Intelligent problem-solving as integrated hierarchical reinforcement learning

Manfred Eppe1,4✉, Christian Gumbsch2,3, Matthias Kerzel1, Phuong D. H. Nguyen1, Martin V. Butz2 and Stefan Wermter1

According to cognitive psychology and related disciplines, the development of complex problem-solving behaviour in biological agents depends on hierarchical cognitive mechanisms. Hierarchical reinforcement learning is a promising computational approach that may eventually yield comparable problem-solving behaviour in artificial agents and robots. However, so far, the problem-solving abilities of many human and non-human animals are clearly superior to those of artificial systems. Here we propose steps to integrate biologically inspired hierarchical mechanisms to enable advanced problem-solving skills in artificial agents. We first review the literature in cognitive psychology to highlight the importance of compositional abstraction and predictive processing. Then we relate the gained insights with contemporary hierarchical reinforcement learning methods. Interestingly, our results suggest that all identified cognitive mechanisms have been implemented individually in isolated computational architectures, raising the question of why there exists no single unifying architecture that integrates them. As our final contribution, we address this question by providing an integrative perspective on the computational challenges to develop such a unifying architecture. We expect our results to guide the development of more sophisticated cognitively inspired hierarchical machine learning architectures.

Humans and several other intelligent animal species have the ability to break down complex problems into simpler, previously learned sub-problems. This hierarchical approach allows them to solve previously unseen problems in a zero-shot manner, that is, without any trial and error. For example, Fig. 1 depicts how a crow solves a non-trivial food-access puzzle that consists of three causal steps: it first picks a stick, then uses the stick to access a stone, and then uses the stone to activate a mechanism that releases food. There exist numerous analogous experiments that attest similar capabilities to primates, octopuses and, of course, humans2,3. Human and animal cognition research suggests that hierarchical learning and problem-solving is critical to the development of such skills2,4. This raises the question of how we can equip intelligent artificial agents and robots with similar hierarchical learning and zero-shot problem-solving abilities.

We address this question from the perspective of reinforcement learning (RL)5–8. Several studies suggest that RL is biologically and cognitively plausible9. Many existing RL-based methods are aiming at zero-shot problem-solving and transfer learning, but this is currently only possible for minor variations of the same or a similar task10 or in simple synthetic domains, such as two-dimensional (2D) grid worlds11,12. A continuous-space problem-solving behaviour that is comparable with the crow’s behaviour in Fig. 1 has not yet been accomplished with any artificial system. We suggest that a focus on hierarchical learning and problem-solving is largely missing, given that research in hierarchical RL (HRL) is under-represented, as suggested by our meta review in Box 2 of the Supplementary Information.

In this Perspective, we address this gap in three steps. First, we evaluate the neurocognitive foundations of hierarchical decision making and identify important mechanisms that enable advanced problem-solving skills. Second, we reveal an integration problem of contemporary HRL methods: we show that the biological mechanisms have mostly been implemented already, but only in isolation and not in an integrative manner. Third, we suggest steps towards combining key methods and mechanisms to overcome the integration challenge for the development of a unifying cognitive architecture.

Neurocognitive foundations

In the following, we propose that the complex problem-solving skills in biological agents can be distinctively characterized by particular cognitive abilities. In our Perspective, these abilities depend on key cognitive mechanisms, including abstraction, intrinsic motivation and mental simulation. Figure 2 outlines how these mechanisms may develop from forward and inverse models, as the fundamental neuro-functional prerequisites.

Cognitive abilities. As distinct cognitive abilities, we focus on the following traits and properties.

Few-shot problem-solving. Few-shot problem-solving is the ability to solve unknown problems with few (≤5) trials. Zero-shot problem-solving is a special case of few-shot problem-solving, where no additional training at all is required to solve a new problem. For example, a crow can use a stick as a tool for a novel food-access problem without further training (Fig. 1), given it has previously solved related problems1. We identify two abilities that are most central for such problem-solving: transfer learning and decision-time planning (Fig. 2).

Transfer learning. Transfer learning allows biological agents to perform few-shot problem-solving by transferring the solution of a previously solved task to novel, but analogous, unseen tasks3. This notably reduces and sometimes eliminates the number of trials required to solve a problem.

1Universität Hamburg, Hamburg, Germany. 2University of Tübingen, Tübingen, Germany. 3Max Planck Institute for Intelligent Systems, Tübingen, Germany. 4Present address: Hamburg University of Technology, Hamburg, Germany. ✉Email: manfred.eppe@tuhh.de
Such analogical reasoning is often considered a critical cornerstone of human cognition. For example, cognitive theory suggests that humans understand mechanical systems by using analogical knowledge from other domains\cite{13}. Furthermore, education theory suggests that human transfer learning improves when explicitly trained to identify analogies between different problems\cite{11}. However, although humans seem skilful in applying analogies for problem-solving, it seems to be more difficult to discover new ones. The phenomenon of functional fixedness describes the human problem-solving tendency to use objects only in the way they are traditionally used\cite{14}.

**Goal-directed planning.** Behaviour is traditionally divided into two categories: stimulus-driven, habitual behaviour and goal-directed, planned behaviour\cite{15,16}. Habitual behaviour refers to forms of reflexive control that are strongly automated, computationally efficient and mainly learned from past reinforcements\cite{17}. One of the first critics of this approach was Tolman, who showed how rats can find rewards in a maze faster when they have visited the maze before, even if their previous visit had not been rewarded\cite{18}. The results suggest that the rats form a representation, or cognitive map, of the maze, which enables them to plan their behaviour once a reward is detected. Advances in artificial intelligence and cognitive science have influenced one another to understand goal-directed planning. One influential model for action control is the TOTE (test–operate–evaluate–exit) model\cite{19}, where actions are selected based on an iterative, hierarchical feedback loop. Herein, a desired anticipated state is compared to the current sensory state to select the next action. Similar to TOTE, many recent theories view planning as a hierarchical process with different levels of abstraction\cite{20,21,22,23}.

But what kind of information is considered while planning? Ideomotor theories of action generation\cite{24} suggest that anticipated action effects determine action selection. Accordingly, the theory of event coding\cite{25} claims that representations of actions and their effects are stored in a common code, whereby action–effect anticipations can directly activate corresponding actions\cite{26}. A range of experimental evidence supports this link between action generation and effect predictions\cite{27,28}. Supplementary Box 5 describes how the continuous learning of action–effect associations questions the traditional dichotomy between planned and habitual behaviour\cite{29}.

**Cognitive mechanisms for transfer learning and planning.** The previous paragraphs discussed how transfer learning and planning enable intelligent few-shot problem-solving. But what enables transfer learning and planning? Our survey suggests three critical mechanisms for transfer learning and planning, namely sensorimotor abstraction, intrinsic motivation and mental simulation.

**Sensorimotor abstraction.** According to theories of embodied cognition, abstract mental concepts are inferred from our bodily-grounded sensorimotor environmental interaction experiences\cite{30,31,32}. Here, we differentiate between action abstraction and state abstraction. Action abstractions refer to a temporally extended sequence of primitive actions, sometimes also referred to as options\cite{33} or movement primitives\cite{34,35}. For example, transporting an object is as an action abstraction because it is composed of a sequence of more primitive actions, such as reaching and lifting (Fig. 3).

State abstractions refer to reduced encodings of (parts of) the environment, where irrelevant details are abstracted away. Accordingly, state abstractions may decompose a scene into task-critical aspects, such as graspsability, containment or hollowness. For example, both wine glasses and teapot have the property of being a container and being graspsable (Fig. 3). Meanwhile, a container implies hollowness and so on, illustrating the hierarchic and compositional nature of such abstractions. Cognitive linguistics and related fields suggest that abstractions benefit from compositionality\cite{36}. Formally, expressions are compositional if they are composed of sub-expressions and rules to determine the semantics of the composition. The language of thought theory\cite{21} transfers the compositionality principle from language to abstract mental representations, claiming that thought must also be compositional. For example, humans can easily imagine previously unseen images by composing their known constituents from linguistic descriptions\cite{38}, even fantastic ones—think of a pink dragon riding a skateboard would look like.

It is still being debated whether the abstract mental representations are encoded as local symbols or by means of distributed encodings\cite{39}, but in both cases there is evidence for the existence of combinatorial generalization. For example, Haynes et al.\cite{40} show that the neural codes in the ventrolateral prefrontal cortex, which encode compound behavioural rules, can be decomposed into neural codes that encode their constituents.

For sensorimotor abstractions, the rules of composition need to be flexibly inferred by our minds to form plausible environmental scenarios in accordance with our world knowledge\cite{41,42,43}. Such common sense compositionality makes abstractions applicable in meaningful manners. The filling in of an abstract component (for example, the target of a grasp) is constrained towards applicable objects (for example, a graspable object such as a teapot), as depicted in Fig. 3. From an algorithmic perspective, common sense

---

**Fig. 1 | A New Caledonian crow solves a food-access problem.** a. First, the crow picks up a stick. b. It then uses the stick to pull a stone out of a tube. c. Finally, it uses the stone to activate a mechanism that releases food. Although the crow has never solved this problem set-up before, it is able to solve it instantly after a brief inspection phase. Such zero-shot problem-solving is a prime example of hierarchical, goal-directed planning that involves mental simulations and relies on task-specific, compositional abstractions. The authors reason that the crow plans ahead because it can only see one side of the cubic set-up at a time. Hence, the crow must mentally simulate the future steps. There is also transfer learning involved: previously learned solutions to similar sub-problems are transferred and combined to solve the new problem. Successful planning and transfer learning can only be expected to occur once abstract compositional mental representations have developed, which, for example, may encode tool manipulation options. To learn different tool manipulations, however, the crow relies on its intrinsic motivation to acquire knowledge by playing and experimenting with the tools. Images courtesy of R. Gruber and A. Taylor.
Compositionality brings enormous computational benefits: abstractions, such as a grasping encoding, are more economical if they can be applied in various situations, for example, for different grasping targets. Common sense compositionality essentially simplifies the identification of analogies by enabling mappings between representations, an advantage that is well known in cognitive theories of creativity and concept blending41,42 (Supplementary Box 4).

Intrinsic motivation.

Intrinsic motivation affects goal-directed reasoning by setting intrinsically motivated goals. From a cognitive development perspective, the term intrinsic motivation was coined to describe the ‘inherent tendency to seek out novelty, [...] to explore and to learn’43. Here, the term intrinsic is used in opposition to extrinsic values, which are associated with the satisfaction of needs or the avoidance of undesired outcomes44. Intrinsic motivations induce a general exploratory behaviour, curiosity and playfulness. Curiosity refers to information gain-oriented exploratory behaviour45–47. Curiosity is closely related to playfulness, where novel interactions are tried out in a skill-improvement-oriented manner. Playfulness can be observed in certain intelligent animal species, such as dogs and corvids48.

Mental simulation.

Mental simulation describes the ability of an agent to anticipate the dynamics of its environment on various representational levels. For example, motor-imagery is deemed to enable athletes to improve the execution of challenging body movements (for example, a backflip) by mentally simulating the movement49, and engineers can rely on mental simulation when developing mechanical systems13. Several researchers suggest that mental simulation also takes place on higher reasoning levels. For example, Kahnemann and Tversky50, Wells and Gavanski51 and later Taylor et al.52 report how mental simulation improves the planning of future behaviour on a causal conceptual level.

Forward and inverse models as functional prerequisites.

Sensorimotor abstraction, intrinsic motivation and mental simulation are complex mechanisms that demand a suitable neuro-functional basis. We propose that this essential basis is constituted by forward and inverse models24. Forward models predict how an action affects the world, and inverse models determine the actions that need to be executed to achieve a goal.

Forward and inverse models for mental simulation.

To perform mental simulation, an agent needs a forward model to simulate how the dynamics of the environment behave, possibly conditioned on the agent’s actions. However, the prediction of all consequences of all possible actions quickly becomes computationally intractable. Hence, action selection requires more direct mappings from state and desired inner mental state to potential actions. This can be accomplished by an inverse model or behaviour policies, which generate candidate actions to accomplish a current goal.
Forward and inverse models for intrinsic motivation. Forward and inverse models are also needed for intrinsic motivation. For example, the learning progress of internal forward models has been demonstrated to be useful for modeling curiosity as an intrinsic reward signal\(^{57,58}\). Friston et al. propose that intrinsically motivated behaviour emerges when using forward models to apply active inference\(^{57,59}\), that is, the process of inferring actions to minimize expected free energy. The authors argue that active inference can lead to epistemic exploration, when an agent acts out different behaviour to reduce the uncertainty about their forward predictions\(^{54,55}\). However, active inference and the free energy principle have been criticized as overly general, and their falsifiability has been questioned\(^{56}\).

Forward and inverse models for abstraction and grounding. Various cognitive theories, including predictive coding\(^{44,56}\), the free energy principle\(^{59}\) and the Bayesian brain hypothesis\(^{60}\), have viewed the brain as a generative machine, which constantly attempts to match incoming sensory signals with its probabilistic predictions\(^{61}\). Within these frameworks, prediction takes place on multiple processing hierarchies that interact bidirectionally: top–down information provides additional information to predict sensory signals, and bottom–up error information is used to correct high-level predictions\(^{62}\). Butz\(^{29}\) proposes that this hierarchical processing leads to the emergence of abstract, compositional encodings on higher processing levels, for example, a container concept, enabling predictions of lower-level features, such as the occlusion and transport of a contained object. Event segmentation theory (EST)\(^{63}\) makes the role of forward predictions for learning sensorimotor abstractions even more explicit. According to EST, transient prediction errors are used to segment the continuous stream of experience into conceptual event encodings, as detailed in Supplementary Box 6.

Computational realizations

Computational HRL mechanisms (Fig. 4) are less integrated than those of biological agents. In the following, we summarize the state of the art of computational approaches in relation to the neurocognitive foundations. We provide an additional overview in Supplementary Table 1.

Transfer learning and planning for few-shot abilities. Our survey of the neurocognitive foundations indicates two foundational cognitive abilities for few-shot problem-solving: transfer learning and planning. But how can we model these abilities computationally?

Fig. 3 | Compositional action and state abstractions. a, Action abstractions describe a sequence of more primitive actions. b, State abstractions encode certain properties of the state space. c, The compositionality of the abstractions enables their general application by instantiating abstract definitions (the graspable target of reaching) with a specific object (teapot).

Transfer learning. Transfer learning is natural for hierarchical architectures in the sense that they can potentially re-use general high-level skills and apply them to different low-level use cases, and vice versa. Therefore, a substantial fraction of the existing transfer learning approaches build on learning reusable low-level policies\(^{64-72}\). Research has not only considered transfer learning between different tasks, but also between different robot and agent morphologies\(^{73,74}\).

Planning. Planning can be categorized into decision-time planning and background planning\(^{75,76}\). Decision-time planning refers to a search over sequences of actions to decide which action to perform next to achieve a specific goal. The search is realized by simulating the actions with an internal predictive model of the domain dynamics. Decision-time planning enables zero-shot problem-solving because the agent only executes the planned actions if an internal prediction model verifies that the actions are successful. Background planning refers to model-based RL as introduced by Sutton\(^{77}\). It is used to train a policy by simulating actions with a forward model. This can improve the sampling efficiency, but not to the extent that it enables few-shot problem-solving.

Hierarchical decision-time planning is a well-known paradigm in classical artificial intelligence\(^{77}\), but approaches that integrate it with HRL are rare. Some approaches integrate action planning with RL by using an action planner for high-level decision making and a reinforcement learner for low-level motor control\(^{78,79}\). These approaches depend on decision-time planning being especially useful for high-level inference in discrete state-action spaces.

Mechanisms behind transfer learning and planning. Our summary of the cognitive principles behind transfer learning and planning reveals three important mechanisms that are critical for the learning and problem-solving abilities of biological agents, namely compositional sensorimotor abstraction, intrinsic motivation and mental simulation.

Compositional sensorimotor abstraction and grounding. HRL has a temporal and a representational dimension, and one can distinguish between action and state abstraction. The temporal abstraction of actions breaks down complex problems into a hierarchy of simpler problems. The probably most influential method for representational action abstraction builds on behaviour primitives\(^{80,81,82,83,84}\), including options\(^{84}\), sub-policies or atomic high-level skills. Such behaviour primitives are represented in an abstract representational space, for example, in a discrete finite space of action labels or
Perspective Nature Machine Intelligence

HRL approaches focus on two layers, where the high-level action \( a^1 \) is either a behavioural primitive or a subgoal. The low-level actions \( a^2 \) are motor commands that affect the environment directly.

indices (Fig. 5a). A more recent type of high-level action representation comprises subgoals in the low-level state space \( \{0,1,2,3\} \) (Fig. 5b). Here, the agent achieves the final goal by following a sequence of subgoals. Supplementary Table 1 shows that there is considerably less contemporary research on state abstraction than on action abstraction.

Many current RL approaches based on visual input use convolutional neural networks to perform a rather generic form of abstraction, where all hierarchical layers use the same abstract visual feature map \( v_1, v_2, v_3 \). The problem with this is that it does not appreciate that different levels of inference require different levels of abstraction: for example, for a transport task, the exact shape and weight of an object is only relevant for low-level motor control. A high-level planning layer only requires abstract facts, such as whether the object is graspable or transportable. Therefore, the representational state abstraction should fit the level of inference for which it is used. We refer to such abstraction as inference-fitted abstraction. Inference-fitted abstraction has been tackled in existing HRL approaches, but many of them rely on manually defined abstraction functions \( v_1, v_2, v_3 \). There exist only a few exceptions where state abstractions are derived automatically in HRL, for example, through clustering \( v_1, v_2, v_3 \), with feature representations \( v_1, v_2, v_3 \) or by factorization \( v_1, v_2, v_3 \).

Cognitive science also suggests that compositionality is an important property of abstract representations. For example, a symbolic compositional action representation "grasp (glass)" allows for modulating the action "grasp with the object glass." Compositionality is also applicable to distributed numerical expressions, such as vector representations. For example, a vector \( v_1 \) is a compositional representation if it is composed of other vectors, for example, \( v_2 \) and \( v_3 \), and if there is a vector operation \( \ast \) that implements interpretation rules, for example, \( v_1 = v_2 \ast v_3 \).

There is notable cognitive evidence that compositionality improves transfer learning. This evidence is computationally plausible when considering that transfer learning relies on exploiting analogies between problems, as described in Supplementary Box 4. Other cognitively inspired computational methods, such as concept blending \( v_1, v_2 \), provide further evidence that a lower number of possible analogy mappings improves transferability. Furthermore, the work by Jiang et al. \( v_1, v_2 \) provides empirical evidence that compositional representations improve learning transfer. The authors use natural language as an inherently compositional representation to describe high-level actions in HRL, and show that this natural language representation improves transfer learning performance compared to non-compositional representations.

Intrinsic motivation. Intrinsic motivation is a useful method to stabilize RL by supplementing sparse extrinsic rewards with additional domain-independent intrinsic rewards. The most common method of HRL with intrinsic motivation provides intrinsic rewards when subgoals or subtasks are achieved.

Other approaches provide intrinsic motivation to identify a collection of behavioural primitives with high diversity \( \{0,1,2,3\} \) and predictability \( \{0,1,2,3\} \). This also includes the identification of primitives suitable for recomposing high-level tasks \( \{0,1,2,3\} \).

Another prominent intrinsic reward model that is commonly used in non-hierarchical RL is based on surprise and curiosity \( \{0,1,2,3\} \). In these approaches, surprise is often modelled as a function of the prediction error of a forward model, and curiosity is realized by providing intrinsic rewards if the agent is surprised. However, only Blaes et al. \( v_1, v_2, v_3 \), Colas et al. \( v_1, v_2, v_3 \) and Röder et al. \( v_1, v_2, v_3 \) use surprise in a hierarchical setting, showing that hierarchical curiosity leads to a substantial improvement of the learning performance. Hierarchical curiosity provides potential synergies because a curious high-level layer automatically acts as a curriculum generator for a low-level layer. For example, it can generate 'Goldilocks' subgoals that are neither too hard nor too easy to achieve for the low-level layer. The difficulty is to identify these Goldilocks conditions and to find an appropriate balance between explorative and exploitative subgoals.

Mental simulation. Mental stimulation enables an agent to anticipate the effects of its own and other actions. Therefore, it is a core mechanism to equip an intelligent agent with the ability to plan ahead. Computational approaches for mental simulation involve model-based RL \( v_1, v_2, v_3 \), action planning \( v_1, v_2, v_3 \) or a combination of both \( v_1, v_2, v_3 \). However, even though there is cognitive evidence that mental simulation happens on multiple representational layers \( v_1, v_2, v_3 \), there is a lack of approaches that use hierarchical mental simulation. Only a few hierarchical approaches that integrate planning with RL build on mental simulation \( v_1, v_2, v_3 \), although the mental simulation mechanisms of these models are only implemented on the high-level planning layer. An exception is the work by Wu et al. \( v_1, v_2, v_3 \), who use mental simulation on the low-level layer to determine the sub-policies to be executed. Another exception is presented by Li et al. \( v_1, v_2, v_3 \), who perform mental simulation for planning on multiple task layers. However, we are not aware of any approach that performs hierarchical background planning, that is, that trains a hierarchical policy by mentally simulating the action execution on multiple levels. One possible reason for this lack is that high-level background planning requires stable low-level representations, and it is difficult to model a respective stability criterion.

Prerequisites for sensorimotor abstraction, intrinsic motivation and mental simulation. Reinforcement learning builds on policies or inverse models that predict the actions to be executed in the current state to maximize reward or to achieve a goal state. By contrast, a forward model predicts a future world state based on the current state and a course of actions. Our review shows that the vast majority of HRL methods use inverse models for both the high-level and low-level layers. Some approaches exist that use an inverse model only for the low-level layer \( v_1, v_2, v_3 \) and use decision-time planning to perform high-level decision making. Our review also implies that a forward model is required for several additional mechanisms that are necessary or at least highly beneficial for transfer learning, planning and few-shot problem-solving. Some non-hierarchical approaches use a forward model to perform sensorimotor abstraction \( v_1, v_2, v_3 \). They achieve this with a self-supervised process, where forward predictions are learned in a latent abstract space. However, we are not aware of any hierarchical method that exploits this mechanism.

A forward model is also useful to model curiosity as an intrinsic motivation mechanism, for example, by rewarding an agent for being surprised, where surprise is a function of the prediction error \( v_1, v_2, v_3 \). Curiosity has been demonstrated to alleviate the sparse...
rewards problem in RL, but, with few exceptions, there is a lack of approaches that use a hierarchical forward model for generating hierarchical curiosity.

Critical discussion and outlook
Based on our considerations above and Supplementary Box 3 and Supplementary Table 1, we now summarize the shortcomings of current HRL approaches. The shortcomings are consequences of specific challenges that arise from the integration of the different mechanisms (Fig. 6). Here, we address these shortcomings with specific algorithmic suggestions.

Shortcomings and challenges of HRL architectures. The following sections identify the major challenges for current methods.

Hierarchical mental simulation requires stable hierarchical representations and precise predictions. Most current few-shot problem-solving methods build on transfer learning, but they neither leverage decision-time planning nor background planning to exploit hierarchical mental simulation. Although numerous non-hierarchical methods show how mental simulation can improve sample efficiency, our summary in Supplementary Table 1 shows that RL with hierarchical mental simulation on multiple layers is implemented in only one recent method. A major reason for this shortcoming is that the low-level policy should have already converged before the high-level forward model starts training. This problem has been addressed for model-free HRL, but it has not yet been investigated with predictive forward models.

Fig. 5 | Types of action abstraction. a, With behaviour primitives, the agent determines the path to the final goal by selecting a sequence of high-level actions, but without specifying explicitly to which intermediate state each action leads. b, With subgoals, the agent determines the intermediate states as subgoals that lead to the final goal, but without specifying the actions to achieve the subgoals.
Intrinsic motivation requires forward models with elaborate uncertainty handling. Supplementary Table 1 show that curiosity and diversity are under-represented intrinsic motivation methods, being implemented in only 9 of 50 surveyed papers. The section Neurocognitive foundations presents substantial cognitive evidence that these mechanisms are critical for intelligent behaviour in biological agents.

A potential reason for this shortcoming is that prediction-based intrinsic motivation suffers from the overly simplistic uncertainty handling of current computational forward models. This is an issue for noisy observations or chaotic and unpredictable system dynamics. For example, Röder et al. achieve curiosity-driven intrinsic motivation by maximizing the prediction error of a forward model. This causes the agent to suffer from the so-called ‘noisy TV problem’, where an agent always strives to encounter unpredictable observations (such as a TV showing only white noise). This interferes with the problem of learning hierarchical representations because it does not address the uncertainties of high-level representations if the learning of these representations has not yet converged.

Compositionality requires elaborate role-filler representations. Only a few HRL methods consider compositionality. Exceptions include natural language-based representations of actions or symbolic logic-based compositional representations. Because compositional representations generally build on variable-value representations, these need to be researched first on a more foundational level to alleviate this shortcoming. Graph neural networks, a class of neural networks that encode entities and their interactions as nodes and edges within a graph, are a promising approach for compositional representations. When provided with the underlying graph structure, graph neural networks seem to generalize better and support relational reasoning. However, discovering a suitably abstracted graph from data remains an open challenge.

Inference-fitted state abstraction requires a strong learning signal. Recent HRL approaches use the same state representation for each layer in the hierarchy. However, it seems cognitively more plausible to fit the level of abstraction to the level of inference needed on each hierarchical level. For example, a high-level layer may only need to know whether an object is graspable—it does not need to consider the precise shape of the object. This more detailed information, however, is relevant for the low-level layer that performs the grasping. There exist RL methods that learn state abstractions from visual data with a convolutional autoencoder. However, such representations are optimized for reconstructing the visual input—not for decision making. Furthermore, in existing recent vision-based architectures, all hierarchical layers of the RL agent are provided with the same image embedding. Therefore, the abstractions are not inference-fitted. A problem for learning inference-fitted abstraction functions is that the reward signal of RL is a relatively weak driver for the learning process compared to other methods such as supervised learning.

Outlook and suggestions for addressing the challenges. In the following, we sketch how computational key methods may help to address the above challenges.

Retrospective analysis of experiences. A central problem of HRL is that stable high-level representations can only emerge once the learning of the lower level has converged. We suggest that a retrospective analysis of experiences can alleviate this issue. This has been demonstrated for training a policy by learning in hindsight, where the achieved world state of a policy is retrospectively pretended to be the desired goal state. We suggest that hindsight learning is a special case of a more general retrospective analysis of memorized experiences. For example, it is potentially possible to also learn hierarchical forward models in hindsight, without the high-level layer having to wait for a low-level layer to converge. We hypothesize that reconsidering the outcome of past erroneous actions of a not yet fully trained low-level layer as a desired outcome can stabilize the high-level representations. Such a technique may also help to alleviate the model bias.

Meta-uncertainty models. Prediction errors play an important role in mental simulation and intrinsic motivation. In the section Neurocognitive foundations we have discussed the free energy principle, which involves two seemingly contradicting mechanisms that build on predictive processing models. On the one hand there is active inference, which seeks to minimize long-term surprise. This can be implemented by minimizing a forward model’s uncertainty or prediction error. On the other hand there is active learning, which seeks to maximize information gain. This can be implemented by maximizing uncertainty through respective intrinsic reward functions. Schwartenbeck et al. address this contradiction by building on expected uncertainty to unify both mechanisms. Expected as well as unexpected uncertainty can then be understood as a meta-uncertainty analysis. We suppose that an intrinsic reward for high meta-uncertainty may lead to exploratory behaviour, causing the exploration of high short-term but anticipated lower long-term meta-uncertainty. Furthermore, the gained certainty about the uncertainty of forward models can lead to improved predictive processes that are aware of their own limits.

Bidirectional dynamics models. Bidirectional models involve both forward and backward inference, and we suppose that this
combination is beneficial for representational abstraction in hierarchical architectures. Our supposition is based on work by Pathak et al. and by Hafner et al., who use a combined forward and inverse model to generate latent representations in a self-supervised manner. The bidirectional model operates on a latent space, generated by an abstraction function. By simultaneously performing inverse and backward prediction in the latent space, the abstraction function gains two important features. First, the latent space is de-noised because only the predictable parts of the latent observation determine the forward and backward prediction. Second, the method is self-supervised and therefore provides a more stable feedback signal than the comparatively weak reward signal. This alleviates the prediction error and strengthens the overall learning signal.

Conclusion
We have provided an overview of the cognitive foundations of hierarchical problem-solving and how these are implemented in current HRL architectures. We consider few-shot problem-solving as the ultimate ability of intelligent beings that learn to solve causally non-trivial problems with as few trials as possible. As our main research question, we ask for the computational prerequisites and mechanisms to enable problem-solving for artificial computational agents on the level of intelligent animals. We further pose the challenge to integrate the different mechanisms.

We appreciate that the cognitive theories we summarize are often contentious and controversial, as indicated in the Neurocognitive foundations section. However, in our discussion we show that particular combinations of the described theories have the potential to yield algorithmic synergies. In other words, we reason that if a particular mechanism is useful for computational architectures, then it is likely to be beneficial for biological cognitive systems as well, and vice versa. In this way, cognitive sciences and artificial intelligence inspire each other, and they will hopefully converge to insights that are relevant for both fields and useful for the future of artificial intelligence and humanity.

Received: 18 December 2020; Accepted: 7 December 2021; Published online: 25 January 2022

References
1. Gruber, R. et al. New Caledonian crows use mental representations to solve metalool problems. Curr. Biol. 29, 686–692 (2019).
2. Butz, M. V. & Katter, E. F. How the Mind Comes into Being (Oxford Univ. Press, 2017).
3. Perkins, D. N. & Salomon, G. in International Encyclopedia of Education (eds. Husen T. & Postelwhite T. N.) 6452–6457 (Pergamon Press, 1992).
4. Botvinick, M. M., Niv, Y. & Barto, A. C. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113, 262–280 (2009).
5. Tomov, M. S., Vagati, S., Kumar, A., Yang, W. & Gershman, S. J. Discovery of hierarchical representations for efficient planning. PLoS Comput. Biol. 16, e1007594 (2020).
6. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: a brief survey. IEEE Signal Process. Mag. 34, 26–38 (2017).
7. Li, Y. Deep reinforcement learning: an overview. Preprint at https://arxiv.org/abs/1701.07274 (2018).
8. Sutton, R. S. & Barto, A. G. Reinforcement Learning: An Introduction 2nd edn (MIT Press, 2018).
9. Nefci, O. & Averbeck, B. B. Reinforcement learning in artificial and biological systems. Nat. Mach. Intell. 1, 133–143 (2019).
10. Eppe, M., Nguyen, P. D. H. & Werment, S. From semantics to execution: integrating action planning with reinforcement learning for robotic causal problem-solving. Front. Robot. AI 6, 123 (2019).
11. Oh, J., Singh, S., Lee, H. & Kohli, P. Zero-shot task generalization with multi-task deep reinforcement learning. In Proc. 34th International Conference on Machine Learning (ICML) (eds. Precup, D. & Teh, Y. W.) 2661–2670 (PMLR, 2017).
12. Riabinin, S., Oh, J. & Lee, H. Hierarchical reinforcement learning for zero-shot generalization with subtask dependencies. In Proc. 32nd International Conference on Neural Information Processing Systems (NeurIPS) (eds Bengio S. et al.) Vol. 31, 7156–7166 (ACM, 2018).
13. Hegarty, M. Mechanical reasoning by mental simulation. Trends Cogn. Sci. 8, 280–285 (2004).
14. Klauer, K. J. Teaching for analogical transfer as a means of improving problem-solving, thinking and learning. Instruct. Sci. 18, 179–192 (1989).
15. Duncker, K. & Lees, L. S. On problem-solving. Psychol. Monographs 58, No.5 (whole No. 270), 85–101 https://doi.org/10.1037/h0093599 (1945).
16. Dayan, P. Goal-directed control and its antipodes. Neural Netw. 22, 213–219 (2009).
17. Dolan, R. J. & Dayan, P. Goals and habits in the brain. Neuron 80, 312–325 (2013).
18. O’Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, reward, and decision making. Annu. Rev. Psychol. 68, 73–100 (2017).
19. Tolman, E. C. & Honzik, C. H. Introduction and removal of reward, and maze performance in rats. Univ. California Publ. Psychol. 4, 257–275 (1930).
20. Butz, M. V. & Hoffmann, J. Anticipations control behavior: animal behavior in an anticipatory learning classifier system. Adaptive Behav. 10, 75–96 (2002).
21. Miller, G. A., Galanter, E. & Pribram, K. H. Plans and the Structure of Behavior (Holt, Rinehart & Winston, 1960).
22. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. Philos. Trans. R. Soc. B Biol. Sci. 369, 20130480 (2014).
23. Wiener, J. M. & Mallot, H. A. ‘Fine-to-coarse’ route planning and navigation in regionalized environments. Spatial Cogn. Comput. 3, 331–358 (2003).
24. Stock, A. & Stock, C. A short history of ideo-motor action. Psychol. Res. 68, 176–188 (2004).
25. Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. The theory of event coding (TEC): a framework for perception and action planning. Behav. Brain Sci. 24, 849–878 (2001).
26. Hoffmann, J. in Anticipatory Behavior in Adaptive Learning Systems: Foundations, Theories and Systems (eds Butz, M. V. et al.) 44–65 (Springer, 2003).
27. Kunde, W., Elsner, K. & Kiesel, A. No anticipation-no action: the role of anticipation in action and perception. Cogn. Process. 8, 71–78 (2007).
28. Barsalou, L. W. Grounded cognition. Annu. Rev. Psychol. 59, 617–645 (2008).
29. Butz, M. V. Toward a unified sub-symbolic computational theory of cognition. Front. Psychol. 7, 925 (2016).
30. Pulvermüller, F. Brain embodiment of syntax and grammar: discrete combinatorial mechanisms spelt out in neuronal circuits. Brain Lang. 112, 167–179 (2010).
31. Sutton, R. S., Precup, D. & Singh, S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif. Intell. 112, 181–211 (1999).
32. Flash, T. & Hochner, B. Motor primitives in vertebrates and invertebrates. Curr. Opin. Neurobiol. 15, 680–686 (2005).
33. Schaal, S. in Adaptive Motion of Animals and Machines (eds Kimura, H. et al.) 261–280 (Springer, 2006).
34. Fieldman, J., Dodge, E. & Bryant, J. In The Oxford Handbook of Linguistic Analysis (eds Heine, B. & Narrog, H.) 111–138 (Oxford Univ. Press, 2009).
35. Fodor, J. A. Language, thought and compositionality. Mind Lang. 16, 1–15 (2001).
36. Frankland, S. M. & Greene, J. D. Concepts and compositionality: in search of the brain’s language of thought. Annu. Rev. Psychol. 71, 273–303 (2020).
37. Hommel, J. E. Getting symbols out of a neural architecture. Connection Sci. 23, 109–118 (2011).
38. Haynes, J. D., Winiewski, D., Gorgen, K., Momennajad, I. & Reverberi, C. FMRI decoding of intentions: compositionality, hierarchy and prospective memory. In Proc. 3rd International Winter Conference on Brain-Computer Interface (BCI), 1-3 (IEEE, 2015).
39. Härnström, P. The Geometry of Meaning: Semantics Based on Conceptual Spaces (MIT Press, 2014).
40. Lakoff, G. & Johnson, M. Philosophy in the Flesh (Basic Books, 1999).
41. Eppe, M. et al. A computational framework for concept blending. Artif. Intell. 256, 105–129 (2018).
42. Turner, M. The Origin of Ideas (Oxford Univ. Press, 2014).
43. Deci, E. L. & Ryan, R. M. Self-determination theory and the facilitation of intrinsic motivation. Am. Psychol. 55, 68–78 (2000).
44. Friston, K. et al. Active inference and epistemic value. Cogn. Neurosci. 6, 187–214 (2015).
45. Berlyne, D. E. Curiosity and exploration. Am. Psychol. 25, 105–129 (1970).
46. Oudeyer, P.-Y., Kaplan, F. & Hafner, V. Intrinsic motivation systems for autonomous mental development. In IEEE Transactions on Evolutionary Computation (eds Coello, C. A. C. et al.) Vol. 11, 265–286 (IEEE, 2007).
47. Pisansky, W. Play and exploration in animals—a comparative analysis. Psychol. Psychol. Bull. 39, 104–107 (2008).
49. Jeannerod, M. Mental imagery in the motor context. *Neuropsychologia* **33**, 1419–1432 (1995).
50. Kahnemann, D. & Tversky, A. In *Judgement under Uncertainty: Heuristics and Biases* (eds Kahneman, D. et al.) Ch. 14, 201–208 (Cambridge Univ. Press, 1982).
51. Wells, G. L. & Gavanski, I. Mental simulation of causality. *J. Personal. Social Psychol.* **56**, 161–169 (1989).
52. Taylor, S. E., Pham, L. B., Rivkin, I. D. & Armor, D. A. A harnessing the imagination: mental simulation, self-regulation and coping. *Am. Psychol.* **53**, 429–439 (1998).
53. Kaplan, F. & Oudeyer, P.-Y. In *Embodied Artificial Intelligence*, Lecture Notes in Computer Science Vol. 3139 (eds Iida, F. et al.) 259–270 (Springer, 2004).
54. Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation. *IEEE Trans. Auton. Mental Dev.* **2**, 230–247 (2010).
55. Friston, K., Mattout, J. & Kijler, J. Action understanding and active inference. *Biol. Cybern.* **104**, 137–160 (2011).
56. Oudeyer, P.-Y. Computational theories of curiosity-driven learning. In *The New Science of Curiosity* (eds. Goren Gordon), 43–72 (Nova Science Publishers, 2018); https://arxiv.org/abs/1802.10546
57. Clark, A. M. & Wright, C. First principles in the life sciences: the free-energy principle, organismism and mechanisms. *Synthese* **198**, 3463–3488 (2021).
58. Huang, Y. & Rao, R. P. Predictive coding. *WIREs Cogn. Sci.* **2**, 580–593 (2011).
59. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
60. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
61. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
62. Clark, A. Surfing Uncertainty: Prediction, Action and the Embodied Mind (Oxford Univ. Press, 2016).
63. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynoldson, J. R. Event perception: a mind/brain perspective. *Psychol. Bull.* **133**, 273–293 (2007).
64. Eysenbach, B., Ibzar, J., Gupta, A. & Levine, S. Diversity is all you need: learning skills without a reward function. In *International Conference on Learning Representations (ICLR, 2019).*
65. Frans, K., Ho, J., Chen, X., Abbeel, P. & Schulman, J. Meta learning shared hierarchies. In *Proc. International Conference on Learning Representations* https://openreview.net/pdf?id=SylX6eWAV (ICLR, 2018).
66. Heess, N. et al. Learning and transfer of modular locomotor controllers. Preprint at https://arxiv.org/abs/1610.05182 (2016).
67. Jiang, Y., Gu, S., Murphy, K. & Finn, C. Language as an abstraction for hierarchical deep reinforcement learning. In *Neural Information Processing Systems (NeurIPS)* (eds. Wallach, H. et al.) 9414–9426 (ACM, 2019).
68. Li, A. C., Florensa, C., Clavera, I. & Abbeel, P. Sub-policy adaptation for hierarchical reinforcement learning. In *Proc. International Conference on Learning Representations* https://openreview.net/forum?id=Syl3w076S (ICLR, 2020).
69. Qureshi, A. H. et al. Composing task-agnostic policies with deep reinforcement learning. In *Proc. International Conference on Learning Representations* https://openreview.net/forum?id=H1eFRlKhV (ICLR, 2020).
70. Sharma, A., Gu, S., Levine, S., Kumar, V. & Hausaman, K. Dynamics-aware unsupervised discovery of skills. In *Proc. International Conference on Learning Representations* https://openreview.net/forum?id=H1gL2Rk4kV (ICLR, 2020).
71. Toseler, C., Givony, S., Zahav, T., Mankowitz, D. I. & Mannor, S. A deep hierarchical approach to lifelong learning in minecraft. In *Proc. 31st AAAI Conference on Artificial Intelligence 1553–1561* (AAAI, 2017).
72. Vezhnevets, A. et al. Strategic attentive writer for learning macro-actions. In *Neural Information Processing Systems (NeurIPS)* (eds. Lee, D. et al.) 3494–3502 (NIPS, 2016).
73. Devin, C., Gupta, A., Darrell, T., Abbeel, P. & Levine, S. Learning modular neural network policies for multi-task and multi-robot transfer. In *Proc. International Conference on Robotics and Automation (ICRA)* (eds. Okamura, A. et al.) 2169–2176 (IEEE, 2017).
74. Hejna, D. J., Abbeel, P. & Pinto, L. Hierarchically decoupled morphological transfer. In *Proc. International Conference on Machine Learning (ICML)* (eds. Daumé III, H. & Singh, A.) 11409–11420 (PMLR, 2020).
75. Hamrick, J. B. et al. On the role of planning in model-based deep reinforcement learning. In *Proc. International Conference on Learning Representations* https://openreview.net/pdf?id=H1e6IDGB2 (ICLR, 2021).
76. Seoane, R. S. Integrative architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proc. 7th International Conference on Machine Learning (ICML)* (eds. Porter, B. W. & Mooney, R. J.) 216–224 (Morgan Kaufmann, 1990).
77. Nau, D. et al. SHOP2: an HTN planning system. *J. Artif. Intell. Res.** Res.*** **20**, 379–404 (2003).
101. Rasmussen, D., Voelker, A. & Eliasmith, C. A neural model of hierarchical reinforcement learning. *PLoS ONE* **12**, e0180234 (2017).

102. Riedmiller, M. et al. Learning by playing—solving sparse reward tasks from scratch. In *Proc. International Conference on Machine Learning (ICML)* (eds. Dy, J. & Krause, A.) Vol. 10, 6910–6919 (PMLR, 2018).

103. Yang, F., Lyu, D., Liu, B. & Gustafson, S. PEORL: integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. In *Proc. 27th International Joint Conference on Artificial Intelligence (IJCAI)* (ed. Lang, J.) 4860–4866 (IJCAI, 2018).

104. Machado, M. C., Bellemare, M. G. & Bowling, M. A Laplacian framework for option discovery in reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)* (eds. Precup, D. & Teh, Y. W.) Vol. 5, 3567–3582 (PMLR, 2017).

105. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proc. 34th International Conference on Machine Learning (ICML)* (eds. Precup, D. & Teh, Y. W.) 2778–2787 (PMLR, 2017).

106. Schillaci, G. et al. Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces. *Adaptive Behav.* **29**, 549–566 (2020).

107. Colas, C., Fournier, P., Sigaud, O., Chetouani, M. & Oudeyer, P.-Y. CURIOUS: intrinsically motivated modular multi-goal reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)* (eds. Chaudhuri, K. & Salakhutdinov, R.) 1331–1340 (PMLR, 2019).

108. Hafner, M. B., Weber, C., Kerzel, M. & Wermter, S. Improving robot dual-system motor learning with intrinsically motivated meta-control and latent-space experience imagination. *Robot. Auton. Syst.* **133**, 103630 (2020).

109. Yamamoto, K., Onishi, T. & Tsuruoka, Y. Hierarchical reinforcement learning with abductive planning. In *Proc. ICMUI/JCAI/AAMAS 2018 Workshop on Planning and Learning (PAL-18)* (2018).

110. Wu, B., Gupta, J. K. & Kochenderfer, M. J. Model primitive hierarchical lifelong reinforcement learning. In *Proc. International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (eds. Agmon, N. et al.) Vol. 1, 34–42 (IFAAMAS, 2019).

111. Li, Z., Narayan, A. & Leong, T. Y. An efficient approach to model-based hierarchical reinforcement learning. In *Proc. 31st AAAI Conference on Artificial Intelligence* 3583–3589 (AAAI, 2017).

112. Hafner, D., Lillicrap, T. & Norouzi, M. Dream to control: learning behaviors by latent imagination. In *Proc. International Conference on Learning Representations* https://openreview.net/pdf?id=S1OYTChlDS (ICLR, 2020).

113. Deisenroth, M. P., Rasmussen, C. E. & Fox, D. Learning to control a low-cost manipulator using data-efficient reinforcement learning. In *Robotics: Science and Systems VII* (RSS) (eds. Durrant-Whyte, H. et al.) 57–64 (Robotics: Science and Systems Foundation, 2011).

114. Ha, D. & Schmidhuber, J. Recurrent world models facilitate policy evolution. In *Proc. 32nd International Conference on Neural Information Processing Systems (NeurIPS)* (eds. Bengio, S. et al.) 2455–2467 (NIPS, 2018).

115. Battaglia, P. W. et al. Relational inductive biases, deep learning and graph networks. Preprint at https://arxiv.org/abs/1806.01261 (2018).

116. Andrychowicz, M. et al. Hindsight experience replay. In *Proc. Neural Information Processing Systems (NIPS)* (eds. Guyon I. et al.) 5048–5058 (NIPS, 2017); https://papers.nips.cc/paper/7090-hindsight-experience-replay.pdf

117. Schwartenbeck, P. et al. Computational mechanisms of curiosity and goal-directed exploration. *eLife* **8**, e41703 (2019).

118. Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proc. International Conference on Machine Learning (ICML)* (eds. Dy, J. & Krause, A.) 1861–1870 (PMLR, 2018).

119. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation and attention. *Neuron* **46**, 681–692 (2005).

120. Baldwin, D. A. & Kosie, J. E. How does the mind render streaming experience as events? *Top. Cogn. Sci.* **13**, 79–105 (2021).

Acknowledgements

We acknowledge funding from the DFG (projects IDEAS, LeCAREbot, TTR169, SPP 2134, RTG 1808 and EXC 2064/1), the Humboldt Foundation and Max Planck Research School IMPRS-IS.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00433-9.

Correspondence should be addressed to Manfred Eppe.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2022