost the last game of a tournament lasting several days against the "AlphaGo" AI system of the development company Google DeepMind. AlphaGo defeated the South Korean champion, 4 games to 1. The event attracted worldwide attention and brought back memories of the victory of IBM's "Deep Blue" against the then reigning world

✉ Holger Lyre
lyre@ovgu.de

[1] Chair of Theoretical Philosophy & Center for Behavioral Brain Sciences, University of Magdeburg, Magdeburg, Germany

chess champion Garri Kasparov some 20 years earlier. And yet the similarity of both events is rather superficial. DeepBlue owed its success to pre-implemented heuristic search combined with brute computational power, a strategy that is impossible for Go due to its sheer complexity. It is said that Go is to chess as chess is to checkers. Consequently, AlphaGo is based on a deep learning neural network (DL network), while DeepBlue was a classic rule-based and symbolic AI system.

DL networks belong to the latest development in neural network research. They are called "deep" as they consist of more than just two or three, sometimes even hundreds of layers. DL networks comprise various types of architectures such as feedforward, recurrent and convolutional neural networks (cf. Goodfellow et al. 2016, LeCun et al. 2015). The breathtaking successes of DL applications in the last 10 years have led to what Sejnowski (2018) calls the "deep learning revolution". What makes these systems special and what in fact distinguishes virtually all neural networks since the perceptron in the late 1950s is their ability to learn or, in view of more recent developments, to *self-learn* by actively exploiting or exploring big training data or by self-interaction with virtual or real environments. The deep learning revolution has led to a new hype in AI over the recent 10 years, be it in science, industry, economy or the media. These developments provide a strong motivation to rethink the question of what constrains the evolution of AI understood as the general quest to develop thinking machines or artificial minds. This is the motivation for the paper. And as we shall see, we come across various foundational issues in the philosophy of mind and cognition.[1]

The goal of the paper is to develop and propose a general model of the *state space of AI*.[2] It proceeds as follows. In Sect. 2, the notion of self-learning will be identified as a first dimension of the AI state space. We touch upon philosophy of mind issues, as the discussion leads to the blockhead objection and the black box problem. The second dimension, generalization, will be the topic of Sect. 3, where we also address the Turing test and its generalization. The central claim of Sect. 4 is that a third dimension consists in semantic grounding. The analysis is partly guided by three well-known philosophical issues: the symbol grounding problem, the Chinese room argument, and use-theoretic considerations of meaning. In Sect. 5, the full model of the state space of AI will be developed and explored. It shall be argued that not only grounding decomposes into three sub-dimensions, but that self-learning is in fact only one of a whole range of here so called *self-x-capacities*. They span the *self-x-subspace* of the AI state space, which, according to our analysis, turns out to be six-dimensional. The notion of self-x-capacity is related to the program of organic computing, where the notion of self-x-property is standardly used (self-repairing

[1] The recent developments in AI only start to get the full attention from philosophy of science and philosophy of mind and cognition that they deserve: cf. Buckner (2018, 2019), López-Rubio (2018), Páez (2019), Schubbach and Arno (2019), and Zednik (2019).

[2] I chose the term "AI state space", or sometimes just "AI space" for short. Indeed, "state space" might be misleading to readers who insist to restrict this term to the states of one system (one AI system, for instance) whereas the way it is used here considers the whole field of AI as one mega system, as it were, with concrete AI systems as states. Other, less catchy but substantially appropriate terms would be "space of possible developments" or "possibility space of AI".

and self-replicating, for instance, count as characteristic self-x-properties). Finally, a 10-dimensional state space of AI with *main dimensions* generalization, grounding and self-x-capacity will be defended.

2 Self-Learning as a Dimension of the Space of AI

AlphaGo and DeepBlue use radically different architectures. In contrast to DeepBlue, AlphaGo is based on a DL network (see LeCun et al. 2015 and Schmidhuber 2015a, b for overviews). The learning-based training of the network proceeded in two steps. In its initial phase, AlphaGo first learned patterns and moves of human Go players from a database of millions of moves. It was then able to give move recommendations similar to those of experts. In a second training phase, the system learned by self-play on the basis of reinforcement learning. This second phase demonstrates the crucial qualitative difference from AlphaGo to DeepBlue: the ability to *self-learn*.

2.1 The Notion of Self-Learning

Classical GOFAI ("Good Old Fashioned AI") builds on the assumption that intelligence and cognition consist of rule-based manipulations of symbols. In this regard, DeepBlue is a classic AI system *par excellence*. To calculate the evaluation function of millions of possible positions based on a given starting position (on average more than 100 million positions per second) DeepBlue relied on the expert knowledge of numerous chess grandmasters implemented in the calculation algorithms. As a non-learning system it was only able to operate within the framework of the given implementation. Such a limitation is typical for a classical GOFAI architecture: DeepBlue was designed for one special purpose, and was therefore unable to perform any other task than playing chess. It is a specialized or „narrow"AI (ANI: artificial narrow intelligence).

The more recent AI development almost reverses the original GOFAI doctrine. The ability to self-learn is precisely what opens up the field of flexible general intelligence. In retrospect, it seems hard to understand how the importance of learning could have been downplayed in the early stages of AI. Paradigmatic for this latter view is the position of Noam Chomsky (1980), according to which the human language ability is not otherwise understandable than under the assumption of a presupposed, allegedly innate deep grammar, i.e. a deeply anchored rule competence that is universal to humans. Chomsky considered it out of the question that such an ability could have arisen through imitation or reinforcement learning. Terrence Sejnowski comments on this very clearly:

> *What is innate is not grammar, but the ability to learn language from experience and to absorb the higher-order statistical properties of utterances in a rich cognitive context. What Chomsky could not imagine was that, when coupled with deep learning of the environment and a deeply learned value function honed by a lifetime of experience, a weak learning system like rein-*

> *forcement learning can indeed give rise to cognitive behaviors, including language….it follows logically from Orgel's second rule: evolution is cleverer than you are, and that includes experts like Chomsky.* (Sejnowski 2018, S. 251).

A few more distinctions about the notion of learning are in order. Machine learning (ML) in general is about developing algorithms to extract regularities or patterns from training data. This leads to learning a function that maps an input to an output. Learning can be either supervised, unsupervised or consist of elements of both. Reinforcement learning is the standard example for an in-between case. Learning algorithms are said to be supervised if both the input and the desired output are given. Backpropagation is probably the most well-known type of a neural network algorithm for supervised learning. The idea is that the difference between the network output and the supervisory teaching signal is used to backpropagate a rule for adjusting the weights of the network. Learning is unsupervised when only the inputs are given. In this case the system has to learn to extract regularities or patterns from the input data in a self-organized manner. Main methods in ML are principle component analysis and cluster analysis. Important types of unsupervised neural networks are self-organizing maps such as Kohonen networks or the Willshaw-Malsburg model. Reinforcement learning is typically considered as a third type of learning besides supervised and unsupervised learning. Here, learning improves on the basis of feedback in terms of reward for good performance (see the standard textbook by Sutton et al. 2018 for an overview).

A notion of special interest that somewhat crosscuts the above distinctions is the notion of *self-learning*. As Demis Hassabis, co-founder and CEO of DeepMind, emphasizes, self-learning systems learn directly either from first principles or from raw data. And they learn for themselves rather than being pre-programed.[3] While the notion of self-learning is sometimes used as synonymous to unsupervised learning, self-learning in the more general and broad sense as it is used here should be understood as the system's ability to exploit or explore the training data by itself, or, as in the case of AlphaGo's self-play, to even generate them. This would also include most cases of reinforcement learning, as the learning reward presupposes that such systems actively interact with their environments (whether artificial or real). And it also includes the important aspect of *meta-learning*, the ability of learning to learn (Botvinick et al. 2019, Schaul & Schmidhuber 2010). Moreover, some recent publications suggest improving AI systems by using more elaborate and biologically more plausible self-learning models inspired by neuroscience and the brain (cf. Bengio et al. 2016, Hassabis et al. 2017, Ullman 2019). Yann LeCun, one of the pioneers of the DL revolution, recently proposed to distinguish the category of *self-supervised learning* from the weaker unsupervised learning and the even weaker reinforcement learning: "in self-supervised learning, the system learns

[3] Demis Hassabis: The Power of Self-Learning Systems. Talk at MIT Center for Brains, Minds, and Machines, March 20, 2019. URL: https://cbmm.mit.edu/news-events/events/cbmm-special-seminar-self-learning-systems.

to predict parts of its input from other parts of its input".[4] Self-supervised learning is thus largely congruent with the general concept of self-learning introduced here (the latter comprising not only self-generated but also self-explored inputs).

2.2 Self-Learning, Blockhead, and the Black Box Problem

Self-learning systems show an obvious family resemblance to the human model. Unlike classic hand-crafted and rule-based GOFAI systems, they can therefore better counter Ned Block's (1981) famous objection against Turing machine functionalism, known as the "blockhead" objection. It aims at the fact that an AI program can very well show intelligent behavior without being truly intelligent. Note, for instance, that the number of sentences in a natural-language conversation is limited and shows various logical interrelationships. It is therefore in principle always possible that a machine system, the blockhead program, might successfully master a conversation by simply retrieving a pre-programmed look-up table of sentences. Block's objection touches on two aspects. First, a critique of a purely behaviorist understanding of intelligence and cognition. Second, the question of what role the internal structure of an intelligent system plays. Both aspects are interrelated. According to Block, a test of cognitive abilities and intelligence based solely on external behavior, such as the Turing test (see Sect. 3.2), is inadequate. There are obviously internal structures that do not produce intelligence.

It is rather evident that the blockhead objection can be directly applied to any system that uses a look-up tree algorithm. But many GOFAI systems are of course more sophisticated than that. DeepBlue, for instance, was partly based on built-in heuristic principles guided by human expert knowledge and a grandmaster game database. This allowed for a complex evaluation function to assess the quality of the positions reached and to explore the search space far more efficiently than by mere brute-force (Hsu 2002). Does DeepBlue really play chess then? The blockhead objection may not apply in a straightforward manner, but still the question can be denied. DeepBlue is merely manipulating internal symbols based on rules. And even if such rules include built-in chess knowledge, the system just relies on combinatorics and prediction on the basis of computational power. It does nothing to explore or to encounter chess in a self-driven or self-exploratory way. Human chess players, who cannot rely on such computational power, have to invoke self-learned and self-trained intuition and knowledge. These considerations will be taken up in Sect. 4, where they will be connected to the issue of grounding.

Applied to a self-learning, flexible system like AlphaGo, which is in part also capable of creativity, the blockhead objection seems implausible. It is, in fact, inappropriate precisely to the extent that the system fulfills the condition of self-learning. DL systems in general are largely free of programmed specifications. But what internal structures, logics or heuristics do these systems use? At this point a novel problem arises that affects much of the developments of the new wave of connectionism.

[4] Yann LeCun, April 30, 2019, on Twitter https://twitter.com/ylecun/status/1123235709802905600 and Facebookhttps://www.facebook.com/722677142/posts/10155934004262143/.

The internal structure of AlphaGo and related DL systems is only known in outline. This is a consequence of the deep structure and complexity of these systems as well as their ability to self-learn. This is known as the problem of *opacity* or the *black box problem* of deep neural networks.

The black box problem has led to *Explainable AI* (XAI) as a unique and novel sub-discipline in machine learning. It is driven by two motivations: on the one hand, the developers want to understand their own systems, on the other hand, the use of DL systems is in many areas–particularly obvious in the case of AI-based decision aids in medicine or self-driving cars–sensitive to whether and to what extent one is able to account for the question of how the systems do what they do. XAI therefore focuses, for example, on the development of tools of visualization and analysis that allow to open the black box and to understand the internal mechanisms of DL or related machine learning systems step-by-step and by means of reverse engineering (cf. Zednik 2019). In computer science, a remarkable interest has emerged to explore the methodology in the field of machine learning and thus to practice a kind of philosophy of science.

2.3 The Dimension of Self-Learning

The discussion so far should have made clear that the ability to self-learn must count as the first dimension of the state space of AI. This dimension unfolds the spectrum from non-learning, rigid rule-based systems over supervised to unsupervised learning and, finally, self-learning systems. The DL revolution has brought about systems, such as AlphaGo, that acquire significant values along this dimension, while classic GOFAI systems achieve low values at best (if at all). Neural networks from the first and second wave of connectionism may be ordered along in-between values.[5]

[5] The history of neural networks or connectionism may be divided into three historical phases or waves, each interrupted by a characteristic phase of stagnation. The 1st wave (1950–early 1960s) consisted of simple feedforward networks, especially the perceptron. After the first „neural winter" during which the GOFAI paradigm was dominant, the 2nd wave (1980–early 1990s) represented a powerful return of (neo-) connectionism with a flood of novel neural network models (e.g. Hopfield nets, Boltzmann nets, backpropagation, self-organizing feature maps, recurrent nets, spiking nets etc.). The renewed decline of connectionism in the 1990s during the 2nd neural winter (1990–early 2000s) can largely be attributed to a kind of „scaling problem": neuronal models designed for a manageable number of neurons (about 10-100) cannot so easily be scaled up to millions or billions of neurons without running into divergence problems. From 2010 onwards, the 3rd wave or DL revolution set in – mainly due to three factors: first, various achievements at the end of the 2000s and beginning of the 2010s made DL networks mathematically controllable and feasible (cf. Hinton & Salakhutdinov 2006, Krizhevsky et al. 2012), second, the rapid development of computing power (such as the development of fast GPUs, i.e. graphics processors) and, third, the huge amount of available training data only made possible by the Internet.

Regarding the usage of the term "neural winter" it should be added that there are in fact different notions of "winter" periods in the history of AI. The term "AI winter", for instance, is often used by GOFAI proponents to indicate the decline of research funding in the second half of the 1980s, while AI had been rather strong before. Almost the reverse is true for connectionism: while neural network research was down in the 1970s, the 1980s brought PDP and neo-connectionism back on stage. For our usage of the term "neural winter" compare Sejnowski (2018, pp. 1,35) for the first neural winter and Bengio (in: Ford et al. 2018, p. 25) for the second.

The self-learning dimension is potentially unlimited and infinite, much as any of the other dimensions that will be proposed here. It is of course also open for the evaluation of non-artificial, natural intelligent systems. Higher animal species, for instance, acquire higher values in terms of their self-learn capacities than today's DL systems (see Sect. 5.3). Humans still perform best, but future AI systems may outweigh natural intelligent systems in this regard.

3 Generalization as an AI Space Dimension

3.1 AGI and the Generalization Dimension

While AlphaGo still relied on human expert knowledge during its initial training phase, DeepMind was able to overcome this weakness just 1 year later with the development of AlphaGoZero. AlphaGoZero defeated AlphaGo in 2017 with an overwhelming 100:0. The accompanying Nature paper speaks of "Mastering the game of Go without human knowledge" (Silver et al. 2017), because AlphaGoZero is a self-learning system from scratch. Not only was it possible, for the first time ever, to create a machine that masters a cognitive task as complex as Go through self-learning alone. AlphaGoZero also masters Go on a super-human level that leaves behind the best human experts by orders of magnitude. A further outstanding feature of AlphaGoZero is that it is able to acquire additional skills through self-learning without any change in the basic architecture. Unlike DeepBlue, an ANI system, AlphaGoZero's capabilities are potentially generalizable within a large domain of tasks. In December 2017, DeepMind introduced the further developed system AlphaZero (Silver et al. 2017). It masters Go, Chess and Shogi on a superhuman level after only a one-day training phase. Other recent developments in the Alpha series include AlphaFold, a system for predicting the 3D structure of proteins, and AlphaStar, a system mastering StarCraft II, one of the most challenging real-time strategy games in e-sports (see deepmind.com, August 2019 release). StarCraft II is even more complex than Go. It demands dealing with incomplete information, and here again human top players have meanwhile been beaten. These are remarkable steps towards so-called General AI (AGI: artificial general intelligence)–the level of machine intelligence that is equal to human intelligence with regard to any task. AGI is thus very often considered as the equivalent to human-level AI (HAI).

Of course, AlphaZero is still a long way from real AGI. The specification of a special target function, with respect to which the system then allows generalization, is still a clear limitation. Nevertheless, DeepMind's developments showed from the outset a remarkable potential for generating creative solutions and not just short-term success-oriented strategies. For example, the 37th move in the second game by AlphaGo against Lee Sedol, in which the system violated a millennium-old Go wisdom, was seen by the experts as spectacular. It not only confused the grandmaster, but also soon proved to be crucial for AlphaGo to win the game. A creative solution that is easier to understand but no less original is already evident in one of the predecessor systems. DeepMind's first AI, a reinforcement learning based system called Deep Q-Network (DQN), successfully learned to play various classic Atari

computer games (Mnih et al. 2013). The game "breakout", for example, is about hitting a virtual stone wall in a kind of 2D-squash-court-setting with racket and ball, and to remove a stone with every wall hit and score points, until finally the whole wall disappears. DQN independently learned the tricky strategy of drilling a tunnel through the wall and pushing the ball into the back of the wall, causing it to remove large amounts of stones (as DeepMind founder Demis Hassabis reports, the developers did not know this trick, which is well-known among Atari gamers; cf. Tegmark 2017, p. 122).

As the examples show and as already implicit in the section before, self-learning systems have the decisive potential to generalize to new tasks and to come up with novel solutions. They may even solve things we don't know how to solve at all. *Generalization* thus counts as another important dimension that spans the state space of AI. It was a big step in the history of computing to go over from special purpose machines (e.g. a mechanical device for calculating the four arithmetic operations) to general purpose Turing machines (as foreseen by Charles Babbage in his Analytical Engine). It is likewise a big step to go over from ANI to AGI. Needless to say, this "step" corresponds to a gradual development, a continuous increase in terms of the capability of AI systems to master more and more general types of problems. As this development is potentially open-ended, the often found equalization of AGI and HAI is pretty misleading. Future AI systems may easily outweigh human intelligence in terms of their generalization capabilities.

3.2 Turing Test and Generalization

According to Turing (1950), if a machine succeeds in being indistinguishable from a human being in its response to arbitrary questions, then intelligence and higher cognitive abilities should be attributed to it. Numerous examples of Turing-like scenarios indicate the weakness and limitations of this test procedure. Weizenbaum's (1976) early experiences with his well-known imitation program ELIZA, which was able to conduct a psychotherapeutic dialogue, are telling (and were frightening for Weizenbaum): ELIZA was based on comparatively simple scripts and structured dictionaries, yet some test persons could not escape the impression of a conversation with a real psychotherapist. A more recent example is the Goostman chatbot. It scored a surprisingly good pass on several Turing test contests in the early 2000s. The bot simulates a 13-year-old Ukrainian boy, taking advantage of the fact that people more easily concede grammatical mistakes and lack of general knowledge to such a personality. These examples show that the Turing test offers no sufficient criterion for meaningful cognition, since it can be passed with too simple and possibly also "dishonest" means (as, for instance, in the form of a blockhead).

Assistance systems such as Siri, Cortana or Google Assist provide the contemporary variant of Turing-like scenarios. At its annual developer conference I/O 2018, Google surprised the general public with presenting Google Duplex, a system currently under development. It is meant to support everyday life, for example by making appointments for the user. Google had tested its system in real life by scheduling a restaurant reservation or calling a hairdresser to book an appointment. The

natural-language performance of the system is shockingly good: the called persons could not have guessed that they actually spoke to a machine. The phone calls were fluent and spontaneous including prosodic and non-verbal elements such as "hmm" and "uh" together with natural intonation and breaks. This provides another example of successfully passing the Turing test, this time by a DL-based AI.

It makes sense to try to generalize the Turing test along the axis of the generalization capabilities of an AI system. A Turing test for AGI should also require the system to have practical skills such as active navigation through natural environments, producing or repairing things, or social interaction and activities (cf. also Sect. 5). Nevertheless, any Turing test remains on the level of purely external functionality, an insight into the internal goings-on of the black box is against the behaviorist spirit of the test. This, however, is in strong contrast to the intuition of the blockhead objection (Sect. 2.2) according to which an understanding of the internal structure and mechanisms of an AI system is indispensable to assign cognitive or intelligent properties to it. Therefore, the Turing test should not be regarded as a sufficient criterion for intelligence. It may still be regarded as a necessary criterion: AI systems should perform functionally and behaviorally equivalent to humans in order to be regarded as cognitive. Sufficient for this attribution, however, is an understanding of the relevant internal structure of the system. At least insofar as it becomes possible to open the black box in part. It should further on be possible to provide information about the system's grounding, a feature that we consider in the next section.

4 Semantic Grounding as an AI Space Dimension

What's wrong with blockhead? After all, the system is able to master a conversation. But it seems clear that it has no *understanding* of what it is talking about. It cannot grasp the *meaning* of the words. In fact, it can't possibly talk *about* the world. It doesn't *refer* to the world, as it never had any contact with the world. In short: it has no *semantic grounding*. Genuine intelligence, however, needs at some point some sort of grounding. Semantic grounding will be our third candidate for an AI state space dimension.

4.1 The Symbol Grounding Problem and the Dimension of Functional Role Grounding

As argued in Sect. 2.2, DeepBlue doesn't really play chess as it essentially won by basic symbol manipulation and calculation without any self-driven exploration of chess. Conversely, how about a rigorous self-learning system like AlphaGo (or AlphaGoZero)? Does AlphaGo actually play Go? As an easier to grasp example, let us consider DQN as already briefly described in Sect. 3.1. It plays Atari breakout at a super-human level. As typical for DL systems, DQN draws on a gigantic number of training examples. In fact, the number of training games exceeds human training and, thus, the experience of human Atari gamers by orders of magnitude. This already suggests that the DL algorithms and network architectures do not strictly

correspond to those that play a role in humans. But this makes them no less successful with regard to certain capabilities. And they can nevertheless be regarded as a "proof of principle" for a biologically inspired connectionism (cf. Hassabis et al. 2017 for faster deep reinforcement learning models inspired by neuroscience).

Consider the training data of DQN. How can they be understood *from the perspective of the system*? For a human player, breakout's block-like pixel world, despite its minimalism, looks like a world of rackets, balls, walls and stones. There is nothing to suggest that this is the case for DQN. The system never had contact with real rackets, balls, walls or stones. From DQN's perspective, the only things that exist, as it were, are tons of pure pixel distributions. With this in mind, the system's superhuman playing abilities, and especially its additional ability to develop creative long-term solution strategies (like digging a tunnel), seem almost scary. At least it might seem that way. And the same seems to apply to AlphaGo or AlphaGoZero.

The problem can be seen as an instance of the *symbol grounding problem*, although this problem, in its original form, is aimed at classical symbolism (cf. Taddeo and Floridi 2005 for a review). A symbol is a physical token that is individuated based on its physical form and that can be linked to other symbols according to syntactic rules. Symbols are therefore elements of symbol systems. Symbolism regards the manipulation of physical symbols as necessary and sufficient for intelligence and cognition. This is compatible with a computational theory of mind according to which the brain as the vehicle of cognition is to be regarded as a computing device or Turing machine. According to Stevan Harnad the symbol grounding problem now consists in the following: „*Suppose you had to learn Chinese as a first language and the only source… you had was a Chinese/Chinese dictionary! … How is symbol meaning to be grounded in something other than just more meaningless symbols? This is the symbol grounding problem*"(Harnad 1990, p. 339–340).

In contrast to symbolism, connectionism emphasizes not only the network architecture of cognitive systems, but also a "subsymbolism" instead of symbol-based information processing. Superficially, neural networks do not operate on symbols but on inputs that represent (micro) features. In the case of DQN or AlphaGo, these are pixels with different gray or color values. However, since it can be shown that important classes of neural networks such as recurrent networks are Turing complete (cf. Siegelmann and Sontag 1995), these systems ultimately also operate symbolically insofar as they can be mapped to the symbolic operations of Turing machines. The question of whether and how the input pixel distributions are meaningful for DQN or AlphaGo amounts to the question as to what extent these distributions have a grounding or anchoring in the world. And superficially, it seems as if they do not have any such grounding. Therefore, neither DQN nor AlphaGo operate meaningfully, they do not understand what they are doing.

But this conclusion falls short because it overlooks an important distinction regarding meaning. Consider chess. How do the chess pieces get their meaning? What makes a knight a knight? Well, two things. First, it must be different in shape from all other types of pieces, such as rook or bishop. Second, it acquires its *meaning in the game* through exactly the role it plays in the game, which in turn is clearly assigned to it by the rules of the game. The knight (or any other chess piece) works like a physical symbol that is manipulated according to rules (in other words,

according to a syntax). And the semantics of the physical symbols, the chess pieces, originates from this syntax. In contrast, consider the words of a spoken language. They allow for sequences according to grammatical rules to form sentences. However, the question of what a particular word, such as the word "tree," refers to, is in no way determined by the grammar. Semantics in the sense of reference does not originate from syntax. We must therefore distinguish between meaning as *functional role*, which is determined by *internal* rules, and meaning in the sense of *external reference*. The symbol grounding problem primarily asks for meaning in the second sense: how can system-internal symbols be grounded in the external world so that they acquire a meaning in the sense of reference?

Consider again the example of Atari breakout. DQN seems not to dispose of the meaning of the terms racket, ball, etc. in the sense of reference. However, it is by no means excluded that the learning performance of DQN consists essentially in the fact that it recognizes certain stable and recurring patterns in pixel distributions and links them to regular behavior. A sufficient XAI analysis could provide exactly this kind of information, as it could show that DQN represents stable pixel configurations in higher layers and thus achieves the concepts racket, ball etc. in the sense of a *functional role semantics* (FRS). It is, moreover, reasonable to assume that, for the purposes of significance *in* the game, everything essential has been achieved by an FRS framework. For if we look for instance at AlphaGo, the question of semantics in the sense of reference does not arise at all, since Go pieces just like chess pieces do not refer to things or states of affairs in the world, but only possess an internal functional role within the system, i.e. a meaning *in* the game.

To conclude: we must expect the dimension of semantic grounding to decompose in at least two parts. Grounding on the basis of meaning as functional role, call it functional role grounding, and a more genuine form of grounding by means of reference-to-world. The latter will be our concern in the following two sections. The former is now identified as the FRS grounding dimension of the AI state space.[6]

4.2 The Chinese Room and the Dimension of Causal Grounding

Harnad's symbol grounding problem was inspired by Searle's related and well-known Chinese room argument (Searle 1980, 1990, cf. Harnad 1989, 2001). In a way, Harnad's argument makes the deeper core of the Chinese room argument explicit. The latter argument aims to show that syntax is not sufficient for semantics, and that the human brain is not a computer in the sense of symbolism. To this end, Searle conceives the Chinese room as the caricature of a Turing machine, where he himself takes over the role of a tape head for reading and writing by sitting in an otherwise empty room and by using a set of rules (the machine table) provided to

[6] One might complain that "FRS grounding" is a misnomer. Doesn't functional role always amount to *internal* functional role, while grounding has to be *external* reference-to-world? However, as we shall see, knowledge about functional role structure can have a bearing on grounding if this internal structure somehow mirrors external world structure (for whatever reasons). The example of Google translate in Sect. 4.2 will be a case in point.

him and allowing him to manipulate Chinese characters that he obtains in the room as input and that he reaches out as output. Since Searle doesn't understand Chinese and since for him Chinese symbols look like "meaningless squiggles", he insists that he can never attain the meaning of Chinese symbols in this way, i.e. by pure syntactic symbol manipulation.

In the 1980s, the Chinese room argument triggered a flood of reactions and discussions. Among the most common objections to the argument are the connectionist critique (Searle 1990) and the criticisms referred to by Searle as *systems reply* and *robot reply* (Searle 1980). Searle basically counters all objections according to the same strategy by showing that they just provide "more of the same". According to the systems reply the whole room rather than the internal operator is proficient in Chinese. Searle, however, argues that he could just as well internalize the whole room (particularly by learning the rule book) and still do nothing but mere syntactic symbol manipulation. According to the connectionist variant of the systems reply we are asked to consider an entire network of operators rather than a single operator. Searle, again, argues that we can as well imagine a "Chinese gym" with lots of operators manipulating symbols according to rules, but that still neither any of the operators nor the whole gym would thereby acquire the meaning of Chinese symbols.

Of particular importance is the robot reply. Would not a robot equipped with the rules of Chinese and operating in the Beijing marketplace gain the meaning of the previously merely syntactic symbols? Wouldn't a system in this way establish the necessary reference-to-world grounding? According to Searle, this is not the case, since the "computer inside the robot" (Searle 1980, 420) is still an analogue of the Chinese room. Be it that the input stems from an external camera and the output is used to control the arms, on the level of the internal computer that controls the robot both input and output still consist of nothing but mere meaningless symbols. This answer is reminiscent of a strange homunculus conception, and the question also arises as to whether a combination of systems and robot reply cannot be reinforced by further arguments from the areas of embodied and situated cognition (cf. Robbins & Aydede 2009). But we do not need to pursue this further here. Since Searle's argument is an argument against GOFAI's symbolism (for Searle: "strong AI"), it merely aims to show that meaningful thinking goes beyond algorithmic symbol manipulation. But if, in order to attain semantics, embodiment, social interaction and situatedness are crucial, this even ultimately strengthens the argument. And it shows, in essence, that the Chinese room argument boils down to the problem of grounding in the sense of reference.

Strangely, the DL technologies now available in the area of machine translation seem to realize the Chinese room scenario, at least in part. Freely available systems such as Google Translate or DeepL have shown a breathtaking improvement in their translation performance in recent years. And yet: one would hardly want to assume that any of these systems truly understand the texts they translate (sometimes in excellent quality). The new systems go beyond earlier forms of either rule-based or statistics-based machine translation. They extract rules of word selection, word order

etc. by self-learning on the basis of voluminous bilingual text corpora.[7] All this suggests the following: syntax is 'almost sufficient' to produce the linguistic behavior that corresponds to the behavior of speakers who are truly semantically grounded. Although syntax is not completely sufficient for semantics, syntax is 'almost sufficient' in the sense that it is sufficient for all practical purposes (but still insufficient from a strict Searlean point of view). This means that, in effect, a syntactic machine can be indistinguishable in its translation performance from a human speaker.

In the light of the above considerations, it follows that DL translation systems do not appear to have any semantic grounding. Nevertheless, these systems acquire rule knowledge and meaning in an FRS sense through self-learning. This alone is remarkable. The crucial step, however, is yet to come. How can it even be possible that a mere symbol based FRS becomes effectively indistinguishable (regarding language behavior) from a truly referential semantics? Should it be possible that, when it comes to semantics, we can do with a pure FRS alone and dismiss any appeal to reference? On the face of it, this amazing possibility seems to be suggested by the translation capabilities of systems such as Google Translate and DeepL. But on closer inspection, it is not. The systems do indeed go beyond a pure FRS. The text corpora used in learning allow for a kind of indirect reference-to-the-world. They were created by human speakers who dispose of a semantics in a full referential sense. Hence, Google Translate and DeepL have no direct but an *indirect grounding*. They refer to the world indirectly. Regularities that can be extracted from text corpora comparisons go beyond grammatical regularities, they provide world regularities as the texts deal with worldly circumstances. This allows to extract a decent amount of structural information about the world.[8]

All of this shows that semantic grounding in the sense of reference, even if in an indirect and tricky way as in our foregoing example, but most decisively in the direct variant of causal contact with the world, is of utmost importance to acquire meaning. It is thus of utmost importance for intelligence, whether artificial or natural. Grounding in the sense of causal reference is a crucial dimension of the AI state space. Note that postulating this dimension means no commitment to any particular theory of meaning or mental representation. It also means neither a realist nor an anti-realist commitment to content or representation. The criterion of grounding in the sense of causal reference is equally fulfilled in programs of naturalized semantics such as causal theories (Fodor 1987) or teleosemantics (Millikan 1984), as in recent accounts of structural representation (Cummins 1996, Ramsey 2007, Shea 2018) that are more in tune with instrumentalism. All such accounts entail causal world connections as central elements, and this is what counts for grounding by causal reference-to-the-world.

[7] A Google representative told George Dyson: "We are not scanning all those books to be read by people… We are scanning them to be read by an AI" (interview in Brockman 2019, p. 64).

[8] It may not allow to extract information about the intrinsicality of things in a rigorous ontological sense. According to the doctrine of structural realism, however, such a concept of intrinsicality is in conflict with our best knowledge about the bottom level and can therefore be regarded as doubtful anyway; cf. Lyre 2010.

4.3 Meaning as Use and the Dimension of Social Grounding

While causal grounding is certainly an element in almost all theories of meaning and representation, there are, however, accounts that downplay the role of reference such as conventionalism and use theories of meaning. Core ideas of the latter have been introduced by Ludwig Wittgenstein (Wittgenstein 1953). The central idea is to trace meaning back to linguistic use and social practice. According to Wittgenstein, the diversity of language can be seen in the variety of ways in which it is used. The focus is on the concept of rules. A classical, rule-based conception of language sees language as regulated by some unambiguous syntax. This applies all the more to formal languages or mathematics (and has tacitly been assumed in our considerations on the relationship between syntax and semantics). As Wittgenstein aims to show in his "rule following" considerations, such a strict Platonic conception of rules leads to an infinite regress. In order to set up the syntactic rules of, say, a certain Turing machine, other rules are required governing the former rules. But these too satisfy further rules–hence, regress.

According to Wittgenstein, language is governed by rules, but these rules presuppose a public practice and only become apparent in use. The rule following problem consists in the fact that language use and practice are always finite, but that no finite number of cases determines the "rules" of language use and thus the meaning of linguistic expressions under *all*, hence infinitely many, circumstances. Language rules are by no means rigid, but depend on the social context. Wittgenstein's bizarre thought experiment of the two-minute man drastically demonstrates the consequences of his conception:

> *Let us imagine a god creating a country instantaneously in the middle of the wilderness, which exists for two minutes and is an exact reproduction of a part of England, with everything that is going on there in two minutes. Just like those in England, the people are pursuing a variety of occupations. Children are in school. Some people are doing mathematics. Now let us contemplate the activity of some human beings during these two minutes. One of these people is doing exactly what a mathematician in England is doing, who is just doing a calculation. - Ought we to say that this two minute man is calculating? Could we for example not imagine a past and a continuation of these two minutes, which would make us call the process something quite different?* (Wittgenstein 1956, VI §34).

Wittgenstein's answer is obvious: the two-minute man "does not calculate" because he is not embedded in the practice and context of mathematics. Against this backdrop, let us consider the question whether AlphaGo actually plays Go. Games, like language, are limited by rules. Wittgenstein suggests a tight analogy between games and language and between the corresponding roles of rules and rule use. Indeed, he speaks of language as a "language game". Just as there is no mathematics or linguistic meaning without a social context, there are no games. Hence, from a Wittgensteinian understanding of use and practice, AlphaGo does not play Go since it lacks social context: the shared and public practice of the game of Go.

In Sect. 4.1 our conclusion was that the functional roles comprise everything that's essential in terms of meanings *in* the game. The main reason for this was that the meaning of moves and pieces is not referential, as for instance chess pieces do not refer to anything in the world. Following Wittgenstein, however, the meaning of games still has a kind of grounding, even if not in the referential sense. Instead, it is a kind of social grounding. Without public practice rules of games won't be subject to external control and, therefore, are no rules at all.

Wittgenstein's reflections on rule following are undoubtedly radical (and accordingly controversial), as the two-minute man scenario drastically demonstrates. Saul Kripke saw himself prompted to a radical rule skepticism, which infects not only rules of games or language, but even the rules of mathematics and logic (Kripke 1982). An in-depth discussion of these questions is far beyond the scope of the current paper. We shall assume that, for all practical purposes, rule knowledge can be set up in an AI machine modulo "Kripkensteinian" doubts.

Thus, for each version of the Alpha series, from AlphaGo to AlphaZero, the respective rules of the games to be learned were unambiguously implemented (Silver et al. 2017). The machine then develops a functional role semantics about the elements and overall setup of the game limited by these pre-determined rules. The systems of the Alpha series have no further grounding. Google Translate or DeepL, on the other hand, already have a rudimentary form of a socially anchored semantics, because these systems acquire an indirect social grounding in the course of their translation learning. After all, the text corpora on the basis of which the systems learn were generated by socially situated speakers, and are therefore parasitic with regard to their social practices. A future AI that combines, for example, the external performance of Google Duplex with the indirect grounding of world knowledge on the basis of Internet data could ultimately become a real part of our social practice of language and, hence, a real part of the language community. There is no convincing reason to assume that such systems would still lack a proper semantic grounding.

To conclude: social grounding is as important as causal grounding. To acquire meaning, intelligent systems must not only be coupled to the world, they must also share social practices. The debate about the ultimate theory of meaning and representation is still open in philosophy of mind and language, but for the time being it seems reasonable to assume both types of grounding as independent dimensions of the AI state space.

5 The State Space of AI

5.1 A Simplified Model Space

As a first shot and according to the foregoing sections, the AI state space is to be conceived as a three-dimensional space spanned by the dimensions:

- Self-learning (from rule-based to learning-based),
- Generalization (from narrow to general AI),
- Grounding (the degree of semantic world anchoring).

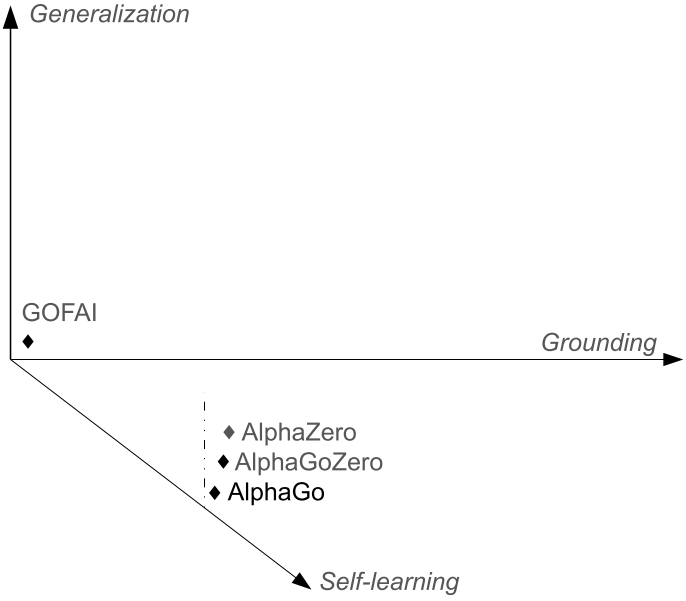

Fig. 1 Simplified AI state space

We already saw in Sect. 4 that the grounding dimension in fact decomposes into three sub-dimensions: functional role grounding, causal grounding and social grounding. Therefore, the full AI space has more than three dimensions. It is nevertheless instructive to look at the simplified three-dimensional model for a first orientation and to locate the systems discussed in this paper in this space: A classic GOFAI system like DeepBlue is close to the origin (see Fig. 1). Such systems are rule-based rather than learn-based, and almost all of them are narrow AI systems (e.g. DeepBlue is confined to chess). At best, a typical GOFAI system has an internal FRS (as DeepBlue captures the functional roles of chess pieces). AlphaGo sits at a much higher position in the self-learning dimension. From there we reach AlphaGoZero and AlphaZero by successive shifts parallel to the generalization axis. But none of the mentioned systems has a semantic grounding beyond FRS. At best, AI assistance systems such as Google Duplex move into this dimension, albeit still weakly at present.

It would be desirable to proceed from the state space topology (dimensionality and neighborhood) to a metric space (to determine distances). Human-level AI is a point of orientation (see Sect. 3.1). HAI has values in all dimensions and can therefore be used to calibrate the coordinate axes. In addition, systems that lie on the extended radial connecting line between origin and HAI (or within a suitably chosen spatial angle range) mark the area of superintelligence or superhuman AI (SAI), as roughly outlined in Fig. 2.

A detailed determination of the metric goes beyond the scope of this paper and is a task of further investigations. Let us, instead, focus once again on the

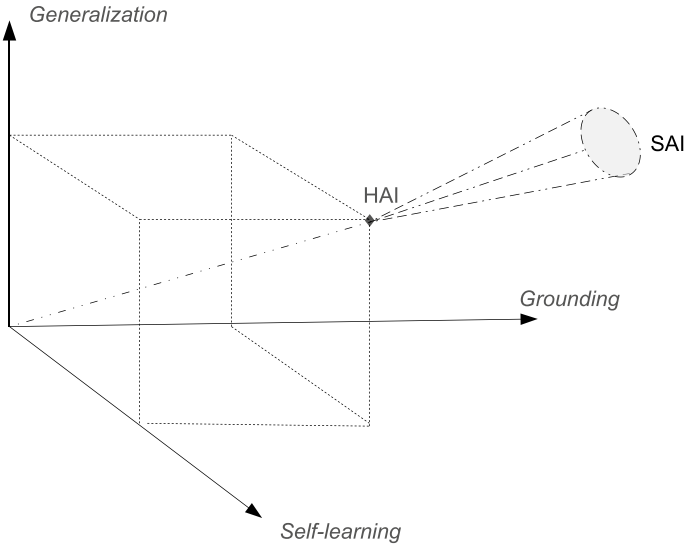

Fig. 2 Simplified AI state space with human-level and superhuman AI

dimensionality. As we already saw, the three-dimensional model offers a first orientation only, but is strictly speaking an approximation. It amounts to a simplified dimensionality reduction. While the generalization dimension is already correctly identified, we saw that the grounding dimension in fact decomposes into three further sub-dimensions. It could be dubbed a "main dimension" and actually represents a subspace of the AI state space. Let us first, however, consider self-learning.

5.2 The Self-x-Capacity Subspace

The dimension of self-learning needs some unpacking. Self-learning determines as to what extent an AI system can change its own configuration by means of *self-organization*. Surely there are further dimensions of self-organization that are crucial for the development of intelligence and cognition. The comparison with biological intelligent systems immediately suggests *self-repair* and *self-replication* as additional dimensions. They are so-called *self-x-properties*, as is the property of self-learning. Self-x-properties are central elements of the program of *organic computing* (cf. Würtz et al. 2008, Müller-Schloer et al. 2017). Here, the key idea is that self-organization or self-organized system configurations play a decisive role for intelligent systems. Depending on the author, the self-x-properties in organic computing include self-learning, self-configuring, self-optimizing, self-healing and self-protecting. Some of these properties are not quite distinct (e.g. self-healing and self-protecting), others are vaguely subsumed under self-configuring, although strictly speaking all self-x-properties are properties of the configuration (from this perspective, self-configuring and self-organizing are, as it were, "meta self-x-properties"). It seems reasonable to denominate, as a first attempt, *self-learning*, *self-repairing* and *self-replicating* as AI space dimensions.

In view of such self-x-properties, the term *self-x-capacity* shall be introduced here. The simplified model of the AI state space is still three-dimensional, but it consists of generalization, grounding and self-x-capacity as its dimensions, where the latter two are *main dimensions* in the above sense representing subspaces of the full AI space. The dimension of self-x-capacity, in particular, decomposes into various sub-capacities represented by the relevant self-x-properties that span the self-x-subspace of the AI state space. An exhaustive identification of all relevant self-x-properties and the necessary analysis of the terms configuration and self-organization goes beyond the scope of the current paper, but a somewhat anticipatory consideration should nevertheless be given.

First, keep in mind that our concern is not just about organic self-organization, but about the most general self-x-properties of any intelligent systems, whether artificial or biological. In this sense, the concept of AI just includes natural and especially human intelligence as special cases. It is therefore much likely that further self-x-properties must be accounted for. Recall Sect. 3.2, where it was argued that a Turing test on AGI should include not only pure responsive but also practical skills, where the system is asked to actively explore its environment. This suggests to add *self-exploratory* to the list of self-x-capacities. But even for Turing tests on AGI the black box problem essentially remains. In contrast, a qualitatively new level would be reached if AI systems were able to provide explanations, self-descriptions and justifications of their own responses and actions. The ability of being *self-explanatory* would then add to the above as a further significant self-x-capacity. And this finally raises the question of whether AI systems should not also be *self-conscious*. Therefore, and as a first attempt, the self-x-subspace appears to be six-dimensional:

- Self-learning
- Self-repairing
- Self-replicating
- Self-exploratory
- Self-explanatory
- Self-conscious

It should be clear that a detailed analysis of the self-x-capacities leads to far-reaching questions that go considerably beyond the scope of the current paper. For our purposes, the self-x-subspace of the AI state space is sufficiently characterized.

5.3 The Grounding Subspace

Let us finally turn to the *main dimension* of grounding and the corresponding AI subspace. Grounding, as in Sect. 4, is to be understood in the general sense of bearing semantics. We considered functional role grounding in Sect. 4.1, causal grounding in Sect. 4.2, and social grounding in Sect. 4.3. Therefore, the AI subspace of grounding is three-dimensional.

Moreover, like any of the other dimensions, the grounding dimensions are understood as continuous dimensions. This is a further important point that can only be

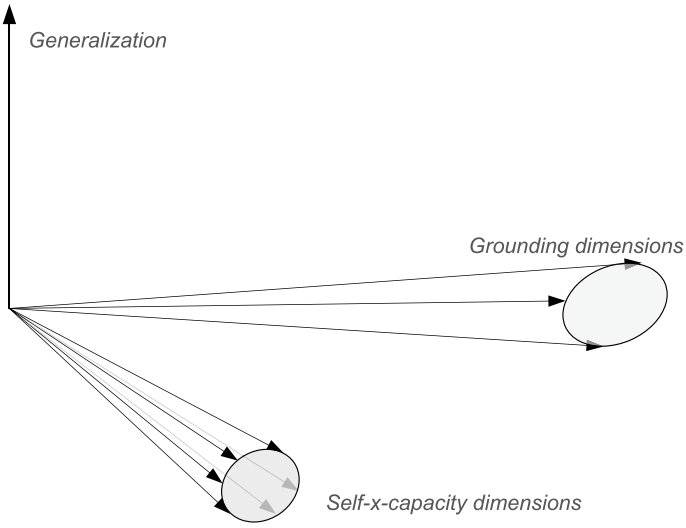

Fig. 3 The full AI state space with three main dimensions, two of which decomposing into sub-dimensions

touched upon here. Semantic grounding isn't on–off. Intelligent systems, whether biological or artificial, may be more or less grounded. The semantic skills of apes outweigh the skills of ravens, which in turn outweigh the skills of ants. The semantic skills pertain the way in which intelligent beings are grounded or anchored in the world in terms of their meaningful grasp and understanding of that world. Humans, in turn, trump the semantic skills of any known animal. But the gradual differences in terms of grounding exist of course also within a species. Healthy human adults exceed the semantic skills of newborns or patients with dementia. Moreover, semantic grounding is open-ended. Future AI systems may likewise outweigh the semantic skills of humans. The consequences of this are largely unknown and speculative. This is one of the pressing questions and, presumably, big worries with the issues of singularity and superintelligence (Bostrom 2013, Tegmark 2017).

5.4 Missing Dimensions?

Our analysis has led to a 10-dimensional AI state space that can be compactified by a three-dimensional model (see Fig. 3) consisting of:

- Self-x-capacity (*main dimension* decomposing into 6 sub-capacities).
- Grounding (*main dimension* decomposing into 3 sub-variants of grounding).
- Generalization.

One could criticize that several important features of natural and artificial intelligent systems do not count as dimensions. And again, while it might be the case that our model must be refined, it should be defended here against all too easy attacks.

A first complaint may be that features like computational power, speed, or accuracy are not captured by our model. To this the answer is that none of such quantitative technical features appear to be important for intelligence, at least not as such. For instance, a simple pocket calculator is much faster and more accurate in doing arithmetic than any (typical) human agent. But this alone doesn't make the technical device smarter or in any sense intelligent. In and of themselves, such quantitative capacities do not contribute to intelligence. Only insofar as such features help to elevate and promote the properties of self-x-capacity, grounding and generalization, are they relevant. And in this sense, they are covered in our model, if only indirectly.

What about core concepts in psychology and cognitive neuroscience such as attention, memory, control, decision-making and the like? As already pointed out in Sect. 2.1, many pioneering AI authors emphasize improving AI by using neuroscience inspired models (cf. Bengio et al. 2016, Hassabis et al. 2017, Ullman 2019). Knowledge and insights from neuroscience about such mechanisms will for sure influence the developments of future AI. But while the insights into such mechanisms may contribute to self-x-capacity, grounding and generalization, each of the mechanisms themselves are no fundamental dimensions of the AI space. This is even the case for such basic cognitive skills as perception, action and language. Ditto the modern concepts of embodiment and situatedness. The idea is that they are all covered indirectly by our more general dimensions. They may be central features of human cognition, but we should not postulate them as fundamental for AI systems in general. For even if it turns out that, say, language is necessary for social grounding, then social grounding will be the more general concept covering language and not vice versa. Or consider embodiment. It will most probably be entailed in the more advanced requirements of self-repair and self-replication. This is the rationale behind our dimensions.

6 Conclusion

The goal of the present paper was to develop and propose a general model of the state space of AI. Our analysis has led to a 10-dimensional space that can be compactified by a three-dimensional model consisting of self-x-capacity, grounding and generalization. These were dubbed as *main dimensions*, as self-x-capacity and grounding decompose into further sub-dimensions that span the corresponding subspaces of the full, 10-dimensional state space. Incorporating self-x-capacity and grounding as explicit features to classify and analyze approaches of AI is a particular value of the present model. While the distinction between narrow and general AI and hence, in our terminology, the generalization dimension is often considered in AI debates, the two other *main dimensions* have not previously been discussed and considered to the extent to which they are considered here. Surely, semantic grounding typically plays a role when AI is discussed from the point of view of philosophy of mind and philosophy of language. But it has seldom be considered as a characteristic feature playing the role of a gradual dimension along which AI systems may be developed and classified. A real novelty of the present model is to consider the various self-x-capacities as crucial for a classification of AI approaches.

While our discussion started with self-learning as an important dimension, we later argued that further capacities such as self-repairing, self-replicating and self-exploratory and ultimately self-explanatory and self-conscious should count as additional dimensions.

The model of the state space of AI developed here is only a first draft. It remains a task for future investigations whether more dimensions should be added or whether postulated dimensions should be changed or deleted. This, however, affects less the *main dimensions* than the sub-dimensions of grounding and self-x-capacity. The analysis of our paper should above all have supported the claim that the number of *main dimensions* is actually three. Compared to this the six-dimensionality of the self-x-capacity subspace should rather be understood as a first proposal open to further examination. What has been argued for, however, is that the dimension of self-learning is one of the central self-x-capacities. Finally, regarding semantic grounding, one further remark should be made on the question of whether three sub-dimensions do suffice. We made a distinction between FRS and social grounding as sub-dimensions. Yet, Block (Block & Ned 1998) has argued that we can distinguish between short-arm and long-arm versions of functional roles (cf. also Lyre 2016). While short-arm roles stop at the boundary of the system, long-arm functional roles extend beyond that boundary comprising social behavior, for instance. From this perspective, there is a gradual transition from a purely internal FRS to a socially anchored externalistic semantics. This then means that the two sub-dimensions above are far from orthogonal but merge into one. Obviously, a further discussion of these topics goes beyond the scope of the present paper. Whether the grounding subspace is truly three-dimensional remains to be seen, but it provides a reasonable working hypothesis.

To conclude, the usefulness of a general model of the state space of AI is obvious. AI developments can be better classified and related to one another by locating them in a state space. The future development could also be taken more clearly into account. Indeed, not only the occupied regions of the state space are of importance, but also those sectors that can possibly never be reached by any AI are of fundamental interest. For example, it could very well turn out that there are no AI systems that have a high degree of generalization but are at the same time not self-learning or grounded to a certain extent. These and many other considerations are the task of further research and exploration of the proposed state space of AI.

Acknowledgements Open Access funding provided by Projekt DEAL. The author would like to thank Carlos Zednik and two anonymous referees for valuable comments.

References

Bengio, Yoshua, Dong-Hyun Lee, Jörg Bornschein, Thomas Mesnard & Zhouhan Lin (2016): Towards Biologically Plausible Deep Learning. arXiv:1502.04156v3.

Block, Ned (1981). Psychologism and Behaviorism. *Philosophical Review, 90*(1), 5–43.

Block, Ned (1998): Semantics, conceptual role. In *The Routledge Encylopedia of Philosophy*, ed. E. Craig. London: Routledge.

Bostrom, Nick (2013). *Superintelligence*. Paths: Oxford University Press.

Botvinick, Matthew M., Ritter, Sam, Wang, Jane X., Kurth-Nelson, Zeb, & Hassabis, Demis (2019). Reinforcement Learning, Fast and Slow. *Trends Cognitive Sci, 23*(5), 408–422.

Brockman, John, editor (2019): *Possible Minds. 25 Ways of Looking at AI*. Penguin Press.

Buckner, Cameron (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese, 195,* 5339–5372.

Buckner, Cameron (2019). Deep learning: a philosophical introduction. *Philosophy Compass, 2019,* e12625.

Chomsky, Noam (1980). Rules and Representations. *Behavioral and Brain Sciences, 3*(127), 1–61.

Cummins, Robert C. (1996). *Representations, Targets, and Attitudes*. Cambridge: MIT Press.

Fodor, Jerry (1987): *Psychosemantics*. MIT Press.

Ford, Martin, editor (2018): *Architects of Intelligence: The truth about AI from the people building it.* Packt Publishing.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville (2016): *Deep Learning*. MIT Press.

Harnad, Stevan (1989). Minds, Machines and Searle. *J Theoretical Exp Artifi Intell, 1,* 5–25.

Harnad, Stevan (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena, 42,* 335–346.

Harnad, Stevan (2001): What's Wrong and Right About Searle's Chinese Room Argument? In M. Bishop & J. Preston (eds.): *Essays on Searle's Chinese Room Argument*. Oxford University Press.

Hassabis, Demis, Kumaran, Dharshan, Summerfield, Christopher, & Botvinick, Matthew (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron, 95,* 245–258.

Hinton, Geoffrey E., & Salakhutdinov, Ruslan R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313,* 504–507.

Hsu, Feng-hsiung (2002): Behind Deep Blue: Building the Computer that Defeated the World Chess Champion. Princeton University Press.

Kripke, Saul A. (1982): *Wittgenstein on Rules and Private Language.* Harvard University Press.

Krizhevsky, Alex, Ilya Sutskever & Geoffrey E. Hinton (2012): ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS 2012), Vol. 1: 1097-1105.

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey E. (2015). Deep learning. *Nature, 521,* 436–444.

López-Rubio, Ezequiel. (2018). Computational functionalism for the deep learning era. *Minds Machines, 28,* 667–688.

Lyre, Holger (2016). Active content externalism. *Rev Philos Psychol, 7*(1), 17–33.

Lyre, Holger (2010): Humean Perspectives on Structural Realism. In: F. Stadler (ed.): The Present Situation in the Philosophy of Science. Springer, p. 381-397.

Millikan, Ruth (1984). *Language, Thought and Other Biological Categories*. Cambridge: MIT Press.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra & Martin Riedmiller (2013): Playing Atari with Deep Reinforcement Learning. arXiv:1312.5602.

Müller-Schloer, Christian & Sven Tomforde (2017): Organic Computing-Technical Systems for Survival in the Real World. Birkhäuser.

Páez, Andrés (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds Mach, 29,* 441–459.

Ramsey, William (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Robbins, Philip & Murat Aydede, editors (2009): *The Cambridge Handbook of Situated Cognition*. Cambridge University Press.

Schaul, Tom, & Schmidhuber, Jürgen (2010). Metalearning. *Scholarpedia, 5*(6), 4650.

Schmidhuber, Jürgen (2015a). Deep Learning in Neural Networks: an Overview. *Neural Networks, 61,* 85–117.

Schmidhuber, Jürgen (2015b). Deep Learning. *Scholarpedia, 10*(11), 32832.

Schubbach, Arno (2019): Judging Machines. Philosophical Aspects of Deep Learning. *Synthese*. https://doi.org/10.1007/s11229-019-02167-z.
Searle, John R. (1980). Minds, brains and programs. *Behavioral Brain Sci, 3,* 417–457.
Searle, John R. (1990). Is the Brain's Mind a Computer Program? *Sci Am, 1,* 26–31.
Sejnowski, Terrence J (2018): *The Deep Learning Revolution*. MIT Press.
Shea, Nicolas (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.
Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *J Comput Syst Sci, 50*(1), 132–150.
Silver, David, et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529,* 484–489.
Silver, David, et al. (2017). Mastering the game of Go without human knowledge. *Nature, 550,* 354–359.
Sutton, Richard & Andrew Barto (2018): *Reinforcement Learning: An Introduction*. 2nd edition. MIT Press.
Taddeo, Mariarosaria, & Floridi, Luciano (2005). Solving the symbol grounding problem: a critical review of fifteen years of research. *J Experimental Theoretical Artifi Intell, 17*(4), 419–445.
Tegmark, Max (2017): *Life 3.0: Being Human in the Age of Artificial Intelligence*. Allen Lane.
Turing, Alan (1950). Computing machinery and intelligence. *Mind, 49,* 433–460.
Ullman, Shimon (2019). Using neuroscience to develop artificial intelligence. *Science, 363*(6428), 692–693.
Weizenbaum, Joseph (1976). *Computer Power and Human Reason*. From Judgement to Calculation. W. H: Freeman.
Wittgenstein, Ludwig (1953): *Philosophical investigations*. Macmillan Publishing Company.
Wittgenstein, Ludwig (1956): *Remarks on the foundations of mathematics*. Blackwell.
Würtz, Rolf P., editor (2008): *Organic Computing (Understanding Complex Systems)*. Springer.
Zednik, Carlos (2019). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philos Technol*. https://doi.org/10.1007/s13347-019-00382-7.