Neural Networks 152 (2022) 267–275

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue on AI and Brain Science: Perspective

Deep learning, reinforcement learning, and world models

Yutaka Matsuo [a], Yann LeCun [b,c], Maneesh Sahani [d], Doina Precup [e,f], David Silver [e], Masashi Sugiyama [g,a], Eiji Uchibe [h], Jun Morimoto [h,i,*]

[a] *The University of Tokyo, Japan*
[b] *New York University, Courant Institute & Center for Data Science, United States of America*
[c] *Facebook AI Research, United States of America*
[d] *Gatsby Computational Neuroscience Unit, University College London, United Kingdom*
[e] *DeepMind, United Kingdom*
[f] *McGill University, Canada*
[g] *RIKEN Center for Advanced Intelligence Project, Japan*
[h] *Advanced Telecommunication Research International (ATR), Japan*
[i] *Kyoto University, Japan*

ARTICLE INFO

Article history:
Available online 19 April 2022

Keywords:
Deep learning
Reinforcement learning
World models
Machine learning
Artificial intelligence

ABSTRACT

Deep learning (DL) and reinforcement learning (RL) methods seem to be a part of indispensable factors to achieve human-level or super-human AI systems. On the other hand, both DL and RL have strong connections with our brain functions and with neuroscientific findings. In this review, we summarize talks and discussions in the "Deep Learning and Reinforcement Learning" session of the symposium, International Symposium on Artificial Intelligence and Brain Science. In this session, we discussed whether we can achieve comprehensive understanding of human intelligence based on the recent advances of deep learning and reinforcement learning algorithms. Speakers contributed to provide talks about their recent studies that can be key technologies to achieve human-level intelligence.

1. Introduction

Although artificial intelligence (AI) systems now show super-human performances in such target regions as image and speech recognition, yet, our brain can do much better than AI in most of the tasks that we are easily dealing with in our daily life. For example, industrial robot motions are sometimes faster than human arm movements. However, such quick performance can be achieved only when robots repeatedly generate pre-designed trajectories without being adapted to unknown situations. Even inside a factory, many uncertainties exist, such as randomly placed components to be assembled on a production line. To handle the uncertain placement of target objects, end-to-end deep learning methods have been explored (Levine, Pastor, Krizhevsky, Ibarz, & Quillen, 2018). Even for the relatively simple object-picking task, many robots were needed to be involved to collect large-scale data from a real environment to acquire reasonable-level policies. Unlike image recognition tasks, collecting data for learning robot controllers is quite time-consuming and sometimes impossible since the robot needs to interact with its physical environment. Using a physical simulation to virtually train a policy and applying the acquired one to a real system would be a promising method to cope with the difficulty of collecting data in the real environment. Although this sim-to-real approach has been successfully implemented, for example, in a hierarchical RL framework (Morimoto & Doya, 2001) or with using domain randomization (Akkaya et al., 2019), so far, there are only limited applications.

To achieve human-level or super-human AI systems for wider applications, deep learning (DL) and reinforcement learning (RL) methods seem to be a part of indispensable factors while other approaches such as Bayesian inference (Ghahramani, 2015) and symbolic reasoning methods (Russell & Peter Norvig, 2020) are also important. On the other hand, both DL and RL have strong connections with our brain functions and with neuroscientific findings. In this review, we summarize talks and discussions in the "Deep Learning and Reinforcement Learning" session of the symposium, International Symposium on Artificial Intelligence and Brain Science (AIBS2020). The symposium aimed to bring together researchers advancing the forefront of AI and neuroscience to identify the next targets in creating brain-like intelligence and further advancing neuroscience. Professor Shun-ichi Amari led the discussion of this session of the symposium and raised the issue that there is no theory of deep learning except for individual minor ones. So far, these individual theories are not sufficient to achieve a comprehensive understanding of our brain.

[*] Correspondence to: Department of Brain Robot Interface, Computational Neuroscience Laboratories, Advanced Telecommunication Research Institute International, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan.
E-mail address: xmorimo@atr.jp (J. Morimoto).

https://doi.org/10.1016/j.neunet.2022.03.037
0893-6080/

Speakers provided talks about their recent studies that can be key technologies to achieve brain-like intelligence.

Each sub-section of this review, by and large, corresponds to each speaker's talk. Therefore, "we" in each section means members of the corresponding speaker's group rather than all the co-authors of this review. The review is organized as follows. We introduce DL-related talks in Section 2 and RL-related talks in Section 3. In Section 4, we overview learning methods and models. In Section 5, we discuss the small-sample learning problem. Finally, in Section 6, we describe possible future directions.

2. Deep learning

2.1. World model for perception, control and language (by Yutaka Matsuo)

A world model is a key to intelligent systems. We humans use a world model as a simulator in our brain. The model is obtained by learning from large amounts of sensorimotor data through interaction in the environment. We can learn the world model using deep generative models.

We humans deal with multimodal sensory information. We want to make better decisions and predictions by integrating different modalities. We try to expand the usual deep generative models to multi-modal. However, a problem exists. There is often a situation in which one modality is missing but other modalities are useful. If the amount of information related to the missing modality is large, it might result in a collapsed representation. This is what we call the missing modality problem. To resolve that difficulty, we propose a joint multimodal autoencoder (Suzuki, Nakayama, & Matsuo, 2016). Encoders for the respective modalities were prepared. Then we apply learning to them to approximate the original Variational Auto Encoder (VAE). After training, we can use each trained encoder for proper inference from a single modality. This method, Joint Multimodal VAE (JMVAE), can obtain the joint representation and well perform bidirectional generation because it explicitly learns to recover a missing modality from the observed modality. We use JMVAE for multi-modal neural machine translation. Our method outperforms other methods such as regular neural machine translation with and without images.

Recently, we have growing interests in RL especially for application to the real world. However, in many real-world applications of RL, the deployment of a new data-collection policy might be associated with several costs and risks. Therefore, it is important to reduce the number of deployments. We propose Behavior Regularized Offline Reinforcement Learning (BREMEN) (Matsushima, Furuta, Matsuo, Nachum, & Gu, 2020). It not only performs better than the state-of-the-art approaches on existing benchmarks, but it can also optimize a policy offline effectively using only a tenth or a twentieth of the data necessary for earlier methods. BREMEN learns a dynamics model, which can be regarded as a world model, from the offline dataset. It interacts with the learned model. The algorithm is based on Dyna-style model-based RL, learning an ensemble of dynamics models in conjunction with a policy using imaginary rollouts. Starting from a randomly initialized policy, it collects experience data and performs offline policy updates. To manage the discrepancy between the true dynamics and the learned model caused by the distribution shift in batch settings, we propose to use iterative policy updates via a trust-region constraint. For this study, we propose (1) the notion of deployment efficiency and (2) a simple means of achieving the goal. We think our results open a new direction, which is learning a world model from others' experiences. However, the physical differences between the learner and others are larger than a certain level, we cannot use the above approach. Such a case remains a challenge as future work.

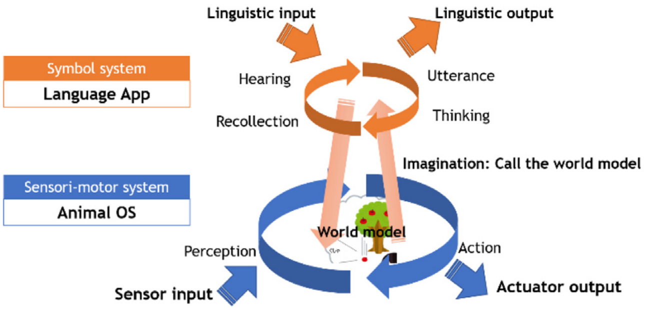

Fig. 1. Language App and Animal OS.

We hypothesize that human intelligence would roughly consists of two major components (Fig. 1). The lower part (sensory-motor system), which comprises world models and a controller, deals with real-world patterns and takes actions based on that. We call it "animal operating system (animal OS)" because the function closely resembles an OS in a computer system. The world model is obtained from multiple modals by methods such as JMVAE and by interaction with sample efficient behaviors such as BREMEN. The upper part (symbol system) deals with language. The system hears others' utterances, thinks, and produces an utterance accordingly. We call it "Language App", which is very specific to human intelligence. The Language App calls and uses the world model of animal OS as a module. For that, world models would be extremely important. Especially, a world model is triggered by language; it is used as a simulator. A deep generative model conditioned on input sentences is used for that purpose. We can imagine many things such as flying cars and mountain-high giants by using language. Thereby, we call it a mental canvas. The mental canvas would play a crucially important role for human intelligence. That is based on the world models trained in real-world situations, but it can be conditioned by language and be generated very flexibly.

2.2. Self-supervised learning (by Yann LeCun)

Progress in Self-Supervised Learning can be the next challenge towards making real progress in AI. Self-supervised learning seems to be one type of learning that we observe in humans and animals. Babies seem to learn basic concepts about the world in the first few months of life. What type of learning is taking place in the brain when babies perform this kind of learning? Trying to figure out the process is the biggest obstacle to making real progress in AI. It is quite possible that this type of learning through observation which does not seem to be task specific results in accumulation of a large amount of knowledge and perhaps this constitutes the basis of what we call common sense. Self-Supervised learning is basically learning to fill in the blanks (video clip, text and so on). There are two uses for self-supervised learning. The first one will be learning hierarchical representations of the world. The learned representations by self-supervised pre-training can be used in supervised learning or RL afterward. The second one is learning predictive (forward) models of the world. The learned predictive forward models can be used for model-predictive control or model-based RL. The essence of intelligence is the ability to predict and the big technical problem we are going to face is how to represent uncertainty/multi-modality in the prediction. For this, Energy-Based Model was proposed (LeCun, Chopra, Hadsell, Marc'Aurelio, & Huang, 2006). There are two types of methods for training Energy-Based Model, contrastive methods and regularized/architectural methods.

Contrasting methods have been extremely successful in recent years, particularly for applications in natural language processing. In the process of predicting missing words in the text, the system will learn good representations of texts that can be used in subsequent tasks (Vincent, Larochell, Bengio, & Manzagol, 2008). There has been a lot of success in natural language processing such as BERT, RoBERTa and so on.

Another form of contrastive learning is one in which the system is trained to learn a common representation between two identical networks. There has been a considerable success with techniques like PIRL (Misra & van der Maaten, 2019), MoCo (He, Fan, Wu, S., & Girshick, 2019), SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020). But the problem with contrastive learning is that it does not scale very well because it takes a lot of computation to train the system.

In regularized/architectural methods, a latent variable model is constrained to be sparse. Therefore, its information content is limited. It limits the volume of space that can take energy. Sparse AE and LISTA, which is a type of sparse AE, or Variational AE are basically models where there is a latent variable that encodes the input and can reproduce the output. But its capacity and its information content are limited. Mikael et al. proposed a predictive video prediction model that is trained to be able to predict what cars around a particular car on a highway are going to do (Henaff, Canziani, & LeCun, 2019). If you want to train the machine to drive itself, it is useful to be able to predict in advance what cars around you are likely to do. We feed a few frames of top-down view of highway and then we asked the system to predict the next frame of the next few frames and then keep predicting. And by choosing the latent variable in various ways, or by sampling it from a distribution, we can predict multiple futures. It uses drop out to regularize the latent variable.

Forward Model makes a prediction about the next state of the world as a function of the action we take and the latent variable that we draw which represents what we do not know about the world. We can enroll the system for multiple timestamps through back propagation. By gradient descent, we can find a sequence of actions that will minimize an objective in optimal control that is called the adjoint state method. What we can do a little more efficiently is to use this method to determine a new action each time. We need to do a process of developing the prediction, coming up with the best action, executing the action, and then repeating the process. And this may have to be done for multiple drawings of latent variables, which may be very costly. One way to accelerate this is, instead of doing a gradient descent with respect to action every time, training a policy network of neural net to predict the right action that will minimize cost. The back propagation will learn a policy that will minimize objective and learn to drive, for example by avoiding to bump into other cars where the cost function indicates how close you are to other cars or far you are from them, ensuring lane, etc.

We want to put together an entire intelligent system. Cost function indicates instantaneous cost of the state of the world. Critic would be a trainable function which is going to estimate or predict in advance what the ultimate cost of an outcome is going to be. Actor is going to either run this policy network or in case that skill has not been completely acquired yet, it will basically infer a sequence of action and optimize the cost through optimization. And there is a need for a perception module that estimates world state (Fig. 2).

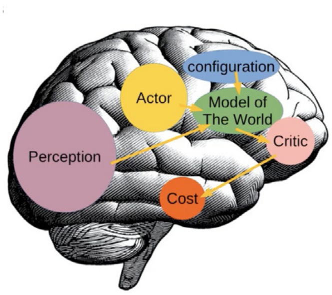

Fig. 2. Architecture of intelligent system.

Self-supervised learning would be the future of AI and machine learning. Model of the world needs to be trained so we need to find ways to represent uncertainty. By learning models of the world, machine will be able to accumulate sufficiently large amounts of knowledge about how the world works, so that some sort of common sense would emerge from it.

2.3. How do neural systems learn to infer? (by Maneesh Sahani)

Here, we think about the nature of a world model in terms of what it means to recognize or infer elements of the world. The process of making inferences involves a combination of a large variety of cues and beliefs based on those cues. How are the neural systems able to learn to parse the world and perform inferences? They do it in large part through data. The dominant stream of sensory data does not include any direct supervisory signals, but models are shaped by predictive comparisons; interaction and manipulation; reinforcement. With little supervision, the obvious signal for learning in a given model architecture is based on maximizing the joint or conditional probability of observations. Models generalize in familiar and unfamiliar environments — suggesting a learnt skeleton of causal interactions that can quickly adapt to new statistics. We want to stress a separation between the way things interact and the statistics of the objects that are present in the world. Though those statistics change from environment to environment, these principles of interaction still pertain.

The kind of model that is able to express this sort of dependence, the skeleton of interactions, is called a graphical model or probabilistic graphical model. There are two kinds of probabilistic graphical model, directed models and undirected models. Directed models connect closely to causality, but undirected models capture related structures of independence. And our question is how neural systems implement models like this.

A key part of this question is how neural activity encodes probability distributions. We are going to work in a framework that we call "distributive distributional coding (DDC)" (Sahani & Dayan, 2003). The idea is that P(z) may be represented by expectations of encoding functions Ψ_{i}(z):

$$\mu_i = E_{P(z)}\left[\Psi_i(z)\right] \ where \ \Psi_i(z) = g\left(w_i \cdot z + b_i\right), \tag{1}$$

Where g is a nonlinear function. This is a generalization of the idea of moments and has clear links to some ideas such as kernel-space mean embedding, predictive state representations, and exponential families.

One way of thinking about this representation is that each one of these expectations places a constraint on the encoded distribution. We can then regard the encoded distribution as the one that satisfies those constraints, but is otherwise as general as possible. What we mean by that might be that it has the maximum entropy of all distributions consistent with our constraints. If we choose that definition, it can be worked out that the distribution p(z) implied by the expectations will always have the following form:

$$p(z) \propto e^{\sum_i \eta_i \Psi_i(z)}. \tag{2}$$

The Lagrange multipliers η_i and normalizer are typically difficult to find from μ_{i}. But exponential-family distributions are, in

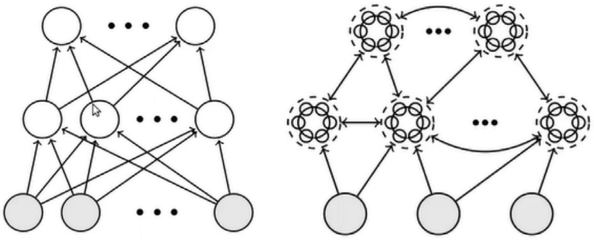

Fig. 3. Changing conceptual sketches.

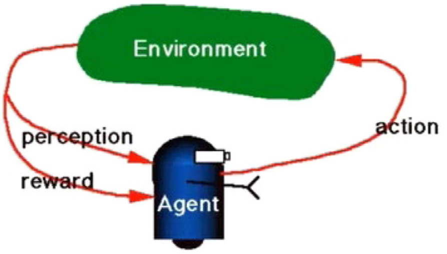

Fig. 4. Reinforcement learning.

general, described equally well by the natural parameters η_i or by the mean parameter μ_{i}. It is difficult to evaluate the density from just the mean parameters, but most of the computations that we want to do are not about calculating densities. What we need is to be able to evaluate an expected value. With a flexible set of basis functions, such expectations can be approximated by linear combinations of mean parameters. We have been using this idea to explore many questions (Vértes & Sahani, 2018, 2019; Wenliang & Sahani, 2019), but mapping directly to the workings of the nervous system is not as easy as we had hoped. In general, they require explicit, parametric generation in tractable form (“deep” conditional exponential family models). And this leads to an asymmetry in the generation and recognition process. It also, in places, depends on the feedforward perceptron conceptual sketch of neural systems, which neglects many salient features of connectivity in the real neural system. An alternative approach takes us from the left conceptual sketch in Fig. 3 to the one on the right, which reflects both the essential recurrent structure of local neural circuits (e.g. Douglas, Martin, K, & Whitteridge, 1989) and the reciprocal connections between local circuits in a partial hierarchy (e.g. Felleman & Fan Essen, 1991), albeit in much simplified forms.

While simplified, this sketch allows us to draw a formal mathematical relationship between such interconnected recurrent circuits and the sorts of structures needed to do probabilistic inference and deep learning in undirected probabilistic graph models. In our new conceptual sketch, the nodes are recurrent nuclei or columns of cortex that are densely interconnected and they themselves then connect to other such columns or nuclei in a reciprocal manner. This connection goes both ways and evidence of sensory experience is fed into some subset of these and then the whole system evolves dynamically to perform inference. Rather than inference being a feedforward pass as it is in the standard view of amortized inference, it becomes a dynamical process. There still is a potential feedforward pass to this dynamical process which may support a very rapid stage of inference so these two views are not necessarily inconsistent. But the focus is on the dynamical inferential process. By doing this, we have established this formal link between interconnected recurrent circuits, inference in graphical models and unsupervised learning. We have separated parameterization and learning into two terms at least, local distributions and interaction weights.

We have tried to create a new conceptual sketch for inference in neural systems, which allows us to map the dynamics of neural systems formally to a mathematical model of inference. It provides a substrate for general (structured) unsupervised learning and inference using this sort of architecture. We think of neural adaptation as basically short timescale learning and mostly of the local prior on the individual variables. Of course, there are multiple variables and these remain coherent because if the distribution of one variable changes and there is association with another then the distribution of the other variables should change and they should co-adapt appropriately.

It is worth stressing there is an interesting middle ground here between a connectionist structure. But there is also room for some innate shaping in the choice of the graphical structure. In principle, it is not the precise mapping of variables or distributions to neurons; variables, distributions and structure all learnt by flexible recurrent circuitry with biological learning rules. Therefore, there is some sort of disentangling necessary. The multiple timescales of learning may be key to disentangling. Good explanations partition explanatory variables into those whose short-term priors vary independently. But architectural scaffold can shape variables and relationships learnt. So, there is room for an evolutionary scaffold that dictates the form of the system that can be learned.

In relation to the world model, the model that leaned with a huge neural network in a self-supervised manner would internally contain the probability graph model. Therefore, the methodology to extract the probability graph can be important. In particular, the current deep learning techniques are still premature in terms of handling time and actions. Thus, the network structure needs to include many loops as it progresses.

3. Reinforcement learning

Reinforcement learning (RL) is a learning framework that improves a policy in terms of a given objective through interaction with an environment where an agent perceives the state of that environment (Sutton & Barto, 2018). A learning agent gets a reward signal depending on its actions, which also alter that environment’s state (Fig. 4). RL was developed at the intersection of ideas in artificial intelligence, neuroscience, and cognitive science. Many behaviorist’s ideas have been converted into concepts used in computational RL algorithms. RL is a general-purpose framework for decision making that can be applied in many kinds of situations whenever an artificial agent is in a situation where it has some action choices. For example, RL has been applied to robot control (Asada, Uchibe, & Hosoda, 1999; Endo, Morimoto, Matsubara, Nakanishi, & Cheng, 2008; Peters & Schaal, 2008). Its goal is simply to choose actions to make decisions that maximize future rewards as much as possible. In the following sections, we introduce previous achievements and key ideas and explain that RL is a strong framework for AI to achieve high-level intelligence.

3.1. Fast reinforcement learning with generalized policy updates (by Doina Precup)

Humans seem to efficiently utilize previously acquired policies in simple tasks to cope with complex problems. Such a divide-and-conquer approach can be very useful to reduce the amount of data and computation needed to solve a large-scale problem that cannot be handled by the standard RL. Generalized policy updates provide divide-and-conquer algorithms with reinforcement learning formalism to efficiently use previously learned policies to cope with novel tasks (Barreto, Hou, Borsa, Silver, &

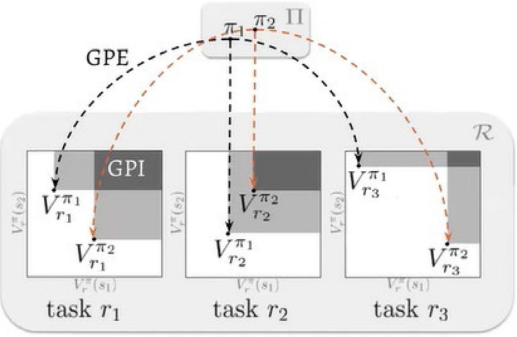

Fig. 5. Generalized policy update.

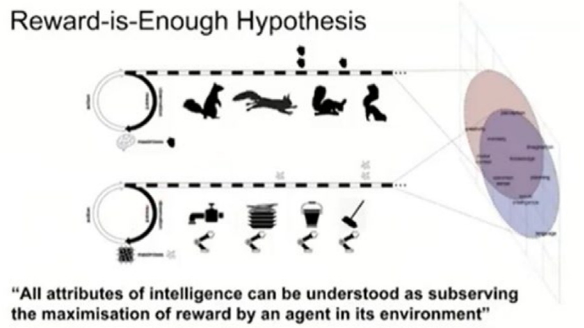
Reward-is-Enough Hypothesis
"All attributes of intelligence can be understood as subserving the maximisation of reward by an agent in its environment"

Fig. 6. Reward-is-Enough hypothesis.

Precup, 2020). While the standard RL provides a framework for learning a single policy to accomplish a single task, Generalized policy update learns novel policies for multiple tasks based on existing multiple policies. Concretely, in the policy evaluation phase, a standard RL evaluates the current policy for a single given task defined by a reward function. In generalized policy updates, multiple policies are evaluated from multiple reward viewpoints associated with multiple tasks in a phase called Generalized policy evaluation (GPE). On the other hand, in the policy improvement phase, only one policy is updated to select actions that improve the corresponding action value function in the standard RL. In Generalized policy update, for a given task, the action values associated with multiple policies are derived for a certain action. Then one policy is selected with the highest action value for that action. After selecting these policies for all the actions, the corresponding action values are compared, and a novel policy is derived to output an action with the highest action value for each state. This phase is called Generalized policy improvement (GPI) (Fig. 5).

Even though GPE and GPI provide systematic ways to derive novel policies for multiple tasks based on existing policies, these procedures are insufficient to achieve a data-efficient learning system because each existing policy needs to be re-evaluated based on the reward function of the target task. This process requires additional data samples and computations. Finding an efficient method to utilize the past experiences even for different tasks is necessary. To cope with this difficulty, the idea of using successor features was introduced. To share previous experiences among different tasks, reward functions are approximated by a combination of weight parameters and features. Then successor features are derived as expected accumulated feature values on the path probability for each policy. With this approximation, a previously acquired successor feature associated with a certain policy can be efficiently utilized to derive action value functions for a novel task by simply multiplying the weight parameters. Combining GPE with the successor representation is the key to achieving fast and computationally efficient reinforcement learning. Furthermore, the successor representation plays a very important role in the brain. This has been explored by a series of papers. Initially, the special case of the successor representation was proposed by Dayan (1993). Stachenfeld, Botvinick, and Gershman (2017) showed that the successor representation is linked to place cell in the hippocampus. Also, Momennejad et al. (2017) used successor representation to explain human responses to manipulations of rewards and transitions. This feature representation explains experimental results better than the standard model-free or model-based RL methods in the passive learning task to verify differential sensitivity to reward and transition revaluation. Therefore, the generalized policy update is not only a computationally and data-efficient RL method, but it can also possibly provide a systematic approach to the life-long learning problem that our brain is also trying to solve.

3.2. Deep reinforcement learning (by David Silver)

Consider the following hypothesis. An RL framework that just try to maximize reward in a continual cycle of action and observation "is" sufficient to yield all the different attributes of intelligence: perception, memory, imagination, creativity, motor control, knowledge, common sense, planning, social intelligence, and language. The question remains: How can a simple framework that optimizes a policy for one goal induces all these different attributes? Consider the process of maximizing a simple reward, such as a kitchen robot is maximizing the cleanliness of its kitchen. To achieve that goal in a very rich and complex environment, all kinds of attributes of intelligence are required. This fact leads to the following hypothesis: all attributes of intelligence can be understood as subserving an agent's maximization of reward in its environment. In other words, we only need one goal to achieve everything that we need by intelligence. This is a powerful hypothesis (Silver, Singh, Precup, & Sutton, 2021) (Fig. 6). Then, what about deep learning? Deep learning is a solution-side method with a universal framework for representing and learning functions. Whatever problem we are working on, assume that it can be specified by an objective. The idea of deep learning is to learn a function optimized for that objective in the end-to-end fashion. Deep learning uses a universal function approximator, which is a very rich class of neural networks that can represent any function with an arbitrary degree of accuracy. It also appears capable of universal accessibility to simply learn such things by a gradient descent of this objective.

The deep reinforcement learning (DRL) method integrates reinforcement and deep learning using neural networks as a function approximator and outputs, actions, values, and policies. The most famous successful application is AlphaGo (Silver, Huang, Maddison, Guez, et al., 2016), which defeated the human European Go champion, where exhaustive searches are infeasible due to the huge search space. AlphaGo maintains a value network that evaluates board positions and a policy network that implements strategies. The value network reduces the depth of searches by truncating search trees based on the value network's output, and the breadth of the search is effectively reduced by sampling subsequent moves based on the policy network's output. To train the networks, AlphaGo first uses supervised learning based on human playing and then reinforcement learning with self-play. In AlphaGo Zero (Silver, Schrittwieser, Simonyan, Antonoglou, et al., 2017), a successor to AlphaGo, a value and policy network was integrated into a single neural network and successfully avoided supervised learning of the initialization of human moves.

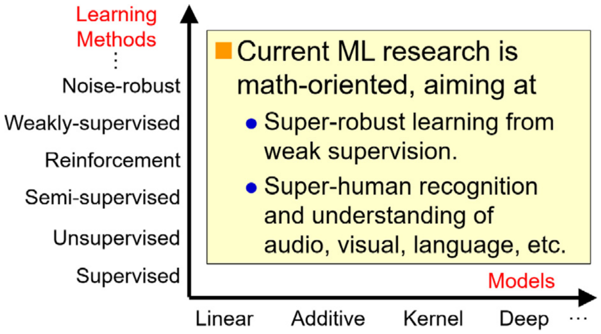
Learning Methods
⋮
Noise-robust
Weakly-supervised
Reinforcement
Semi-supervised
Unsupervised
Supervised
Current ML research is math-oriented, aiming at
Super-robust learning from weak supervision.
Super-human recognition and understanding of audio, visual, language, etc.
Models
Linear
Additive
Kernel
Deep
⋯

Fig. 7. Learning methods and models.

AlphaGo Zero completely vanquished AlphaGo: 100 games to 0. Then AlphaZero (Silver, Hubert, Schrittwieser, Antonoglou, et al., 2018), an extension of AlphaGo Zero, showed that Chess, Japanese Shogi, and Go can be played by a single algorithm and a single network architecture. These results suggest that AlphaZero is both applicable to other perfect games and a promising approach to solve real-world problems. The above algorithms are model-based approaches, in which the game rules are given in advance. MuZero (Schrittwieser, Antonoglou, Hubert, et al., 2020) learned a partial model essential to its decision-making processes. The learned model is used with AlphaZero's search algorithm and a search-based policy iteration algorithm.

While it has been argued that Go is a highly constrained game and DRL do not truly understand the game they are playing or transfer to other tasks (Mitchell, 2019), the AlphaGo series did achieve impressive performances.

4. Learning methods and models (by Masashi Sugiyama)

Although scientists is interested in making AI more intelligent, a major engineering challenge is how AI can be more useful for humans and contribute to human society. Towards human-like intelligence, various learning frameworks have been developed: supervised learning, unsupervised learning, semi-supervised learning (Sakai, du Plessis, Niu, & Sugiyama, 2017), reinforcement learning (Tangkaratt, Charoenphakdee, & Sugiyama, 2021), weakly-supervised learning (Kiryo, du Plessis, Niu, & Sugiyama, 2017; Lu, Niu, Menon, & Sugiyama, 2019), and noise-robust learning (Han et al., 2018; Ishida, Yamane, Sakai, Niu, & Sugiyama, 2020). Inspired by neuroscience studies, such learning algorithms as Hebbian learning and the backpropagation algorithm have been widely implemented. For learning models, simple linear models, more complex kernel models, and deep neural network models have been investigated. In deep neural network models, again inspired by neuroscience studies, convolutional neural networks, ReLU, LSTM, and attention mechanisms have been implemented (Fig. 7). Based on these innovations, AI has acquired sufficient high-level intelligence that can beat humans at the games of chess, go, and shogi. On the other hand, generating human-level agile movements or dexterous manipulation in real environments remains unsolved by the current AI since its sample efficiency lags far behind that of humans. Therefore, learning-from-humans (Osa et al., 2018; Schaal, 1999) or human-in-the-loop approaches (Ross, Gordon, & Bagnell, 2011; Teramae, Ishihara, Babič, Morimoto, & Oztop, 2018) remain very useful for motor learning problems.

5. Discussion

While AlphaGo series were quite successful, a huge amount of training data is required to reach super-human performance. The sample inefficiency of deep reinforcement learning is a significant drawback and precludes its application to many real-world problems. Next we discuss how to cope with this issue in the following sections.

5.1. World models and planning methods

The concept of using or learning the "world model" may not be very new. An approach using a simulation model to virtually generate physical interaction data has been widely adopted to pre-train a policy before applying it to the real environment, i.e., the sim-to-real strategy (Akkaya et al., 2019; Morimoto & Doya, 2001). However, unlike previous studies, recent "world model" learning methods claim that the dynamical models of environments can be acquired from image sequences (Ha & Schmidhuber, 2018) rather than from proprioceptive inputs such as joint or inertial measurement information. Furthermore, low-dimensional latent state representations can be also learned. For example, in the ATARI game scenario, MuZero (Schrittwieser et al., 2020) could learn the latent dynamical models of the games from observed image data and action sequences and successfully acquired game playing policies using the Monte-Carlo Tree Search (MTSC). Dreamer V2 (Hafner, Lillicrap, Norouzi, & Ba, 2021), a world-model-based RL method, was able to efficiently learn policies to play ATARI games with limited computational resources. Learning latent locally linear models also seems to be a useful approach for robot control (Finn et al., 2016; Karl, Soelch, Bayer, & Smagt, 2017; Watter, Springenberg, & Riedmiller, 2015; Zhang, Vikram, Smith, Abbeel, Johnson, & Levine, 2019).

On the other hand, if we have the model of the environment, for the robot control, we can use a planning method such as model predictive control (MPC). MPC-based algorithms have been applied to a variety of robots, e.g., mobile robots (Williams, Drews, Goldfain, Rehg, & Theodorou, 2018), drone robots (Bouffard, Aswani, & Tomlin, 2012), and humanoid robots (Ishihara, T.D. Itoh, & Morimoto, 2020; Tassa, Erez, & Todorov, 2012). However, especially for real robot control, we need rich computational resources for real-time planning by iteratively solving an optimal control problem.

5.2. Generating/reusing data

Self-play is a promising approach for cheaply generating training data without domain-specific knowledge. It has become always applicable. Experience replay (Lin, 1991), which is a common technique for improving sample efficiency in deep reinforcement learning (Mnih, Kavukcuoglu, et al., 2015), enables a learning agent to store and reuse past experiences. However, such off-policy reinforcement learning algorithms as Deep Q-Networks (Mnih et al., 2015), Deep Deterministic Policy Gradient (Lillicrap, Hunt, Pritzel, Heess, Erez, Tassa, Silver, & Wierstra, 2016), Soft Q-learning (Haarnoja, Tang, & Levine, 2017), and Soft Actor–Critic (SAC) (Haarnoja, Zhou, Abbeel & Levine, 2018) should be selected. For efficiently reusing previously acquired samples, importance-sampling-based methods have been proposed (Zhao, Hachiya, Tangkaratt, Morimoto, & Sugiyama, 2013) and applied to real humanoid robot control (Sugimoto et al., 2016). The idea of importance-sampling was also adopted in a sample-efficient RL framework for large-scale computing environments (Espeholt et al., 2018). Using a simulation model and updating the simulated environment is also a promising approach to efficiently use the data acquired from the real environment (Morimoto & Atkeson, 2009; Sugimoto & Morimoto, 2013).

5.3. Entropy regularization

Entropy-regularization is another important technique in modern deep reinforcement learning. It was originally used to improve exploration by seeking to maximize the entropy of the policy as well as total rewards, but it also improves sample efficiency. SAC considers the entropy of the policy, and Haarnoja, Zhou, Hartikainen, et al. (2018) evaluated SAC on locomotion for a quadrupedal robot and valve rotation with a 3-finger dexterous robotic hand. SAC achieved state-of-the-art performance for sample efficiency and asymptotic performance. Tsurumine, Cui, and Matsubara (2019) proposed deep dynamic policy programming with dueling architecture that incorporates Kullback–Leibler divergence between the current and learned policies and applied it to two real robotic cloth manipulation tasks with a dual-arm robot: turning over a handkerchief and folding a t-shirt. Since KL divergence limits overly large policy updates, it results in stable and efficient learning. Kozuno and Doya (2019) showed that RL algorithms with both entropy and KL divergence regularization share gap-increasing and softmax operator properties. These algorithms show both noise and error tolerance and avoid poor asymptotic performance. Recently, Vieillard, Kozuno, Scherrer, Pietquin, Munos, and Geist (2020) investigated the effect of entropy regularization and provided a performance bound that showed a linear dependency to the horizon length. Several modern algorithms, such as Soft Q-learning, SAC, Trust Region Policy Optimization (TRPO) (Schulman, Levine, Abbeel, Jordan, & Moritz, 2015), and Maximum a Posteriori policy Optimization (MPO) (Abdolmaleki et al., 2018), can also be unified under entropy regularization.

Imitation learning is a powerful framework for designing a policy from demonstrations. Although how it is implemented in our brain remains controversial (Charpentier, Iigaya, & O'Doherty, 2020; Collette, Pauli, & O'Doherty, 2017; Najar, Bonnet, & Palminteri, 2020), it is usually more sample efficient than vanilla reinforcement learning. Entropy regularization is critical for inverse reinforcement learning (Jeon et al., 2021; Uchibe, 2018; Ziebart, Maas, Bagnell, & Dey, 2008) to mitigate the issues of ambiguity and degeneracy. Recent imitation learning, inspired by generative adversarial networks and imitation learning (Uchibe & Doya, 2021). These studies show that efficient exploration can be achieved using the reward function estimated by inverse reinforcement learning if we can prepare expert's behaviors.

5.4. Hierarchical architecture/composite control

Although DRL methods have successfully generated a variety of human character motions in simulated environments (Heess et al., 2017; Merel et al., 2019), large-scale data are required as well as a many more motor learning trials than for humans. The human brain has a hierarchical structure, which might be a key architecture to allow humans to efficiently acquire policies for coping with complex motor control tasks (Gazzaniga, Ivry, & Mangun, 2008; Merel, Botvinick, & Wayne, 2019b). Therefore, hierarchical RL (HRL) is another promising approach to data-efficient learning as well as a way to implement the divide-and-conquer strategy (Dayan & Hinton, 1993; Morimoto & Doya, 2001; Sutton, Precup, & Singh, 1999). Another strategy to efficiently use previously derived controllers is to combine a set of learned policies to create a new policy that is applicable to a new task. Since the optimal value function is the solution of the nonlinear Bellman equation, a weighted linear summation of optimal value functions is not optimal for the Bellman equation in which the reward function is created by weighted linear summation. Todorov (2009a) proposed a framework of linearly solvable Markov Decision Process that makes the nonlinear Bellman equation linear. Then Todorov (2009b) developed the compositionality theory based on the property of the superposition of linear equations. Compositionality theory has been applied to control problems of character animation (Da Silva, Durand, & Popović, 2009), quadruped robot walking (Uchibe & Doya, 2014), a one-dimensional double-slit task (Matsubara, Asakura, & Sugimoto, 2015), and assisting human movements (Furukawa & Morimoto, 2021). When a set of learned policies is given for related tasks, compositionality theory creates a policy that performs a new task. A composite policy with additional learning was obtained much faster than learning an optimal policy from scratch. A compositionality theory was recently generalized by entropy-regularized reinforcement learning, and Haarnoja, Pong et al. (2018) described the relationship between a composite value function and a value function trained with a composite reward function.

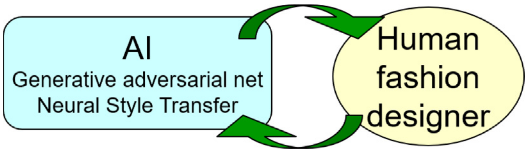

Fig. 8. Human-inclusive AI.

6. Future directions

6.1. Learning to learn

A deep learning framework allows AI agents to extract appropriate features for given tasks instead of using hand-designed features crafted by an experimenter and now AI has made a big leap ahead. As the next AI challenge, lifelong learning problems are getting much attention. In such problems, AI agents need to efficiently learn many policies to accomplish a wide variety of tasks for survival. Meta-learning methods have been developed to adapt a policy to different tasks only through a limited number of learning trials. To achieve different task goals, such meta-parameters as the learning rate or the discount factor also need to be adapted by the AI agent rather than using manually tuned parameters for specific tasks. Elfwing, Uchibe, Doya, and Christensen (2011) proposed an evolutionary computation approach to find meta-parameters for RL and described how the method works using cyber rodent robots. Recently, meta-gradient learning methods have been proposed that update these meta-parameters, cope with multiple tasks, or even find surrogate objectives (Finn, Abbeel, & Levine, 2017; Kirsch, Steenkiste, & Schmidhuber, 2020; Xu, van Hasselt, Hessel, Oh, Singh, & Silver, 2020). Recent studies are now developing a method to allow an AI agent to find learning rules by itself without explicitly providing the idea of value function which has been considered a mandatory component for RL (Oh et al., 2020). These new meta-RL studies that embrace the learning-to-learn principle (Lansdell & Kording, 2019) may motivate AI technologies to take the next huge leap and greatly impact society.

6.2. Human-inclusive AI

We must broach how human-like AI benefits society. Perhaps AI might be a different kind of entity. For creativity fields like arts and design where clear answers to target problems do not exist, providing new ideas that are born from different perspectives is critical. As a concrete example, a fashion designer, Ema Rie, corroborated with AI researchers at the University of Tokyo and RIKEN to produce new dress designs with an AI system (EMarie, AIP, & Tokyo, 2019). Ema provided dress designs to the AI system, which in turn proposed new ideas to her based on the first designs. Through this iteration, novel designs emerged. Such a human-inclusive AI direction might enrich human society (Fig. 8).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Preparation of this review article was supported by JSPS KAKENHI, Japan Grant Numbers JP16H06562, JP16H06565, JP19H05001. Maneesh Sahani was supported by the Gatsby Charitable Foundation, United Kingdom and the Simons Foundation, United States (SCGB 543039). Masashi Sugiyama was supported by the International Research Center for Neurointelligence, Japan (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., & Riedmiller, M. (2018). Maximum a posteriori policy optimization. In *Proc. of the 6th International conference on learning representations*.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., et al. (2019). Solving Rubik's cube with a robot hand. arXiv:1910.07113.

Asada, M., Uchibe, E., & Hosoda, K. (1999). Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development. *Artificial Intelligence*, *110*(2), 275–292.

Barreto, A., Hou, S., Borsa, D., Silver, D., & Precup, D. (2020). Fast reinforcement learning with generalized policy updates. In *Proceedings of the national academy of sciences*.

Bouffard, P., Aswani, A., & Tomlin, C. (2012). Learning-based model predictive control on a quadrotor: Onboard implementation and experimental results. In *2012 IEEE International conference on robotics and automation* (pp. 279–284). http://dx.doi.org/10.1109/ICRA.2012.6225035.

Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, *106*(4), 687–699.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv:2002.05709.

Collette, S., Pauli, W. M., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *ELife*, *6*.

Da Silva, M., Durand, F., & Popović, J. (2009). Linear Bellman combination for control of character animation. *ACM Transactions on Graphics*, *28*(3).

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Dayan, P., & Hinton, G. E. (1993). Feudal reinforcement learning. In S. Hanson, J. Cowan, & C. Giles (Eds.), *Advances in neural information processing systems, vol. 5*. Morgan-Kaufmann.

Douglas, R., Martin, A. C., K, & Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Computation*, *1*(4), 480–488.

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2011). Darwinian embodied evolution of the learning ability for survival. *Adaptive Behavior*, *19*(2), 101–120.

EMarie, AIP, RIKEN, & Tokyo, U. (2019). Fashion show held in u. Tokyo Komaba Campus on March 20.

Endo, G., Morimoto, J., Matsubara, T., Nakanishi, J., & Cheng, G. (2008). Learning CPG-based biped locomotion with a policy gradient method: application to a humanoid robot. *International Journal of Robotics Research*, *27*(2), 213–228.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., et al. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th international conference on machine learning, in PMLR 80* (pp. 1407–1416).

Felleman, D. J., & Fan Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, *1*(1), 1–47.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th international conference on machine learning, in PMLR, vol. 70* (pp. 1126–1135).

Finn, C., Yu, X., Duan, Y., Darrell, T., Levine, S., & Abbeel, P. (2016). Deep spatial autoencoders for visuomotor learning. In *Proc. of the 2016 IEEE Int. conf. on robotics and automation, May 16-21, Stockholm* (pp. 512–519).

Furukawa, J., & Morimoto, J. (2021). Composing an assistive control strategy based on linear bellman combination from estimated user's motor goal. *IEEE Robotics and Automation Letters*, *6*(2), 1051–1058.

Gazzaniga, M., Ivry, R. R., & Mangun, G. R. (2008). *Cognitive neuroscience: The biology of the mind*. MIT Press.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Ha, D., & Schmidhuber, J. (2018). World models. arXiv:1803.10122.

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., & Levine, S. (2018). Composable deep reinforcement learning for robotic manipulation. In *Proc. of IEEE International conference on robotics and automation*.

Haarnoja, T., Tang, H., & Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *Proc. of the 34th international conference on machine learning* (pp. 1352–1361).

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proc. of the 35th international conference on machine learning* (pp. 1861–1870).

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., et al. (2018). Soft actor-critic algorithms and applications. arXiv:1812.05905.

Hafner, D., Lillicrap, T. P., Norouzi, M., & Ba, J. (2021). Mastering atari with discrete world models. In *International conference on learning representations*.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., et al. (2018). Co-teaching: Robust training deep neural networks with extremely noisy labels. In *Advances in neural information processing systems, vol. 31* (pp. 8527–8537).

He, K., Fan, H., Wu, Y., S., Xie., & Girshick, R. (2019). Momentum contrast for unsupervised visual representation learning. arXiv:1911.05722.

Heess, N., T.B., D, Sriram, S., Lemmon, J., Merel, J., Wayne, G., et al. (2017). Emergence of locomotion behaviours in rich environments. Cs. http://arXiv.org/abs/1707.02286.

Henaff, M., Canziani, A., & LeCun, Y. (2019). Model-predictive policy learning with uncertainty regularization for driving in dense traffic. arXiv:1901.02705.

Ishida, T., Yamane, I., Sakai, T., Niu, G., & Sugiyama, M. (2020). Do we need zero training loss after achieving zero training error? In *Proceedings of 37th international conference on machine learning* (pp. 4604–4614). online, Jul. 13-18.

Ishihara, K., T.D. Itoh, D., & Morimoto, J. (2020). Full-body optimal control toward versatile and agile behaviors in a humanoid robot. *IEEE Robotics and Automation Letters*, *5*(1), 119–126. http://dx.doi.org/10.1109/LRA.2019.2947001.

Jeon, W., Su, C.-Y., Barde, P., Doan, T., Nowrouzezahrai, D., & Pineau, J. (2021). Regularized inverse reinforcement learning. In *Proc. of the 9th international conference on learning representations*.

Karl, M., Soelch, M., Bayer, J., & Smagt, P. (2017). Deep variational Bayes filters: Unsupervised learning of state space models from raw data. In *International conference on learning representations*.

Kirsch, L., Steenkiste, S. V., & Schmidhuber, J. (2020). Improving generalization in meta reinforcement learning using learned objectives. In *The international conference on learning representations*.

Kiryo, R., du Plessis, M. C., Niu, G., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems, vol. 30* (pp. 1674–1684).

Kozuno, T. E. Uchibe, & Doya, K. (2019). Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In *Proc. of the 22nd international conference on artificial intelligence and statistics* (pp. 2995–3003).

Lansdell, B. J., & Kording, K. P. (2019). Towards learning-to-learn. *Current Opinion in Behavioral Sciences*, *29*, 45–50.

LeCun, Y., Chopra, S., Hadsell, R., Marc'Aurelio, R., & Huang, F. (2006). *A tutorial on energy-based learning. predicting structured data*. MIT Press.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, *37*(4–5), 421–436.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2016). Continuous control with deep reinforcement learning. In *Proc. of the 4th international conference on learning representations*.

Lin, L.-J. (1991). Programming robots using reinforcement learning and teaching. In *Proc. of the 9th national conference on artificial intelligence (AAAI)* (pp. 781–786).

Lu, N., Niu, G., Menon, A. K., & Sugiyama, M. (2019). On the minimal supervision for training any binary classifier from only unlabeled data. In *Proceedings of seventh international conference on learning representations*.

Matsubara, T., Asakura, K., & Sugimoto, T. (2015). Dynamic linear bellman combination of optimal policies for solving new tasks. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E98.A*(10), 2187–2190.

Matsushima, T., Furuta, H., Matsuo, Y., Nachum, O., & Gu, S. (2020). Deployment-efficient reinforcement learning via model-based offline optimization. arXiv:2006.03647.

Merel, J., Botvinick, M., & Wayne, G. (2019b). Hierarchical motor control in mammals and machines. *Nature Communications*, *10*(1), 5489.

Merel, J., Hasenclever, L., Galashov, A., Ahuja, A., Pham, V., Wayne, G., et al. (2019). Neural probabilistic motor primitives for humanoid control. In *International conference on learning representations*.

Misra, I., & van der Maaten, L. (2019). Self-supervised learning of pretext-invariant representations. arXiv:1912.01991.

Mitchell, M. (2019). Artificial intelligence: A guide for thinking humans. In *Farrar straus & giroux*.

Mnih, V., Kavukcuoglu, K., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692.

Morimoto, J., & Atkeson, G. (2009). Nonparametric representation of an approximated poincare map for learning biped locomotion. *Autonomous Robots*, *27*(2), 131–144.

Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*(1).

Najar, A., Bonnet, E., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS Biology*, *18*(12), Article e3001028.

Oh, J., Hessel, M., Czarnecki, W. M., Xu, Z., van Hasselt, H. P., Singh, S., et al. (2020). Discovering reinforcement learning algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), *Advances in neural information processing systems, vol. 33* (pp. 1060–1070).

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., & Peters, J. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, *7*, 1–2.

Peters, J., & Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural Networks*, *21*(4), 682–697.

Ross, S., Gordon, G. J., & Bagnell, J. A. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*.

Russell, S., & Peter Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th Edition). Pearson.

Sahani, M., & Dayan, P. (2003). Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, *15*(10), 2255–2279, 2003 Oct.

Sakai, T., du Plessis, M. C., Niu, G., & Sugiyama, M. (2017). Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of 34th international conference on machine learning* (pp. 6–12).

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, *3*(6), 233–242.

Schrittwieser, J., Antonoglou, I., Hubert, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839) 604–609.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In *Proc. of the 32nd International Conference on Machine Learning* (pp. 1889–1897).

Silver, D., Huang, A., Maddison, C. J., Guez, A., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, Article 103535.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653.

Sugimoto, N., & Morimoto, J. (2013). Trajectory-model-based reinforcement learning: Application to bimanual humanoid motor learning with a closed-chain constraint. In *IEEE-RAS International conference on humanoid robots* (pp. 429–434).

Sugimoto, N., Tangkaratt, V., Wensveen, T., Zhao, T., Sugiyama, M., & Morimoto, J. (2016). Trial and error: Using previous experiences as simulation models in humanoid motor learning. *IEEE Robotics & Automation Magazine*, *23*(1), 96–105.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211.

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. arXiv:1611.01891.

Tangkaratt, V., Charoenphakdee, N., & Sugiyama, M. (2021). Robust imitation learning from noisy demonstrations. In *Proceedings of 24th international conference on artificial intelligence and statistics* (pp. 298–306). online, Apr. 13-15.

Tassa, Y., Erez, T., & Todorov, E. (2012). Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ international conference on intelligent robots and systems, vol. 2012* (pp. 4906–4913). http://dx.doi.org/10.1109/IROS.2012.6386025.

Teramae, T., Ishihara, K., Babič, J., Morimoto, J., & Oztop, E. (2018). Human-in-the-loop control and task learning for pneumatically actuated muscle based robots. *Frontiers in Neurorobotics*.

Todorov, E. (2009a). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(28), 11478–11483.

Todorov, E. (2009b). Compositionality of optimal control laws. *Advances in Neural Information Processing Systems*, *22*, 1856–1864.

Tsurumine, Y., Cui, Y., & Matsubara, T. (2019). Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation. *Robotics and Autonomous Systems*, *112*, 72–83.

Uchibe, E. (2018). Model-free deep inverse reinforcement learning by logistic regression. *Neural Processing Letters*, *47*(3), 891–905.

Uchibe, E., & Doya, K. (2014). Combining learned controllers to achieve new goals based on linearly solvable MDPs. In *Proc. of the IEEE International Conference on Robotics and Automation* (pp. 5252–5259).

Uchibe, E., & Doya, K. (2021). Forward and inverse reinforcement learning sharing network weights and hyperparameters. *Neural Networks*, *144*, 138–153.

Vértes, E., & Sahani, M. (2018). Flexible and accurate inference and learning for deep generative models. In *Advances in neural information processing systems, vol. 31*.

Vértes, E., & Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. In *Advances in neural information processing systems, vol. 32*.

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., & Geist, M. (2020). Leverage the average: an analysis of KL regularization in RL. In *Advances in neural information processing systems, vol. 33*.

Vincent, P., Larochell, H., Bengio, Y., & Manzagol, A. P. (2008). Extracting and composing robust features with denoising autoencoders. *2008*, In *ICML '08 Proceedings of the 25th international conference on machine learning* (pp. 1096–1103).

Watter, H., Springenberg, J. T., & Riedmiller, M. (2015). Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems, vol. 28*.

Wenliang, K. L., & Sahani, M. (2019). A neurally plausible model for online recognition and postdiction. In *Advances in neural information processing systems, vol. 32*.

Williams, G., Drews, P., Goldfain, B., Rehg, J. M., & Theodorou, Evangelos A. (2018). Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Transactions on Robotics*, *34*(6), 1603–1622.

Xu, Z., van Hasselt, H. P., Hessel, M., Oh, J., Singh, S., & Silver, D. (2020). Meta-gradient reinforcement learning with an objective discovered online. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems, Vol. 33* (pp. 15254–15264). Curran Associates, Inc..

Zhang, M., Vikram, S., Smith, L., Abbeel, P., Johnson, M., & Levine, S. (2019). SOLAR: deep structured representations for model-based reinforcement learning. In *Proceedings of the 36th international conference on machine learning*.

Zhao, T., Hachiya, H., Tangkaratt, V., Morimoto, J., & Sugiyama, M. (2013). Efficient sample reuse in policy gradients with parameter-based exploration. *Neural Computation*, *25*(6), 1512–1547.

Ziebart, B. D., Maas, A., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proc. of the 23rd AAAI conference on artificial intelligence* (pp. 1433–38).