In: Science for a Better Tomorrow: Curious 2024 Insights Actions, Springer 2025.

Towards Conscious Service Robots

Sven Behnke[0000−0002−5040−7525]

Abstract Deep learning's success in perception, natural language processing, etc. inspires hopes for advancements in autonomous robotics. However, real-world robotics face challenges like variability, high-dimensional state spaces, non-linear dependencies, and partial observability. A key issue is non-stationarity of robots, environments, and tasks, leading to performance drops with out-of-distribution data. Unlike current machine learning models, humans adapt quickly to changes and new tasks due to a cognitive architecture that enables systematic generalization and meta-cognition. Human brain's System 1 handles routine tasks unconsciously, while System 2 manages complex tasks consciously, facilitating flexible problem-solving and self-monitoring. For robots to achieve human-like learning and reasoning, they need to integrate causal models, working memory, planning, and metacognitive processing. By incorporating human cognition insights, the next generation of service robots will handle novel situations and monitor themselves to avoid risks and mitigate errors.

1 Introduction

Industries like car manufacturing impressively demonstrate the utility of robots. Recent developments in sensing, actuation, and – most importantly – artificial intelligence (AI) make it conceivable that robots will revolutionize many new application domains such as the flexible production of small lots, logistics, agriculture, security, inspection, professional services, and personal assistance. All of these domains, however, require, cognitive capabilities far beyond those of today's robots.

Current robotic systems rely on structuring the environments and tasks, e.g. by providing objects in well-defined locations. In open-ended settings, such as our ev-

Sven Behnke
Autonomous Intelligent Systems, Computer Science Institute VI – Intelligent Systems and Robotics, Center for Robotics and the Lamarr Institute for Machine Learning and Artificial Intelligence, University of Bonn, Germany. e-mail: behnke@cs.uni-bonn.de

eryday environments, robot-friendly structuring is impossible. Instead, autonomous service robots must instantiate models of their environment from sensor measurements, plan actions to achieve goals, carry out plans in the presence of disturbances, and monitor their execution. They also must familiarize with new objects and tools and need to improve their behavior through learning. Finally, they must communicate with persons in a human-understandable way, to receive instructions, answer questions, and explain their behavior.

The tremendous success of deep learning [64, 98] in visual perception, speech recognition, natural language processing, vision-language tasks, and multimodal tasks gives rise to hope that these methods will lead to revolutionary advances in autonomous robot performance.

2 Related Work

Personal service robots Personal robots that assist handicapped or elderly persons in their activities of daily living have attracted much attention in robotics research. An increasing number of research groups are working on robots for service applications. Examples include PR2 [74] that was used in a household marathon experiment [52], Everyday Robots' mobile manipulator [42], and Toyota Human Support Robot [125], a standard platform in the international RoboCup@Home competitions. My team NimbRo won these competitions 2024 with two PAL Robotics TIAGo++ robots [75] and 2011–2013 with our cognitive service robot Cosero [105, 106], demonstrating a large variety of domestic service tasks. Further examples include Care-O-Bot 4 [55], Armar-6/7 [2], HRP-5P [59], TORO [53], E2-DR [127], and our Centauro robot [56, 57] which demonstrated challenging locomotion and manipulation tasks including the use of tools. For the ANA Avatar XPRIZE competition [8,40], capable systems have been developed, including iCub3 [20], Pollen Robotics Reachy, and our winning Avatar system [66, 99]. Scene perception. In order to act in complex indoor environments, service robots must perceive the room structure, obstacles, persons, objects, etc. To this end, they are equipped with cameras and depth sensors. Estimating the sensor poses and registering the measurements yields environment maps [87]. In addition to modeling the environment geometry and appearance, semantic perception is needed [95].

Deep learning For pattern recognition, deep learning [64] methods are extremely successful. They revolutionized visual perception [23, 92], speech recognition [88, 91], natural language processing [1, 85], vision-language tasks [69, 107], and multimodal tasks [33, 39]. Supervised deep learning requires large annotated data sets like ImageNet-21K [25], JFT-3B [129], and Kinetics [14], though, which are expensive to obtain. To address variability that should not change output, data augmentation methods such as image transformations [101] and generative models [128] are used to generate variants of training examples. For robotic tasks such as mobile manipulation, large-scale annotated datasets do not exist. To avoid the need

for large labeled datasets, much research focuses on methods that can adapt to new conditions through transfer learning and domain adaptation. Transfer learning [58] uses representations learned from large data to learn a related task from small data, e.g. by continuation of training. Semi-supervised, weakly supervised, and unsupervised learning methods use fewer, low quality, and no labels at all, respectively. One example of semi-supervised methods is the student-teacher approach [124], where a teacher is trained on a small labeled data set and then generates pseudo labels for a large unlabeled data set to train the student. Because unlabeled data is much easier to obtain than annotated data, unsupervised methods are often used to pre-train models [10, 16]. The hope is to discover useful structure in the data which might aid target tasks. A promising subclass of unsupervised learning is self-supervised learning [37], which requires only unlabeled data to formulate a pretext task, for which a target objective can be computed without supervision. These pretext tasks must be designed in a way that high-level data understanding is useful for solving them, e.g. prediction of occluded image parts [41] or future video frames. As a result, the intermediate layers of trained models encode high-level semantic representations that are useful for solving downstream tasks. One form of self-supervision is contrastive learning [15], where two different data augmentations are applied to an image and a model is trained to maximize agreement between the outputs and minimize agreement with outputs for other images. Contrastive learning of dense descriptors for object surface elements has been applied to learn visuomotor manipulation policies [28] that generalize within a category of objects and are able to handle deformable objects. Other possibilities for self-supervised learning are to maximize mutual information between input and model output [45] and joint embeddings of two inputs with variance-invariance-covariance regularization [3].

Large language models Self-supervised training is the basis for the impressive performance of recent large language models (LLMs) such as GPT-4 [85] and Palm 2 [1] that continue text in plausible ways. Their large transformer networks [114] were trained on massive data to predict the next token. In contrast to recurrent sequence models, transformers flexibly re-route and combine information from relevant parts of the sequence through learned self-attention, which is implemented using content-based access to information values by matching keys to queries. Recently, autoregressively trained LLMs have shown sparks of artificial general intelligence [13]. Such models can acquire human-like systematic generalization through meta-learning [61], but this requires generating a training set of systematic generalization example problems. On the other hand, LLMs often lack common sense, hallucinate facts, fail at arithmetic, have difficulty reasoning, and cannot make proper plans. For these reasons, LLMs are combined with external tools [89] such as search engines, calculators, planners [68], etc. Multimodal models such as PaLM-E [26] and ImageBind [33] combine text and other modalities such as images in joint embeddings. Generative models are not restricted to producing text but are also used to generate, e.g., images [86] and video [90] from text inputs.

3D models To address the 3D nature of scenes, Neural Radiance Fields (NeRF) have been proposed, which learn a neural network mapping 5D coordinates to

density and color by predicting images from multiple views through volumetric rendering [77]. In our recent work PermutoSDF [93], the 3D shape of objects is represented by a neural signed distance function (SDF). By modeling individual objects as permutation-invariant slots, object representations can be learned through novel-view synthesis [94]. If conditioned on latent variables, category-level shape spaces can be learned, e.g., for articulated human bodies [24]. Compositional generative scene models that represent objects and their relations can be learned without image-level supervision [31].

Scene prediction In dynamic scenes, the motion of objects and persons must be estimated and predicted. Scenes with moving agents (e.g., humans or robots) can be represented with 3D dynamic scene graphs [47]. Motion is the strongest cue for perceptual grouping and predictive models are widely used to explain human visual perception [30]. Consequently, optimization of a prediction loss can be used to segment moving objects in videos [119]. SlotFormer [121] models spatio-temporal object relationships and predicts object states. Our recent work on object-centric video prediction decouples the processing of temporal dynamics and object interactions [116]. This facilitates learning of tasks that require understanding of object relations [81]. A fundamental problem when predicting the future is that often multiple plausible futures exist. MultiPath++ [113] predicts a distribution of future trajectories of road users parameterized as a Gaussian Mixture Model (GMM). Multiverse [67] predicts the distribution over multiple possible future paths of persons using convolutional recurrent neural networks (RNNs) over graphs. World modeling. Prediction of future scene states and planning own actions require world models that are conditioned on actions. Playable Video Generation [76] learns a discrete set of actions from unlabeled video that are used to interactively generate video from actions. This task has been extended to Playable Environments [76] that can control multiple objects in 3D scenes with action labels that are discovered in an unsupervised way. DayDreamer [120] learns action-conditioned forward models in a latent space for multiple robots. GenAD [126] and GameNGen [112] are world models for autonomous driving and a video game, respectively.

Deep reinforcement learning Reinforcement learning (RL) addresses the development of situated agents that learn how to behave while interacting with the environment [108]. This problem is formulated as an agent-centric optimization in which the objective is to select actions based on the estimated state in order to obtain as much reward from the environment as possible in the long run. Impressive success has been achieved by combining this approach with deep learning. One example is MuZero [97] which combines tree-based search with a learned model and achieves superhuman performance in a range of challenging and visually complex domains (Atari games, Go, chess, and shogi), without any prior knowledge of their underlying dynamics. MuZero learns a model that predicts the quantities relevant to action planning: the reward, the action-selection policy, and the value function. AlphaStar [117] learned the multi-agent game StarCraft II from ~500K human games and 120M self-played games. Gran Turismo Sophy [123] learned from carefully engineered state and reward in more than five years of simulated driving hours to compete with

the world's best drivers. Playing soccer with humanoid agents was learned from decades of match simulations [70]. Soccer skills for a humanoid robot and 1v1 play were learned from 2.5 years of simulated experience [38].

These numbers indicate that it would be impractical to collect that much experience with a real robotic system. Consequently, real-robot reinforcement learning mostly focuses on individual skills. For example, Google X learned grasping from cluttered bins with a simple manipulator under closed-loop monocular vision-based control [50]. They operated seven experimental setups for four months to collect 580K real-world grasp attempts to train a deep neural network Q-function with over 1.2M parameters and report a 96% grasp success rate on unseen objects. The method learned regrasping strategies, probing or repositioning objects to find the most effective grasps, performing other non-prehensile pre-grasp manipulations, and responding dynamically to disturbances and perturbations. Sorting recyclables and trash was learned from simulation and 9,527 hours of real-robot experience obtained with a fleet of 23 mobile manipulators [42].

Real-robot RL needs suitable inductive biases [43] to learn from little experience. These biases represent domain knowledge and can take many forms, e.g., the structure of the agent-environment interface and the policy generation mechanism. To improve the data efficiency of RL, transfer learning has been investigated. By pre-training on RoboNet [21], a data set providing 15M video frames from seven different robot platforms, and fine-tuning on a held-out target platform, it has been demonstrated that simple manipulation tasks such as pushing and pick-and-place can be learned from limited experience. Multi-task learning amortizes experience over multiple tasks [51]. It generalizes to structurally similar tasks and acquires distinct new tasks more quickly. One way to address the combinatorial complexity of multi-object scenes is to factorize them into objects. Object-centric perception, prediction, and planning [115] learns to discover objects in visual scenes and models their dynamics and appearance without supervision. A model-based reinforcement learner that predicts and plans block stacking on this abstract level generalizes to novel configurations and more objects. Action Schema Networks [110] learn generalized policies for probabilistic planning problems. By mimicking the relational structure of planning problems, they generalize over all instances of a given planning domain. Manipulation inherently involves contact and often requires both haptic and visual feedback. Lee et al. [65] use self-supervision to learn a compact and multimodal representation of sensory inputs, which is then used to improve the sample efficiency of policy learning as demonstrated for peg insertion. To avoid random exploration, imitation of human experts can be used. RT-1 [12] is a transformer-based controller trained on 130K demonstrations of a large variety of pick and place tasks in kitchen environments. Open X-Embodiment [84] is a large data set of camera images and end-effector movements from 22 different robots, demonstrating 527 skills (160,266 tasks). The RT-X model trained on this data exhibits positive transfer and improves the capabilities of multiple robots by leveraging experience from other platforms. 970k episodes from this data set were used to train OpenVLA [54], starting from a large language model and a visual encoder. OpenVLA demonstrates generalist manipulation capabilities and can be adapted to new robots via fine-tuning.

3 Challenges

Despite much research and progress, capable mobile manipulation robots that can cope with the complexity of open-ended real-world applications have not yet been realized. Developing such robots is a tremendous challenge, due to the typical characteristics of these applications.

Many sources of variability There are many sources of variability that a mobile manipulation robot must cope with. These include varying shape, texture, and physical properties of objects – even within a category. Furthermore, the 6D object pose, speed, and articulation state may vary. Environmental conditions, such as lighting, and surface properties, such as shininess, transparency, texturelessness, or non-reflectivity greatly impact appearance in camera images and consequently the completeness and precision of depth estimates. The variability of single objects is exponentiated by the infinite possibilities for multi-object arrangements. Similarly, the robot environments such as rooms and apartments vary greatly in layout, geometry, surface properties, and other factors. The manipulation and locomotion tasks that capable robots need to perform are highly variable as well. Hence, learning methods are needed that generalize to novel, unseen situations.

High-dimensional state and action spaces Input and output of mobile manipulation robot controllers are high-dimensional. Typical camera images are of size 1920×1080×3, already more than 6M dimensions. Depth cameras, 3D LiDARs, force-torque & haptic, inertial, and joint sensors add many more input dimensions. The sensors measure at high rates, e.g., at 30 Hz, producing hundreds of million measurements per second. The output dimensionality is high as well, with typically more than 50 DoF for anthropomorphic robots. These joints need to be controlled at high rates with target positions, velocities, or torques. Hence, learning methods are needed that can cope with high-dimensional state and action spaces.

Hybrid discrete-continuous variables Some variables, such as the presence of objects or the task category, are discrete while other variables, such as 6D object poses or task parameters, are continuous. This creates the need for learning methods that can cope with both discrete and continuous variables.

Non-linear dependencies Objects are typically in contact with support surfaces and with each other and the robot must make and release contacts with its end-effectors or other body parts to manipulate them. This induces highly non-linear constraints. While objects may be easily moved away from the contact point, moving them further towards the colliding surface is not possible. Similarly, occlusion effects and the transition between stick friction and sliding are highly non-linear. Learning methods must address such non-linearities.

Stochasticity There is much randomness in the world. Unmodeled environmental factors and other agents might also be perceived as non-determinism from our robot's point of view. Furthermore, robot sensors are noisy and unreliable; and robot actuators are imperfect and induce stochasticity. Hence, the state must be estimated

from unreliable observations and predictions are hard to make and become more and more uncertain for larger time horizons. Learning methods must cope with such uncertainties.

Partial observability Due to the projection of the 3D world onto 2D cameras and other sensors, limited sensor ranges, resolutions, accuracies, etc., not all state variables that would be needed for action planning are directly accessible. Hence, learning methods must consider the distribution of possible states and must generate actions to acquire more information, for example changing the camera pose to see occluded objects, touching objects to sense physical properties like weight and stiffness, and opening containers to see what is inside.

Underactuation Robots have limited action capabilities to influence the state of the environment. Their drives have limited speed and acceleration, their manipulators have limited reach, strength, and dexterity. Some environmental variables cannot be influenced directly, but only through indirect means like tools. Learning methods must respect these constraints and generate behavior such as improvised tool use to overcome them.

Multimodality Mobile manipulation involves multiple modalities, such as vision, distance measurements, forces, and haptics on the input, and also multiple outputs, such as mobility, manipulation, and active sensing. Hence, learning methods must jointly address these modalities and come up, for example, with grasping strategies that transition smoothly from vision-based scene understanding and grasp selection, to visual tracking and correction of the approaching motion, to grasping execution and re-grasping based on haptic feedback.

Non-stationarity One unique challenge is the non-stationarity of robots and their environments. Not only do robot bodies change due to wear and tear, also the open-ended environments in which they operate and the tasks they perform are constantly changing. Already the ancient philosopher Heraclitus noted that the only constant in life is change. Such changes violate the fundamental assumption underlying current machine learning that a learned model will be used on the same distribution of data it has been trained on. When using a trained model on a different distribution (out-of-distribution, OOD), one cannot expect good performance [80]. In fact, seemingly small changes can lead to catastrophic failure [17].

4 Human Cognitive Functions

Humans are able to cope with such changes and quickly learn new tasks. My hypothesis is that the cognitive architecture of the human mind has evolved to continuously interact with changing environments and that equipping robots with key elements of this architecture will enable flexible handling of OOD data and *systematic generalization*. Systematic generalization was first studied in linguistics [5,60] because it is a core property of language that meaning for a novel composition of existing concepts

(e.g. words) can be derived systematically from the meaning of the composed concepts and the way they are composed. Humans exhibit systematic generalization also when understanding a new object by combining properties or parts which compose it [62]. *Compositionality* is the principle that complex objects can be described by their constituent parts and their relations to each other [29]. It allows to generate infinite variants from a finite set of building blocks, enables open-world zero-shot learning [71], and even makes it possible to generalize to new combinations that have zero probability under the training distribution.

While humans perform many routine tasks like walking or riding a bike without much attention, object manipulation, communication, and handling novelty are different. Cognitive science distinguishes *habitual* and *controlled* processing [11]. Habitual processing effortlessly generates default behaviors that are performed routinely. In contrast, controlled processing requires attention and mental effort to generate non-routine behaviors. Kahneman [49] introduced the framework of fast and slow thinking and corresponding processing systems in our brain (see Fig. 1). Routine, habitual tasks can be achieved quickly in parallel without *conscious* attention using only System 1 abilities, whereas more complex tasks also require System 2 that is more capable but slower, serial and involves conscious processing. System 2 uses explicit, verbalizable knowledge and explicit processing while System 1 relies on implicit, non-verbalizable, intuitive knowledge. We can act in fast and precise habitual ways without having to think consciously, but the reverse is not true: conscious processing builds on the unconscious System 1.

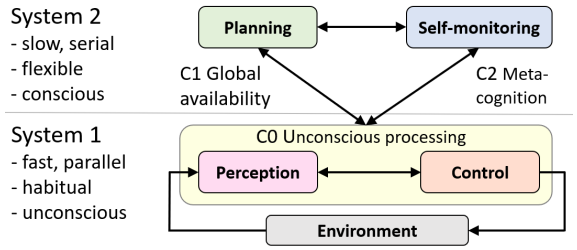

Fig. 1 Human cognitive functions according to Kahneman [49] (System 1, System 2) and Dehaene et al. [22] (C0, C1, C2).

System 2 is very flexible and powerful. It allows to solve novel problems creatively by recombining existing pieces of knowledge, to discover and use causal dependencies, to imagine future outcomes, to plan actions, to find explanations, to reason, etc. It is also at that level that we communicate with others through natural language, e.g. to receive task specifications or new knowledge and rules that we can apply immediately. The capacity of System 2 is very limited, though. Our working memory can only hold 3-5 meaningful items active simultaneously [19]. Baars introduced the *Global Workspace Theory* [4] that identifies conscious processing as the communication bottleneck between selected parts of the brain that are called upon when addressing a current task. There is a threshold of relevance beyond which information that was previously handled unconsciously gains access to this bottleneck and is instantiated in working memory. When this happens, the information is broadcast throughout the brain, allowing its various relevant parts to synchronize and choose configurations and interpretations of their piece of information that are globally coherent with the configurations chosen in other parts of the brain.

While this severe communication bottleneck might appear to be a weakness, it can also be advantageous. Firstly, there is pressure to combine multiple lower-level items that frequently occur together to larger, composite items, facilitating abstracting away irrelevant detail and providing compositionality. Secondly, when focusing on a few relevant items of a scene for conscious planning, we essentially ignore all other items, which are irrelevant for the task at hand. This leads to systematic generalization, because we can reuse the task knowledge in infinitely many novel situations in which the irrelevant items change. System 2 is slow, has limited capacity, and involves conscious effort; hence, there is pressure to migrate tasks to System 1 wherever possible. Through rehearsal, frequently performed tasks become habitual.

Dehaene et al. [22] characterize consciousness further. They distinguish unconscious processing (C0) and two orthogonal dimensions of conscious computations: *global availability of information (C1)* and *meta-cognition (C2)*.

C1 – global availability – is a consequence of the distributed organization of the brain as a deep hierarchy of specialized subsystems that must be synchronized and of the need to act, which means that we cannot stick to a diversity of probabilistic interpretations and action options, but must decide in favor of a single course of action. Such decision-making requires efficient pooling over all available sources of information, considering the available action options and selecting the best one, sticking to this choice over time, and coordinating internal and external processes towards the achievement of that subgoal. Attention – selective processing of information – is crucial for items entering consciousness, but attention is not limited to conscious processing. Unconscious C0 processing also includes bottom-up and top-down attentional mechanisms, which operate in parallel to prioritize and flexibly route information – often without bringing it to consciousness. The hierarchical system of sieves that operate unconsciously computes probability distributions, but only a single sample drawn from these becomes conscious at a given time, making it available globally to all specialized modules. Alternative interpretations might become conscious at other points in time, thus, C1 consciousness is causally responsible for our serial information-processing bottleneck. Attention also implements variable binding [36]. The association of information elements to roles in relations and rules is crucial for applying these templates to varying input and, hence, for multi-step inference and systematic generalization.

Consciousness in the second sense (C2) is characterized by the ability to reflexively represent oneself. When making decisions, we feel more or less confident about our choices. Our brain does not only make perception and action decisions, but also estimates its degree of confidence. State estimation and learning also rely on confidence, for example, we weigh existing knowledge versus new evidence, like a Kalman filter [109]. Error detection is another example of self-monitoring: just after responding, we sometimes realize that we made an error and change our minds. This might be explained by further evidence that arrived after the decision or by slower C2-processing monitoring fast C0 sensory-motor execution. We don't just have knowledge, but we also know what we don't know. Such meta-knowledge is crucial for assessing our limits and for learning.

C1 and C2 are largely orthogonal and complementary dimensions of consciousness. Their joint possession may have synergistic benefits to organisms and robots. Bringing probabilistic metacognitive C2 information into the global C1 workspace allows it to be held over time, integrated into explicit long-term reflection, and shared with others. On the other hand, the possession of an explicit repertoire of one's own abilities (C2) improves the efficiency with which C1 information is processed.

5 The Need for Conscious Robots

Despite tremendous progress in C0-like deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, truly human-like learning and thinking machines will need to go beyond current engineering trends in both what they learn and how they learn it [62]. They need to build causal models of their environment that support explanation and understanding and must harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations [96]. For this, equipping machines with C1 and C2 conscious processing will be crucial. Upon success, they would behave as if they were conscious; e.g., they would know that they are seeing something, express confidence in it, report it to others, and may even experience the same perceptual illusions as humans.

Of course, *traditional symbolic AI* systems (GOFAI), like Hierarchical Task Network Planners [82] and CRAM [7], exhibit some of the properties that are associated with conscious System 2 processing, like compositionality. However, such symbol manipulation systems often lack semantic grounding of the higher-level concepts in terms of the lower-level observations and actions. Whereas pure symbolic representations put every symbol at the same distance from every other symbol, learned embeddings represent concepts through a vector of attributes – with related concepts being close-by and interpolations being meaningful. GOFAI systems often are too rigid to account for real-world data with outliers, etc. Further, GOFAI search and inference are generally intractable and need to be approximated. Here, learning representations together with inference procedures is needed to generate fast habitual C0 behavior. Finally, GOFAI approaches often do not handle uncertainty, which is crucial for partially observable, stochastic environments.

In recent years, *neuro-symbolic* approaches [32, 44, 73] have been proposed that integrate symbolic and subsymbolic representations, inference, and learning. However, hybrid neuro-symbolic systems [18, 34, 63, 72, 83, 100, 102, 111, 130] inherently use different representations and tools for neural and symbolic computations, which are difficult to integrate tightly.

Neurocompositional computing [103] is based on the principles of compositionality and continuity. It encodes structures in vectors that are processed by neural networks and shows promising results by quickly learning tasks from small data sets that require systematic generalization. The Differentiable Tree Machine [104] compiles high-level symbolic tree operations into subsymbolic matrix operations on

tensors. Here, an agent learns to sequentially select tree operations to execute tree transformations with the help of a tree memory.

Recently, autoregressively trained embodied multimodal models have been used for generating robotic skills such as grasping and placing objects [12, 54] and for higher levels of robot control [26]. These models lack System 2 conscious processing, though. They need much data [84] and computing power; and the addressed scenarios are still relatively simple.

My hypothesis is that using insights from human cognition for the cognitive architectures of robots by incorporating C1 global availability and C2 metacognition will enable the next level of robot capabilities.

Because it extends highly successful C0 processing without a change in tools, I am convinced that a bottom-up way towards consciousness-inspired higher-level cognitive functions for service robots is the way to go.

6 Objectives

My overall goal is to develop methods for learning higher-level cognitive functions for service robots, which go beyond unconscious routine tasks by incorporating conscious processing to cope with novel situations and self-monitor.

Unconscious perception and control The System 1 / C0 routine processing directly interacts with the environment and is hence the basis for any higher-level cognitive functions.

Starting from raw sensory measurements, such as video, depth, forces, and haptics, structured representations of mobile manipulation robot workspaces shall be learned on multiple levels of spatio-temporal abstraction. Abstraction will be realized by coarser spatio-temporal scales and more expressive, sparser representations on the higher levels. The elements of these representations will correspond to increasingly larger entities (parts, objects, groups of objects) in the scene and will be increasingly semantic. The learned representations shall transition from sensor coordinate systems (e.g. the camera frame) to 3D representations of the scene and joint multimodal embeddings.

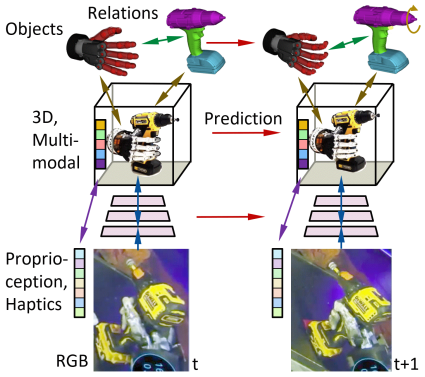

Fig. 2 Scene perception and prediction on three levels: in the sensor coordinate frame (bottom), in 3D multimodal embeddings (center), and with objects and their relations (top).

They shall model individual objects and the robot end-effectors in their own canonical frame. This will enable learning of category-level shape and appearance spaces within a hierarchical categorization. Scene parsing shall instantiate these

models and estimate object parameters, such as pose, shape descriptors, and appearance descriptors. Predictive models for these scene representations shall be learned on all levels of abstraction (see Fig. 2). Prediction of low-level detail shall be done only for short time horizons. Higher-layer representations shall be predicted with coarser temporal granularity over longer time horizons. These predictions shall be based on individual object dynamics models and on pairwise relational models to account for object interactions, such as contact. The graph of object relations shall be sparsely instantiated according to the relevant object interactions. The predicted representations shall be compared to the feed-forward interpretation of new measurements, such that prediction error can be used to update the representations on all levels.

The learned predictive models shall be extended by conditioning them on robot actions. This will allow for the rollout of possible futures of robot-environment interaction. Coarse-to-fine model-predictive control of routine skills that do not require conscious attention shall be learned from imagined rollouts on the multiple levels of abstraction. Higher layers shall plan abstract actions longer into the future, which are concretized on lower levels for shorter time horizons. A large variety of skills shall be learned for modular behaviors that activate coarse-to-fine actors according to the situation. Binding objects or places to roles shall yield parametrizable skills, such as grasping or placing an object or navigating towards a waypoint while avoiding obstacles.

Conscious prediction and planning Methods for selecting a small set of elements from the highest-level C0 representations and for maintaining them in a working memory (WM) shall be learned. This WM will be the basis for learning action-conditioned predictions, based on binding selected elements to variables of applicable rules.

Structured predictions of state transition rewards, value, and action selection probabilities shall be learned from interactions with simulated and real environments. LLMs shall be incorporated as oracles.

As illustrated in Fig. 3, the learned WM world models shall be used for efficient action planning by sequential search. A spatio-temporal action abstraction shall be learned, such that subplans are reused in different contexts. The learned models shall be used for autonomous operation of mobile manipulation in complex novel situations.

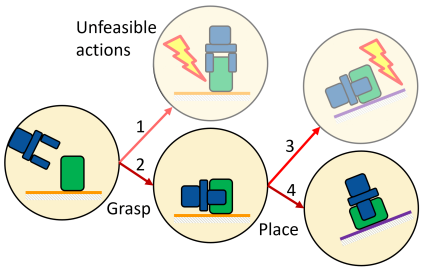

Fig. 3 Conscious planning. The WM state is rolled out using actions 1–4. Actions 1 and 3 are unfeasible (unreachable top grasp and unstable placement, respectively).

Conscious self-monitoring Methods for assessing the confidence of perceptions and predictions shall be developed. They shall be based on learning the distributions of latent variables on which multiple plausible futures can be conditioned (see Fig. 4).

By sampling from these variables, a tree-manifold of state-action rollouts can be generated, such that not only average value, but also its variance and worst-case return can be estimated. These quantities shall be incorporated into perception and action selection, to obtain policies that collect more information when needed and avoid dangers. Furthermore, the execution of low-level skills shall be monitored by comparing the current percept to expected outcomes to detect errors and to mitigate them.

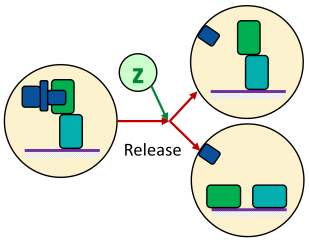

Fig. 4 Predicting multiple plausible futures conditioned on latent variable z.

7 Methodology

My approach will be to add suitable *inductive biases* to deep reinforcement learning (DRL), such that structured representations and conscious processing are enforced, which will enable systematic generalization and self-monitoring. Inductive biases reflect assumptions about the statistics of modeled scenes and robot-environment interactions and are necessary for generalization [78]. For instance, hierarchical convolutional neural networks (CNNs) [64] hardwire local dependencies, translation equivariance, hierarchical structure, and invariance to local deformations; whereas recurrent neural networks [46] exploit equivariance over time.

Further biases are needed for higher cognitive functions. One example of these is choosing the *appropriate frame of reference* for modeling. Commonly, deep neural networks represent visual scenes in a sensor coordinate system. Describing objects in object-centered canonical frames normalizes away the variability induced by the 6D object pose [118]. In such object-centered frames, shape and appearance spaces can be learned much easier. Of course, such canonical frames are also useful for individual parts of objects, for which the 6D pose relative to the object-centered frame must be modeled. The *projection of the 3D world to 2D images* induces occlusions and discontinuities at object boundaries. Modeling the scene in 2.5D by individual depth layers or directly in 3D allows for more complete, continuous representations where occluded parts are present and hence unoccluded parts can be predicted. Much of the image motion can be explained by *camera motion*. Hence, approaches that explicitly model the projection from 3D to the variable camera view can represent this dependency compactly. One particularly powerful assumption is *relational inductive bias* [6]. It expresses the observation that scenes can often be described in terms of entities (objects, parts, groups) and their sparse pairwise interactions (relations). Relations describe interactions between entities on adjacent levels of abstraction, e.g. between the whole object and its parts, which represents the *compositional structure* of the world. Such compositional hierarchy can also be found on the action side, where tasks are composed of multiple subtasks and subtasks are composed of individual skills (see Fig. 5).

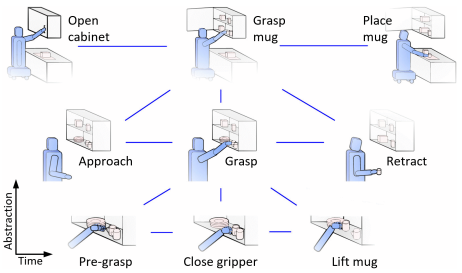

Fig. 5 Modeling actions on multiple levels of abstraction.

Relations are also present between closely interacting entities on the same level of description, e.g. for objects that are in contact or for adjacent subtasks. In *hierarchical categorization*, objects or actions are categorized on multiple granularities [25]. This allows for pooling instances of multiple finer categories to learn models of coarser categories. *Planning in the now* [48] refers to the assumption that typical planning problems are not like mazes but can be solved by plans that consist of only a few steps which are described on a detailed, concrete level for the immediate future and on coarser, more abstract levels for the more distant future. This makes planning exponentially more efficient than detailed long-horizon planning.

Crucial is the *consciousness prior* proposed by Bengio [9]. It assumes that from subconscious, massively parallel computed representations, a small subset of elements is selected by attention mechanisms for sequential processing. This corresponds to a sparse factor graph on a conceptual, symbolic level, which affords abstract reasoning. Generic factors are probabilistic analogs of logical rules with quantifiers, i.e., with variables or arguments that can be bound [35]. On this level, graph neural networks [122] are applicable, which exploit *equivariance over entities and relations*.

Self-attention used in transformer networks [114] provides flexible information routing and learns sparse features with the sample complexity scaling only logarithmically with the context size [27]. High-level representations that describe verbalizable concepts as semantic variables that play a causal role can be encouraged with an *inductive bias towards words* [35]. When planning is restricted to maintaining a single state consisting of few elements and search is serial and can consist of few action-conditioned WM state predictions only, there is *pressure to aggregate* elements to higher-level entities by discovering new concepts and macro actions.

Inductive biases alone will not suffice to address complex, open-ended real-world domains, because real-robot experience is expensive to obtain and cannot be collected in large enough quantities. Fortunately, *foundation models* for vision [23, 92], language [1, 85], and multimodal data [33, 39] are available. They have been trained on Internet-scale data and summarize much more experience than real robots could make. Incorporating this knowledge through distillation will be a crucial factor for success. Another source of knowledge that I will incorporate is *human guidance and demonstrations*. Still, it will be necessary to develop methods first in a *photorealistic physics-based simulation* [79], where experience can be made cheaply in large quantities without endangering the real robot, before transferring them to the real world.

I will not approach human-like mobile manipulation in its full generality in a single step, but will increase the level of difficulty gradually along the following

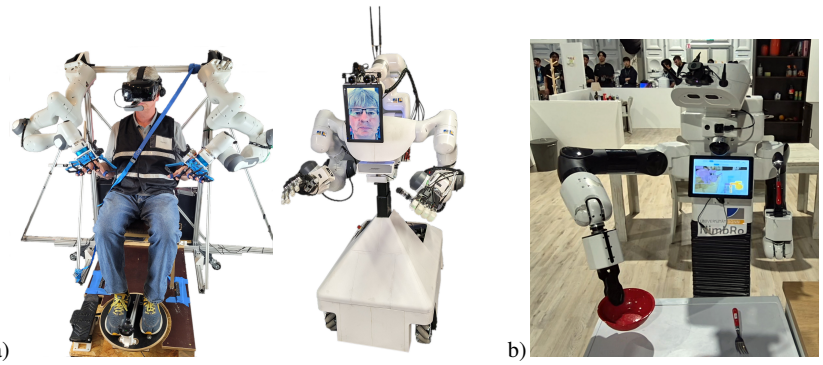

Fig. 6 a) NimbRo Avatar system [66]; b) NimbRo@Home robot [75].

dimensions: *Number of objects:* Starting without an object (learning a self-model), proceeding with a single object (grasping, placing, pushing), and finally considering the manipulation of two objects and the use of tools. *DoF of the robot:* From single-handed object manipulation over bimanual tasks to mobile manipulation tasks. *DoF of objects:* From rigid objects, to articulated objects, to deformable objects. *Feedback modalities:* Starting with RGB-D camera-based feedback, adding simple robot state and force-torque sensing, to rich multimodal feedback incorporating haptic measurements and 3D LiDAR. *Familiarity of objects and tasks:* From known objects and tasks, over variations of known requiring parameter adaptations, to unfamiliar objects and tasks requiring compositional generalization. *Level of abstraction:* Starting with motion control on a fast time scale, proceeding with movement primitives, continuing with skills, such as grasping or placing an object, and finally considering chaining of skills to solve entire tasks. *Dynamics:* Starting with quasi-static motion by considering only kinematics, proceeding with slow, compliant motion with interaction forces, and finally modeling dynamic effects of fast movements.

Intuitive immersive telepresence systems enable transporting human presence to remote locations in real time. My team NimbRo developed the winning entry for the ANA Avatar XPRIZE competition [66] (see Fig. 6a). Telepresence also provides a rich source of environment interaction data for learning structured perception and autonomous behavior.

My team NimbRo develops perception, planning, and learning for anthropomorphic mobile manipulation robots providing personal assistance [105] and benchmarks them in the RoboCup@Home league, where we recently won the German Open 2024 and RoboCup 2024 OPL competitions [75] (see Fig. 6b).

8 Conclusions

By incorporating insights from human cognition, the next generation of service robots will systematically generalize their knowledge to cope with novelty. This

new generation of robots will also monitor themselves to obtain more information when needed, to avoid risks, and to detect and mitigate errors. Conscious service robots have much potential for numerous open-ended application domains, including assistance in everyday environments. Moreover, artificial conscious processing will contribute to a better understanding of consciousness in humans and other animals.

References

1. Anil, R., Dai, A.M., Firat, O., et al.: PaLM 2 technical report. CoRR (2023). ArXiv:2305.10403
2. Asfour, T., Paus, F., et al.: ARMAR-6: A high-performance humanoid for human-robot collaboration in real-world scenarios. IEEE Robotics Autom. Mag. **26**(4), 108–121 (2019)
3. Assran, M., Duval, Q., Misra, I., et al.: Self-supervised learning from images with a joint-embedding predictive architecture. In: IEEE/CVF Conf. on CVPR (2023)
4. Baars, B.J.: A cognitive theory of consciousness. Cambridge University Press (1993)
5. Bahdanau, D., Murty, S., Noukhovitch, M., et al.: Systematic generalization: What is required and can it be learned? In: 7th Int. Conf. on Learning Representations (ICLR) (2019)
6. Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V.F., et al.: Relational inductive biases, deep learning, and graph networks. CoRR (2018). ArXiv:2406.09246
7. Beetz, M., Kazhoyan, G., Vernon, D.: The CRAM cognitive architecture for robot manipulation in everyday activities. CoRR (2023). ArXiv:2304.14119
8. Behnke, S., Adams, J.A., Locke, D.: The $10 million ANA Avatar XPRIZE competition: How it advanced immersive telepresence systems. IEEE Robotics & Automation Magazine (RAM) **30**(4), 98–104 (2023)
9. Bengio, Y.: The consciousness prior. CoRR (2017). ArXiv:1709.08568
10. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**(8), 1798–1828 (2013)
11. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D.: Conflict monitoring and cognitive control. Psychological Review **108**(3), 624 (2001)
12. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., et al.: RT-1: Robotics transformer for real-world control at scale. In: Robotics: Science and Systems XIX (RSS) (2023)
13. Bubeck, S., Chandrasekaran, V., Eldan, R., et al.: Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR (2023). ArXiv:2303.12712
14. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: 37th Int. Conf. on Machine Learning (ICML) (2020)
16. Chen, Y., Yun, Z., Ma, Y., et al.: Minimalistic unsupervised representation learning with the sparse manifold transform. In: 11th Int. Conf. on Learning Representations (ICLR) (2023)
17. Chen, Z., Li, B., Wu, S., Jiang, K., Ding, S., Zhang, W.: Content-based unrestricted adversarial attack. In: Advances in Neural Information Processing Systems 36 (NeurIPS) (2023)
18. Cornelio, C., Diab, M.: Recover: A neuro-symbolic framework for failure detection and recovery. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS) (2024)
19. Cowan, N.: The magical mystery four: How is working memory capacity limited, and why? Current Directions in Psychological Science **19**(1), 51–57 (2010)
20. Dafarra, S., Pattacini, U., Romualdi, G., et al.: iCub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots. Science Robotics **9**(86) (2024)
21. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., et al.: RoboNet: Large-scale multi-robot learning. In: Conference on Robot Learning (CoRL), pp. 885–897 (2019)
22. Dehaene, S., Lau, H., Kouider, S.: What is consciousness, and could machines have it? Science **358**(6362), 486–492 (2017)

23. Dehghani, M., Djolonga, J., Mustafa, B., et al.: Scaling vision transformers to 22 billion parameters. In: Int. Conf. on Machine Learning (ICML), pp. 7480–7512 (2023)
24. Deng, B., Lewis, J.P., Jeruzalski, T., et al.: NASA: Neural articulated shape approximation. In: 16th European Conference on Computer Vision (ECCV), pp. 612–628 (2020)
25. Deng, J., Dong, W., Socher, R., et al.: ImageNet: A large-scale hierarchical image database. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
26. Driess, D., Xia, F., Sajjadi, M.S.M., et al.: PaLM-E: An embodied multimodal language model. In: Int. Conf. on Machine Learning (ICML), pp. 8469–8488 (2023)
27. Edelman, B.L., Goel, S., Kakade, S.M., Zhang, C.: Inductive biases and variable creation in self-attention mechanisms. In: Int. Conf. on Machine Learning (ICML) (2022)
28. Florence, P.R., Manuelli, L., Tedrake, R.: Self-supervised correspondence in visuomotor policy learning. IEEE Robotics Autom. Lett. (RA-L) **5**(2), 492–499 (2020)
29. Fodor, J.: Language, thought and compositionality. Royal Institute of Philosophy Supplements **48**, 227–242 (2001)
30. Friston, K.: Does predictive coding have a future? Nature Neuroscience **21**(8), 1019–1021 (2018)
31. Gao, G., Liu, W., et al.: GraphDreamer: Compositional 3D scene synthesis from scene graphs. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
32. d'Avila Garcez, A., Lamb, L.C.: Neurosymbolic AI: the 3rd wave. Artificial Intelligence Review **56**(11), 12387–12406 (2023)
33. Girdhar, R., El-Nouby, A., et al.: ImageBind one embedding space to bind them all. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2023)
34. Glanois, C., Jiang, Z., Feng, X., et al.: Neuro-symbolic hierarchical rule induction. In: Int. Conf. on Machine Learning (ICML), pp. 7583–7615 (2022)
35. Goyal, A., Bengio, Y.: Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A **478**(2266), 20210068 (2022)
36. Greff, K., van Steenkiste, S., Schmidhuber, J.: On the binding problem in artificial neural networks. CoRR (2020). ArXiv:2012.05208
37. Gui, J., Chen, T., et al.: A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **46**(12), 9052–9071 (2024)
38. Haarnoja, T., Moran, B., Lever, G., et al.: Learning agile soccer skills for a bipedal robot with deep reinforcement learning. Science Robotics **9**(89) (2024)
39. Han, J., Gong, K., Zhang, Y., et al.: OneLLM: One framework to align all modalities with language. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
40. Hauser, K., Watson, E., Bae, J., Bankston, J., Behnke, S., et al.: Analysis and perspectives on the ANA Avatar XPRIZE competition. Int. Journal of Social Robotics (SORO) (2024)
41. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2022)
42. Herzog, A., Rao, K., Hausman, K., et al.: Deep RL at scale: Sorting waste in office buildings with a fleet of mobile manipulators. In: Robotics: Science and Systems XIX (RSS) (2023)
43. Hessel, M., van Hasselt, H., Modayil, J., Silver, D.: On inductive biases in deep reinforcement learning. CoRR (2019). ArXiv:1907.02908
44. Hitzler, P., Sarker, M.K., Eberhart, A.: Compendium of Neurosymbolic Artificial Intelligence. IOS Press (2023)
45. Hjelm, R.D., Fedorov, A., et al.: Learning deep representations by mutual information estimation and maximization. In: 7th Int. Conf. on Learning Representations (ICLR) (2019)
46. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
47. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In: Robotics: Science and Systems (RSS) (2022)
48. Kaelbling, L.P., Lozano-Pérez, T.: Hierarchical task and motion planning in the now. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1470–1477 (2011)
49. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
50. Kalashnikov, D., Irpan, A., Pastor, P., et al.: Scalable deep reinforcement learning for vision-based robotic manipulation. In: Conference on Robot Learning (CoRL), pp. 651–673 (2018)

51. Kalashnikov, D., Varley, J., Chebotar, Y., Swanson, B., et al.: Scaling up multi-task robotic reinforcement learning. In: Conference on Robot Learning (CoRL), pp. 557–575 (2021)
52. Kazhoyan, G., Stelter, S., Kenfack, F.K., et al.: The robot household marathon experiment. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 9382–9388 (2021)
53. Kheddar, A., Roa, M.A., Wieber, P., et al.: Humanoid robots in aircraft manufacturing: The Airbus use cases. IEEE Robotics Automation Magazine (RAM) **26**(4), 30–45 (2019)
54. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., et al.: OpenVLA: An open-source vision-language-action model. CoRR (2024). ArXiv:2406.09246
55. Kittmann, R., Fröhlich, T., Schäfer, J., Reiser, U., Weißhardt, F., Haug, A.: Let me introduce myself: I am Care-O-bot 4, a gentleman robot. In: Mensch und Computer (MuC) (2015)
56. Klamt, T., Kamedula, M., Karaoguz, H., Kashiri, N., Laurenzi, A., Lenz, C., et al.: Flexible disaster response of tomorrow: Final presentation and evaluation of the CENTAURO system. IEEE Robotics & Automation Magazine (RAM) **26**(4), 59–72 (2019)
57. Klamt, T., Schwarz, M., Lenz, C., Baccelliere, L., Buongiorno, D., Cichon, T., others, Behnke, S.: Remote mobile manipulation with the Centauro robot: Full-body telepresence and autonomous operator assistance. Journal of Field Robotics (JFR) **37**(5), 889–919 (2020)
58. Kolesnikov, A., Beyer, L., Zhai, X., et al.: Big transfer (bit): General visual representation learning. In: 16th European Conference on Computer Vision (ECCV), pp. 491–507 (2020)
59. Kumagai, I., et al.: Toward industrialization of humanoid robots: Autonomous plasterboard installation to improve safety and efficiency. IEEE Robotics Autom. Mag. **26**(4), 20–29 (2019)
60. Lake, B.M., Baroni, M.: Generalization without Systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In: Int. C. on Machine Learning (ICML) (2018)
61. Lake, B.M., Baroni, M.: Human-like systematic generalization through a meta-learning neural network. Nature **623**(7985), 115–121 (2023)
62. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Sciences **40**, e253 (2017)
63. Lamb, L.C., d'Avila Garcez, A.S., et al.: Graph neural networks meet neural-symbolic computing: A survey and perspective. In: Int. J. Conf. on Artificial Intelligence (IJCAI) (2020)
64. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
65. Lee, M.A., Zhu, Y., Zachares, P., Tan, M., Srinivasan, K., Savarese, S., Fei-Fei, L., Garg, A., Bohg, J.: Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. IEEE Trans. Robotics **36**(3), 582–596 (2020)
66. Lenz, C., Schwarz, M., Rochow, A., Pätzold, B., Memmesheimer, R., Schreiber, M., Behnke, S.: NimbRo wins ANA Avatar XPRIZE immersive telepresence competition: Human-centric evaluation and lessons learned. International Journal of Social Robotics (2023)
67. Liang, J., Jiang, L., et al.: The garden of forking paths: Towards multi-future trajectory prediction. In: Conf. on Computer Vision and Pattern Recognition (CVPR) (2020)
68. Liu, B., Jiang, Y., Zhang, X., et al.: LLM+P: empowering large language models with optimal planning proficiency. CoRR (2023). ArXiv:2304.11477
69. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. in NeurIPS 36 (2023)
70. Liu, S., Lever, G., Wang, Z., et al.: From motor control to team play in simulated humanoid football. Science Robotics **7**(69) (2022)
71. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5222–5230 (2021)
72. Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., Raedt, L.D.: Neural probabilistic logic programming in DeepProbLog. Artificial Intelligence **298**, 103504 (2021)
73. Marra, G., Dumancic, S., Manhaeve, R., Raedt, L.D.: From statistical relational to neurosymbolic artificial intelligence: A survey. Artificial Intelligence **328**, 104062 (2024)
74. Meeussen, W., Wise, M., Glaser, S., et al.: Autonomous door opening and plugging in with a personal robot. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 729–736 (2010)
75. Memmesheimer, R., Nogga, J., Pätzold, B., Kruzhkov, E., Bultmann, S., Schreiber, M., Bode, J., Karacora, B., Park, J., Savinykh, A., Behnke, S.: RoboCup@Home 2024 OPL winner NimbRo: Anthropomorphic service robots using foundation models for perception and planning. In: RoboCup 2024: Robot World Cup XXVII. Springer (2025)

76. Menapace, W., Lathuilière, S., Tulyakov, S., Siarohin, A., Ricci, E.: Playable video generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
77. Mildenhall, B., Srinivasan, P.P., Tancik, M., et al.: NeRF: Representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2022)
78. Mitchell, T.M.: The need for biases in learning generalizations. Tech. rep., Department of Computer Science, Rutgers University, New Brunswick, NJ, USA (1980)
79. Mittal, M., Yu, C., Yu, Q., et al.: Orbit: A unified simulation framework for interactive robot learning environments. IEEE Robotics Autom. Lett. (RA-L) **8**(6), 3740–3747 (2023)
80. Moreno-Torres, J.G., Raeder, T., Alaíz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognit. **45**(1), 521–530 (2012)
81. Mosbach, M., Ewertz, J.N., Villar-Corrales, A., Behnke, S.: SOLD: Reinforcement learning with slot object-centric latent dynamics. CoRR (2024). ArXiv:2410.08822
82. Nau, D.S., Au, T., Ilghami, O., Kuter, U., Murdock, J.W., Wu, D., Yaman, F.: SHOP2: an HTN planning system. Journal of Artificial Intelligence Research **20**, 379–404 (2003)
83. Nye, M.I., Tessler, M.H., et al.: Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. In: Advances in NeurIPS 34 (2021)
84. O'Neill, A., Rehman, A., et al.: Open X-Embodiment: Robotic learning datasets and RT-X models. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 6892–6903 (2024)
85. OpenAI: Gpt-4 technical report. CoRR (2023). ArXiv:2303.08774
86. Pernias, P., Rampas, D., et al.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: 12th Int. Conf. on Learning Representations (ICLR) (2024)
87. Placed, J.A., Strader, J., Carrillo, H., et al.: A survey on active simultaneous localization and mapping: State of the art and new frontiers. IEEE Trans. Robotics **39**(3), 1686–1705 (2023)
88. Pratap, V., Tjandra, A., Shi, B., et al.: Scaling speech technology to 1,000+ languages. J. Mach. Learn. Res. (JMLR) **25**, 97:1–97:52 (2024)
89. Qin, Y., Liang, S., Ye, Y., et al.: ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In: 12th Int. Conf. on Learning Representations (ICLR) (2024)
90. Qin, Y., Shi, Z., Yu, J., et al.: WorldSimBench: Towards video generation models as world simulators. CoRR (2024). ArXiv:2410.18072
91. Radford, A., Kim, J.W., Xu, T., et al.: Robust speech recognition via large-scale weak supervision. In: Int. Conf. on Machine Learning (ICML), pp. 28492–28518 (2023)
92. Ravi, N., Gabeur, V., Hu, Y., et al.: SAM 2: Segment anything in images and videos. CoRR (2024). ArXiv:2408.00714
93. Rosu, R.A., Behnke, S.: PermutoSDF: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8466–8475 (2023)
94. Sajjadi, M.S.M., Duckworth, D., Mahendran, A., et al.: Object scene representation transformer. In: Advances in Neural Information Processing Systems 35 (NeurIPS) (2022)
95. Schmid, L., Abate, M., et al.: Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments. In: Robotics: Science and Systems (RSS) (2024)
96. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
97. Schrittwieser, J., Antonoglou, I., Hubert, T., et al.: Mastering Atari, Go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020)
98. Schulz, H., Behnke, S.: Deep learning – Layer-wise learning of feature hierarchies. Künstliche Intelligenz (Artificial Intelligence) **26**(4), 357–363 (2012)
99. Schwarz, M., Lenz, C., Memmesheimer, R., Pätzold, B., Rochow, A., Schreiber, M., Behnke, S.: Robust immersive telepresence and mobile telemanipulation: NimbRo wins ANA Avatar XPRIZE finals. In: 22nd IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids) (2023)
100. Shindo, H., Pfanschilling, V., Dhami, D.S., Kersting, K.: αILP: thinking visual scenes as differentiable logic programs. Machine Learning **112**(5), 1465–1497 (2023)
101. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**, 60 (2019)
102. Smet, L.D., Venturato, G., Raedt, L.D., Marra, G.: Relational neurosymbolic Markov models. In: 39th Conference on Artificial Intelligence (AAAI) (2025)

103. Smolensky, P., McCoy, R.T., et al.: Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems. AI Magazine **43**(3), 308–322 (2022)
104. Soulos, P., Hu, E.J., McCurdy, K., et al.: Differentiable tree operations promote compositional generalization. In: Int. Conf. on Machine Learning (ICML), pp. 32499–32520 (2023)
105. Stückler, J., Schwarz, M., Behnke, S.: Mobile manipulation, tool use, and intuitive interaction for cognitive service robot Cosero. Frontiers Robotics AI **3**, 58 (2016)
106. Stückler, J., Holz, D., Behnke, S.: RoboCup@Home: Demonstrating everyday manipulation skills in RoboCup@Home. IEEE Robotics & Automation Magazine **19**(2), 34–42 (2012)
107. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: EVA-CLIP-18B: Scaling CLIP to 18 billion parameters. CoRR (2024). ArXiv:2402.04252
108. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT Press (2018)
109. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D.: How to grow a mind: Statistics, structure, and abstraction. Science **331**(6022), 1279–1285 (2011)
110. Toyer, S., Trevizan, F.W., Thiébaux, S., Xie, L.: Action schema networks: Generalised policies with deep learning. In: 32nd Conf. on Artificial Intelligence (AAAI), pp. 6294–6301 (2018)
111. Trinh, T.H., Wu, Y., Le, Q.V., He, H., Luong, T.: Solving olympiad geometry without human demonstrations. Nature **625**(7995), 476–482 (2024)
112. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. CoRR (2024). ArXiv:2408.14837
113. Varadarajan, B., et al.: MultiPath++: Efficient information fusion and trajectory aggregation for behavior prediction. In: Int. Conf. on Robotics and Automation (ICRA) (2022)
114. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS), pp. 5998–6008 (2017)
115. Veerapaneni, R., Co-Reyes, J.D., Chang, M., Janner, M., et al.: Entity abstraction in visual model-based reinforcement learning. In: Conference on Robot Learning (CoRL) (2019)
116. Villar-Corrales, A., Wahdan, I., Behnke, S.: Object-centric video prediction via decoupling of object dynamics and interactions. In: IEEE International Conference on Image Processing (ICIP), pp. 570–574 (2023)
117. Vinyals, O., Babuschkin, I., Czarnecki, W.M., et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
118. Wang, H., et al.: Normalized object coordinate space for category-level 6D object pose and size estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
119. Wang, X., Misra, I., Zeng, Z., et al.: VideoCutLER: Surprisingly simple unsupervised video instance segmentation. In: IEEE/CVF Conf. on CVPR (2024)
120. Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: DayDreamer: World models for physical robot learning. In: Conference on Robot Learning (CoRL), pp. 2226–2240 (2022)
121. Wu, Z., Dvornik, N., Greff, K., et al.: SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In: 11th Int. Conf. on Learning Representations (ICLR) (2023)
122. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Networks Learn. Syst. **32**(1), 4–24 (2021)
123. Wurman, P.R., Barrett, S., Kawamoto, K., et al.: Outracing champion Gran Turismo drivers with deep reinforcement learning. Nature **602**(7896), 223–228 (2022)
124. Xie, Q., Luong, M., et al.: Self-training with noisy student improves ImageNet classification. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2020)
125. Yamamoto, T., Takagi, Y., et al.: Human support robot as research platform of domestic mobile manipulator. In: RoboCup 2019: Robot World Cup XXIII. Springer (2019)
126. Yang, J., Gao, S., Qiu, Y., et al.: Generalized predictive model for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
127. Yoshiike, T., et al.: The experimental humanoid robot E2-DR: A design for inspection and disaster response in industrial environments. IEEE Robo. Autom. Mag. **26**(4), 46–58 (2019)
128. Yu, T., Xiao, T., Tompson, J., et al.: Scaling robot learning with semantically imagined experience. In: Robotics: Science and Systems XIX (RSS) (2023)
129. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1204–1213 (2022)
130. Zhou, Y., Feinman, R., Lake, B.M.: Compositional diversity in visual concept learning. Cognition **244**, 105711 (2024)