Intrinsically Motivated Goal-Conditioned Reinforcement Learning: a Short Survey

Cédric Colas, Tristan Karch, Olivier Sigaud, Pierre-Yves Oudeyer

Intrinsically Motivated Goal-Conditioned Reinforcement Learning: a Short Survey

Cédric Colas CEDRIC.COLAS@INRIA.FR
INRIA and Univ. de Bordeaux; Bordeaux (FR)

Tristan Karch TRISTAN.KARCH@INRIA.FR
INRIA and Univ. de Bordeaux; Bordeaux (FR)

Olivier Sigaud OLIVIER.SIGAUD@UPMC.FR
Sorbonne Université; Paris (FR)

Pierre-Yves Oudeyer PIERRE-YVES.OUDEYER@INRIA.FR
INRIA; Bordeaux (FR) and ENSTA Paris Tech; Paris (FR)

Abstract

Building autonomous machines that can explore open-ended environments, discover possible interactions and autonomously build repertoires of skills is a general objective of artificial intelligence. Developmental approaches argue that this can only be achieved by autonomous and intrinsically motivated learning agents that can generate, select and learn to solve their own problems. In recent years, we have seen a convergence of developmental approaches, and developmental robotics in particular, with deep reinforcement learning (RL) methods, forming the new domain of *developmental machine learning*. Within this new domain, we review here a set of methods where deep RL algorithms are trained to tackle the developmental robotics problem of the *autonomous acquisition of open-ended repertoires of skills*. Intrinsically motivated goal-conditioned RL algorithms train agents to learn to represent, generate and pursue their own goals. The self-generation of goals requires the learning of compact goal encodings as well as their associated goal-achievement functions, which results in new challenges compared to traditional RL algorithms designed to tackle pre-defined sets of goals using external reward signals. This paper proposes a typology of these methods at the intersection of deep RL and developmental approaches, surveys recent approaches and discusses future avenues.

1. Introduction

Building autonomous machines that can explore large environments, discover interesting interactions and learn open-ended repertoires of skills is a long-standing goal in Artificial Intelligence. Humans are remarkable examples of this lifelong, open-ended learning. They learn to recognize objects and crawl as infants, then learn to ask questions and interact with peers as children. Across their lives, humans build a large repertoire of diverse skills from a virtually infinite set of possibilities. What is most striking, perhaps, is their ability to invent and pursue their own problems, using internal feedback to assess completion. We would like to build artificial agents able to demonstrate equivalent lifelong learning abilities.

We can think of two approaches to this problem: developmental approaches, in particular developmental robotics, and reinforcement learning (RL). Developmental robotics takes inspirations from artificial intelligence, developmental psychology and neuroscience to model cognitive processes in natural and artificial systems [Asada et al., 2009, Cangelosi and Schlesinger, 2015]. Following the idea that intelligence should be *embodied*, robots are often used to test learning models. Reinforcement learning, on the other hand, is the field interested in problems where agents learn to behave by experiencing the consequences of their actions under the form of rewards and costs. As a result, these agents are not explicitly taught, they need to learn to maximize cumulative rewards over time by trial-and-error [Sutton and Barto, 2018]. While developmental robotics is a field oriented towards answering particular questions around sensorimotor, cognitive and social development (e.g. how can we model language acquisition?), reinforcement learning is a field organized around a particular technical framework and set of methods.

Now powered by deep learning optimization methods leveraging the computational efficiency of large computational clusters, RL algorithms have recently achieved remarkable results including, but not limited to, learning to solve video games at a super-human level [Mnih et al., 2015], to beat chess and go world players [Silver et al., 2016], or even to control stratospheric balloons in the real world [Bellemare et al., 2020]. Although standard RL problems often involve a single agent learning to solve a unique task, RL researchers recently extended RL problems to *multi-goal* RL *problems*. Instead of pursuing a single goal, agents can now be trained to pursue goal distributions [Kaelbling, 1993, Sutton et al., 2011, Schaul et al., 2015]. As the field progresses, new goal representations emerge: from the specific goal states to the high-dimensional goal images or the abstract language-based goals [Luketina et al., 2019]. However, most approaches still fall short of modeling learning abilities of natural agents because they build agents that learn to solve predefined sets of tasks, via external and hand-defined learning signals.

Developmental robotics directly aims to model children learning and, thus, takes inspiration from the mechanisms underlying autonomous behaviors in humans. Most of the time, humans are not motivated by external rewards but spontaneously explore their environment to discover and learn about what is around them. This behavior seems to be driven by *intrinsic motivations* (IMs) a set of brain processes that motivate humans to explore for the mere purpose of experiencing novelty, surprise or learning progress [Berlyne, 1966, Gopnik et al., 1999, Kidd and Hayden, 2015, Oudeyer and Smith, 2016, Gottlieb and Oudeyer, 2018]. The integration of IMs into artificial agents thus seems to be a key step towards autonomous learning agents [Schmidhuber, 1991, Kaplan and Oudeyer, 2007]. In developmental robotics, this approach enabled sample efficient learning of high-dimensional motor skills in complex robotic systems [Santucci et al., 2020], including locomotion [Baranes and Oudeyer, 2013b, Martius et al., 2013], soft object manipulation [Rolf and Steil, 2013, Nguyen and Oudeyer, 2014], visual skills [Lonini et al., 2013] and nested tool use in real-world robots [Forestier et al., 2017]. Most of these approaches rely on *population-based* optimization algorithms, non-parametric models trained on datasets of (policy, outcome) pairs. Population-based algorithms cannot leverage automatic differentiation on large computational clusters, often demonstrate limited generalization capabilities and cannot easily handle high-dimension

perceptual spaces (e.g. images) without hand-defined input pre-processing. For these reasons, developmental robotics could benefit from new advances in deep RL.

Recently, we have been observing a convergence of these two fields, forming a new domain that we propose to call *developmental machine learning*, or *developmental artificial intelligence*. Indeed, RL researchers now incorporate fundamental ideas from the development robotics literature in their own algorithms, and reversely developmental robotics learning architecture are beginning to benefit from the generalization capabilities of deep RL techniques. These convergences can mostly be categorized in two ways depending on the type of intrinsic motivation (IMs) being used [Oudeyer and Kaplan, 2007]:

- **Knowledge-based IMs** compare the situations experienced by the agent to its current knowledge and expectations, and reward it for experiencing dissonance (or resonance). This family includes IMs rewarding prediction errors [Schmidhuber, 1991, Pathak et al., 2017], novelty [Burda et al., 2018, Bellemare et al., 2016], surprise [Achiam and Sastry, 2017], negative surprise [Berseth et al., 2019], learning progress [Lopes et al., 2012, Kim et al., 2020] or information gains [Houthooft et al., 2016], see a review in Linke et al. [2019]. This type of IMs is often used as an auxiliary reward to organize the exploration of agents in environments characterized by sparse rewards. It can also be used to facilitate the construction of world models [Lopes et al., 2012, Kim et al., 2020, Sekar et al., 2020].

- **Competence-based IMs**, on the other hand, reward agents to solve self-generated problems, to achieve self-generated goals. In this category, agents need to represent, select and master self-generated goals. As a result, competence-based IMs were often used to organize the acquisition of repertoires of skills in task-agnostic environments [Baranes and Oudeyer, 2010, 2013b, Santucci et al., 2016, Forestier and Oudeyer, 2016, Nair et al., 2018, Warde-Farley et al., 2018, Colas et al., 2019, Pong et al., 2019, Colas et al., 2020c]. Just like knowledge-based IMs, competence-based IMs organize the exploration of the world and, thus, might be used to facilitate learning in sparse reward settings [Colas et al., 2018] or train world models [Baranes and Oudeyer, 2013a, Chitnis et al., 2021].

RL algorithms using *knowledge-based* IMs leverage ideas from developmental robotics to solve standard RL problems. On the other hand, RL algorithms using competence-based IMs organize exploration around self-generated goals and can be seen as targeting a developmental robotics problem: the *open-ended and self-supervised acquisition of repertoires of diverse skills*. *Intrinsically Motivated Goal Exploration Processes* (IMGEP) is the family of algorithms that bake competence-based IMs into learning agents [Forestier et al., 2017]. IMGEP agents generate and pursue their own goals as a way to explore their environment, discover possible interactions and build repertoires of skills. This framework emerged from the field of developmental robotics [Oudeyer and Kaplan, 2007, Baranes and Oudeyer, 2009b, 2010, Rolf et al., 2010] and originally leveraged population-based learning algorithms (POP-IMGEP) [Baranes and Oudeyer, 2009a, 2013b, Forestier and Oudeyer, 2016, Forestier et al., 2017]. Recently, goal-conditioned RL agents were also endowed with the ability to generate and pursue their own goals and learn to achieve them via self-generated rewards. We argue that this set of methods form a sub-category of IMGEPs that we call goal-conditioned

IMGEPs or GC-IMGEPs. In contrast, one can refer to externally-motivated goal-conditioned RL agents as GC-EMGEPs.

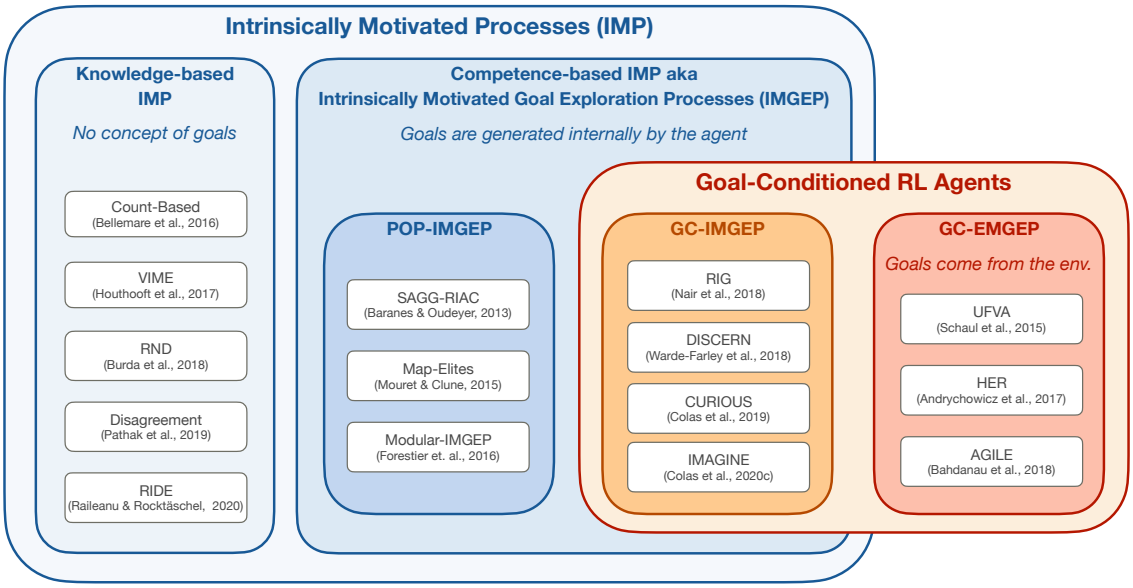


Figure 1: A typology of intrinsically-motivated and/or goal-conditioned RL approaches. POP-IMGEP, GC-IMGEP and GC-EMGEP refer to *population-based intrinsically motivated goal exploration processes*, *goal-conditioned* IMGEP and *goal-conditioned externally motivated goal exploration processes* respectively. POP-IMGEP, GC-IMGEP and GC-IMGEP all represent goals, but knowledge-based IMs do not. While IMGEPs (POP-IMGEP and GC-IMGEP) generate their own goals, GC-EMGEPs require externally-defined goals. This paper is interested in GC-IMGEPs, the intersection of *goal-conditioned* RL *agents* and *intrinsically motivated processes* that is, the set of methods that train learning agents to generate and pursue their own goals with goal-conditioned RL algorithms.

This paper proposes a formalization and a review of the GC-IMGEP algorithms at the convergence of RL methods and developmental robotics objectives. Figure 1 proposes a visual representation of intrinsic motivations approaches (knowledge-based IMs vs competence-based IMs or IMGEPs) and goal-conditioned RL (externally vs intrinsically motivated). Their intersection is the family of algorithms that train agents to generate and pursue their own goals by training goal-conditioned policies. We define goals as the combination of a compact goal representation and a goal-achievement function to measure progress. This definition highlights new challenges for autonomous learning agents. While traditional RL agents only need to learn to achieve goals, GC-IMGEP agents also need to learn to represent them, to generate them and to measure their own progress. After learning, the resulting goal-conditioned policy and its associated goal space form a *repertoire of skills*, a repertoire of behaviors that the agent can represent and control. We believe organizing past GC-RL methods at the

convergence of developmental robotics and RL into a common classification and towards the resolution of a common problem will help organize future research.

Scope of the survey We are interested in algorithms from the GC-IMGEP family as algorithmic tools to enable agents to acquire repertoires of skills in an open-ended and self-supervised setting. Externally motivated goal-conditioned RL approaches do not enable agents to generate their own goals and thus cannot be considered IMGEPs. However, these approaches can often be converted into GC-IMGEPs by integrating the goal generation process within the agent. For this reason, we include some GC-EMGEPs approaches when they present interesting mechanisms that can directly be leveraged in intrinsically motivated agents.

What is not covered. This survey will not discuss some related but distinct approaches including multi-task RL [Caruana, 1997], RL with auxiliary tasks [Riedmiller et al., 2018, Jaderberg et al., 2016] and RL with knowledge-based IMs [Bellemare et al., 2016, Pathak et al., 2017, Burda et al., 2018]. None of these approaches do represent goals or see the agent's behavior affected by goals. The subject of intrinsically motivated goal-conditioned RL also relates to *transfer learning* and *curriculum learning*. This survey does not cover transfer learning approaches, but interested readers can refer to Taylor and Stone [2009]. We will discuss automatic curriculum learning approaches that organize the generation of goals according to the agent's abilities in Section 6 but, for a broader picture on the topic, readers can refer to the recent review Portelas et al. [2020].

Survey organization We first formalize the notion of *goals* and the problem of the *open-ended and self-supervised acquisition of skill repertoires*, building on the formalization of the RL and multi-goal RL problems. After presenting a definition of the GC-IMGEP family, we organize the surveyed literature along three axes: 1) What are the different types of goal representations? (Section 4); 2) How can we learn goal representations? (Section 5) and 3) How can we prioritize goal selection? (Section 6). From this coherent picture of the literature, we identify properties of what humans call *goals* that were not addressed in the surveyed approaches. This serves as the basis for a discussion of potential future avenues for the design of new GC-IMGEP approaches.

2. Self-Supervised Acquisition of Skills Repertoires with Deep RL

This section builds towards the definition of our main objective: *enabling agents to acquire repertoires of skills in an open-ended and self-supervised setting*. To this end, we present the traditional reinforcement learning (RL) problem, propose a formal generalized definition of *goals* in the context of RL, and define the multi-goal RL problem which will serve as a basis to introduce our problem.

2.1 The Reinforcement Learning Problem.

In a reinforcement learning (RL) problem, the agent learns to perform sequences of actions in an environment so as to maximize some notion of cumulative reward [Sutton and Barto, 2018]. RL problems are commonly framed as Markov Decision Processes (MDPs): $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0, R\}$ [Sutton and Barto, 2018]. The agent and its environment, as well as

their interaction dynamics are defined by the first components $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0\}$, where $s \in \mathcal{S}$ describes the current state of the agent-environment interaction and ρ_0 is the distribution over initial states. The agent can interact with the environment through actions $a \in \mathcal{A}$. Finally, the dynamics are characterized by the transition function $\mathcal{T}$ that dictates the distribution of the next state s' from the current state and action $\mathcal{T}(s' \mid s,\ a)$. The objective of the agent in this environment is defined by the remaining component of the MDP: R. R is the reward function, it computes a reward for any transition: $R(s, a, s')$. Note that, in a traditional RL problem, the agent only receives the rewards corresponding to the transitions it experiences but does not have access to the function itself. The objective of the agent is to maximize the cumulative reward computed over complete episodes. The aggregation of rewards over time is often modulated by the discounting function Γ so that the total aggregated reward from time step t, R_t^{tot}, is computed as $R_{\text{tot}} = \sum_{i=t}^{\infty} \Gamma(s_i, a_i, i) \times R(s_{i-1}, a_i, s_i)$. Γ is usually an exponentially-decreasing function of time: $\Gamma(t) = \gamma^t$ with a constant discount factor $\gamma \in [0, 1[$. Each instance of an MDP implements an RL problem, also called a *task*.

2.2 Defining *Goals* for Reinforcement Learning

This section takes inspiration from the notion of *goal* in psychological research to inform the formalization of *goals* for reinforcement learning.

***Goals* in psychological research.** Working on the origin of the notion *goal* and its use in past psychological research, Elliot and Fryer [2008] propose a general definition:

> *A goal is a cognitive representation of a future object that the organism is committed to approach or avoid* [Elliot and Fryer, 2008].

Because goals are *cognitive representations*, only animate organisms that represent goals qualify as goal-conditioned. Because this representation relates to a *future object*, goals are cognitive imagination of future possibilities: goal-conditioned behavior is proactive, not reactive. Finally, organisms *commit* to their goal, their behavior is thus influenced directly by this cognitive representation.

Generalized *goals* for reinforcement learning. RL algorithms seem to be a good fit to train such goal-conditioned agents. Indeed, RL algorithms train learning agents (*organisms*) to maximize (*approach*) a cumulative (*future*) reward (*object*). In RL, goals can be seen as a set of *constraints* on one or several consecutive states that the agent seeks to respect. These constraints can be very strict and characterize a single target point in the state space (e.g. image-based goals) or a specific sub-space of the state space (e.g. target x-y coordinate in a maze, target block positions in manipulation tasks). They can also be more general, when expressed by language for example (e.g. '*find a red object or a wooden one*'). To represent these goals, RL agents must be able to 1) have a compact representation of them and 2) assess their achievement. This is why we propose the following formalization for RL goals: each goal is a $g = (z_g, R_{\mathcal{G}})$ pair where z_g is a compact *goal parameterization* or *goal embedding* and $R_{\mathcal{G}}$ is a *goal-achievement* function – also called *goal-parameterized* or *goal-conditioned reward function* – that measures progress towards goal achievement and is shared across goals. We note $R_g(\cdot) = R_{\mathcal{G}}(\cdot \mid z_g)$ the reward function measuring the achievement of goal g. With this definition we can express a diversity of goals, see Section 4 and Table 1.

The goal-achievement function and the goal-conditioned policy both assign *meaning* to a goal. The former defines what it means to achieve the goal, it describes how the world looks like when it is achieved. The latter characterizes the process by which this goal can be achieved, what the agent needs to do to achieve it. In this search for the meaning of a goal, the goal embedding can be seen as the map: the agent follows this map and via the two functions above, experiences the meaning of the goal.

2.3 The Multi-Goal Reinforcement Learning Problem.

By replacing the unique reward function R by the space of reward functions $\mathcal{R}_\mathcal{G}$, RL problems can be extended to handle multiple goals: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0, \mathcal{R}_\mathcal{G}\}$. The term *goal* should not be mistaken for the term *task*, which refers to a particular MDP instance. As a result, *multi-task* RL refers to RL algorithms that tackle a set of MDPs that can differ by any of their components (e.g. $\mathcal{T}$, R, $\mathcal{S}_0$, etc.). The *multi-goal* RL problem can thus be seen as the particular case of the multi-task RL problem where MDPs differ by their reward functions. In the standard multi-goal RL problem, the set of goals – and thus the set of reward functions – is pre-defined by engineers. The experimenter sets goals to the agent, and provides the associated reward functions. GC-RL is the set of agents targeting multi-goal problems.

2.4 The Intrinsically Motivated Skills Acquisition Problem

In the *intrinsically motivated skills acquisition problem*, the agent is set in an open-ended environment without any pre-defined goal and needs to acquire a repertoire of skills. Here, a skill is defined as the association of a goal embedding z_g and the policy to reach it Π_g. A repertoire of skills is thus defined as the association of a repertoire of goals $\mathcal{G}$ with a goal-conditioned policy trained to reach them $\Pi_\mathcal{G}$. The intrinsically motivated skills acquisition problem can now be modeled by a reward-free MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0\}$ that only characterizes the agent, its environment and their possible interactions. Just like children, agents need to represent and generate their own goals, to measure their progress and to learn to achieve them.

Evaluation It is often straightforward to evaluate externally motivated agents as the set of possible interactions is known in advance. One can, for instance, easily measure generalization by computing the agent's success rate on a held-out set of testing goals. Similarly, exploration can be estimated by several metrics such as the count of task-specific state visitations. In contrast, intrinsically motivated agents evolve in open-ended environments and learn to represent and form their own set of skills. In this context, the space of possible behaviors might quickly become intractable for the experimenter, which is maybe the most interesting feature of such agents. For these reasons, designing evaluation protocols is not trivial.

Interestingly, the question of how to evaluate intrinsically motivated agents is quite similar to the question of how to evaluate self-supervised learning systems such as Generative Adversarial Networks (GAN) [Goodfellow et al., 2014] or self-supervised language models [Devlin et al., 2019, Brown et al., 2020]. In both cases, learning is *task-agnostic* and it is often hard to compare models in terms of their outputs (e.g. comparing the quality of GAN

output images, or comparing output repertoires of skills in intrinsically motivated agents). Let us list some approaches to evaluate such models:

- **Measuring exploration:** one can compute task-agnostic exploration proxies such as the entropy of the visited state distribution, or measures of state coverage (e.g. coverage of the high-level x-y state space in mazes) [Florensa et al., 2018]. Exploration can also be measured as the number of interactions from a set of *interesting* interactions defined subjectively by the experimenter (e.g. interactions with objects in Colas et al. [2020c].
- **Measuring generalization:** The experimenter can subjectively define a set of relevant target goals and prevent the agent from training on them. Evaluating agents on this held-out set at test time provides a measure of generalization [Ruis et al., 2020], although it is biased towards what the experimenter assesses as *relevant* goals.
- **Measuring transfer learning:** Here, we view the intrinsically motivated exploration of the environment as a pre-training phase to bootstrap learning in a subsequent downstream task. In the downstream task, the agent is trained to achieve externally-defined goals. We report its performance and learning speed on these goals. This is akin to the evaluation of self-supervised language models, where the reported metrics evaluate performance in various downstream tasks (e.g. Brown et al. [2020]).
- **Opening the black-box:** Investigating internal representations learned during intrinsically motivated exploration is often informative. One can investigate properties of the goal generation system (e.g. does it generate out-of-distribution goals?), investigate properties of the goal embeddings (e.g. are they disentangled?). One can also look at the learning trajectories of the agents across learning, especially when they implement their own curriculum learning (e.g. Florensa et al. [2018], Colas et al. [2019], Pong et al. [2019], Akakzia et al. [2020]).
- **Measuring robustness:** Autonomous learning agents evolving in open-ended environment should be robust to a variety of properties than can be found in the real-world. This includes very large environments, where possible interactions might vary in terms of difficulty (trivial interactions, impossible interactions, interactions whose result is stochastic thus prevent any learning progress). Environments can also include distractors (e.g. non-controllable objects) and various forms of non-stationarity. Evaluating learning algorithms in various environments presenting each of these properties allows to assess their ability to solve the corresponding challenges.

3. Intrinsically Motivated Goal Exploration Processes with Goal-Conditioned Policies

Until recently, the IMGEP family was powered by population-based algorithms (POP-IMGEP). The emergence of goal-conditioned RL approaches that generate their own goals gave birth to a new type of IMGEPs: the goal-conditioned IMGEPs (GC-IMGEP). We build from traditional RL algorithms and goal-conditioned RL algorithms towards intrinsically motivated goal-conditioned RL algorithms (GC-IMGEP).

3.1 Reinforcement Learning Algorithms for the RL Problem

RL methods use transitions collected via interactions between the agent and its environment $(s, a, s', R(s, a, s'))$ to train a *policy* π: a function generating the next action a based on the current state s so as to maximize a cumulative function of rewards. The deep RL family (DRL) leverages deep neural networks as function approximators to represent policies, reward and value functions. Other set of methods can also be used to train policies. Imitation Learning (IL) leverages demonstrations, i.e. transitions collected by another entity (e.g. Ho and Ermon [2016], Hester et al. [2018]). Evolutionary Computing (EC) is a group of population-based approaches where populations of policies are trained to maximize cumulative rewards using episodic samples (e.g. Lehman and Stanley [2011], Mouret and Clune [2015], Forestier et al. [2017], Colas et al. [2020d]). Finally, model-based RL approaches can be used to 1) learn a model of the transition function $\mathcal{T}$ and 2) perform planning towards reward maximization in that model (e.g. Chua et al. [2018], Charlesworth and Montana [2020]). This removes the need to represent a policy explicitly. In this survey, we focus on DRL methods that represent a policy. Most of the goal-related mechanisms discussed in this paper can be directly applied to other optimization methods as well (e.g. IL or model-based RL).

3.2 Goal-Conditioned RL Algorithms

Goal-conditioned agents see their behavior affected by the goal they pursue. This is formalized via goal-conditioned policies, that is policies which produce actions based on the environment state and the agent's current goal: $\Pi : \mathcal{S} \times \mathcal{Z}_\mathcal{G} \rightarrow \mathcal{A}$, where $\mathcal{Z}_\mathcal{G}$ is the space of goal embeddings corresponding to the goal space $\mathcal{G}$ [Schaul et al., 2015]. Note that ensembles of policies can also be formalized this way, via a meta-policy Π that retrieves the particular policy from a one-hot goal embedding z_g (e.g. Kaelbling [1993], Sutton et al. [2011]). The idea of using a unique RL agent to target multiple goals dates back to Kaelbling [1993]. Later, the HORDE architecture proposed to use interaction experience to update one value function per goal, effectively transferring to all goals the knowledge acquired while aiming at a particular one [Sutton et al., 2011]. In these approaches, one policy is trained for each of the goals and the data collected by one can be used to train others. Building on these early results, Schaul et al. [2015] introduced Universal Value Function Approximators (UVFA). They proposed to learn a unique goal-conditioned value function and goal-conditioned policy to replace the set of value functions learned in HORDE. Using neural networks as function approximators, they showed that UVFAs enable transfer between goals and demonstrate strong generalization to new goals. The idea of *hindsight learning* further improves knowledge transfer between goals [Kaelbling, 1993, Andrychowicz et al., 2017]. Learning by hindsight, agents can reinterpret a past trajectory collected while pursuing a given goal in the light of a new goal. By asking themselves, *what is the goal for which this trajectory is optimal?*, they can use the originally failed trajectory as an informative trajectory to learn about another goal, thus making the most out of every trajectory [Eysenbach et al., 2020]. This ability dramatically increases the sample efficiency of goal-conditioned algorithms and is arguably an important driver of the recent interest in goal-conditioned RL approaches.

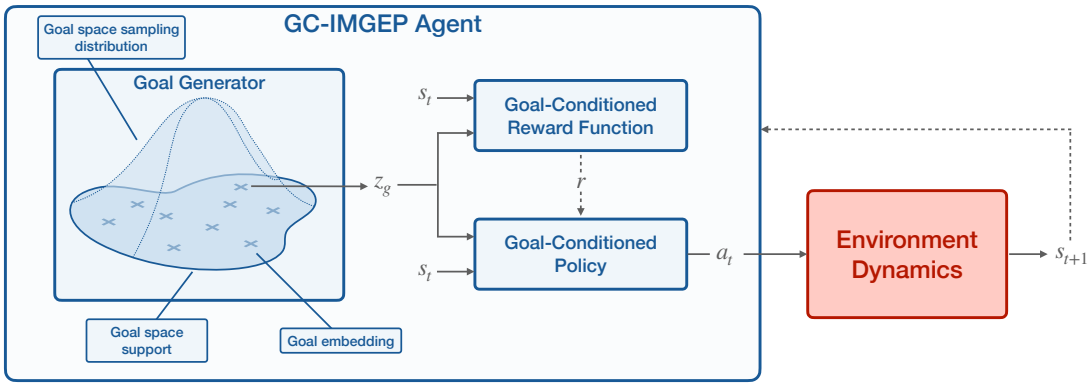
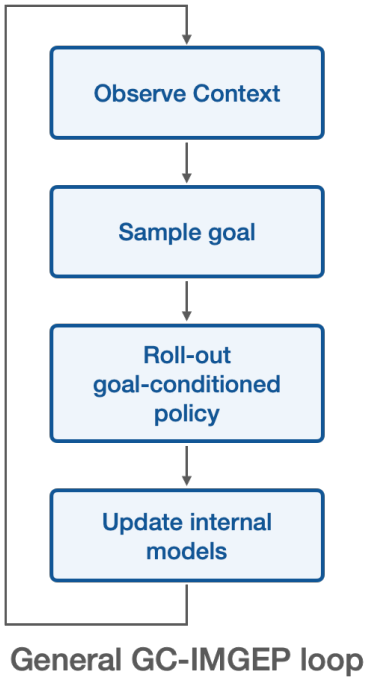


Figure 2: Representation of the different learning modules in a GC-IMGEP algorithm. In contrast, externally motivated goal exploration processes (GC-EMGEPs) only train the goal-conditioned policy and assume *external* goal generator and goal-conditioned reward function.

Algorithm 1 Goal-Conditioned IMGEP

Require: environment $\mathcal{E}$

Initialize empty memory $\mathcal{M}$,
goal-conditioned policy $\Pi_\mathcal{G}$, goal-conditioned reward $R_\mathcal{G}$,
goal space $\mathcal{Z}_\mathcal{G}$, goal sampling policy GS.
loop
  ▷ *Observe context*
  Get initial state: $s_0 \leftarrow \mathcal{E}$.reset()
  ▷ *Sample goal*
  Sample goal embedding $z_g = GS(s_0, \mathcal{Z}_\mathcal{G})$.
  ▷ *Roll-out goal-conditioned policy*
  Execute a roll-out with $\Pi_g = \Pi_\mathcal{G}(\cdot \mid z_g)$
  Store collected transitions $\tau = (s, a, s')$ in $\mathcal{M}$.
  ▷ *Update internal models*
  Sample a batch of B transitions: $\mathcal{M} \sim \{(s, a, s')\}_B$.
  Perform Hindsight Relabelling $\{(s, a, s', z_g)\}_B$.
  Compute internal rewards $r = R_\mathcal{G}(s, a, s' \mid z_g)$.
  Update policy Π via RL on $\{(s, a, s', z_g, r)\}_B$.
  Update goal representations $\mathcal{Z}_\mathcal{G}$.
  Update goal-conditioned reward function $R_\mathcal{G}$.
  Update goal sampling policy GS.
return $\Pi_\}, R_\mathcal{G}, \mathcal{Z}_\mathcal{G}$



3.3 GC-IMGEP

GC-IMGEP are intrinsically motivated versions of goal-conditioned RL algorithms. They need to be equipped with mechanisms to represent and generate their own goals in order to solve the intrinsically motivated skills acquisition problem, see Figure 2. Concretely, this means that, in addition to the goal-conditioned policy, they need to learn: 1) to represent goals g

by compact embeddings z_g; 2) to represent the support of the goal distribution, also called *goal space* $\mathcal{Z}_\} = \{z_g\}_{g\in\mathcal{G}}$; 3) a goal distribution from which targeted goals are sampled $\mathcal{D}(z_g)$; 4) a goal-conditioned reward function $\mathcal{R}_\mathcal{G}$. Algorithm 1 details the pseudo-code of GC-IMGEP algorithms.

4. A Typology of Goal Representations in the Literature

Now that we defined the problem of interest and the overall framework to tackle it, we can start reviewing relevant approaches from the literature and how they fit in this framework. This section presents a typology of the different kinds of goal representations found in the literature. Each goal is represented by a pair: 1) a *goal embedding* and 2) a goal-conditioned reward function.

4.1 Goals as choices between multiple objectives

Goals can be expressed as a list of different objectives the agent can choose from.

Goal embedding. In that case, goal embeddings z_g are one-hot encodings of the current objective being pursued among the N objectives available. z_g^i is the i^{th} one-hot vector: $z_g^i = (\mathbb{1}_{j=i})_{j=[1..N]}$. This is the case in Oh et al. [2017], Mankowitz et al. [2018], Codevilla et al. [2018].

Reward function. The goal-conditioned reward function is a collection of N distinct reward functions $R_\mathcal{G}(\cdot) = R_i(\cdot)$ if $z_g = z_g^i$. In Mankowitz et al. [2018] and Chan et al. [2019], each reward function gives a positive reward when the agent reaches the corresponding object: reaching guitars and keys in the first case, monsters and torches in the second.

4.2 Goals as target features of states

Goals can be expressed as target features of the state the agent desires to achieve.

Goal embedding. In this scenario, a state representation function φ maps the state space to an embedding space $\mathcal{Z} = \varphi(\mathcal{S})$. Goal embeddings z_g are target points in $\mathcal{Z}$ that the agent should reach. In manipulation tasks, z_g can be target block coordinates [Andrychowicz et al., 2017, Nair et al., 2017, Plappert et al., 2018, Colas et al., 2019, Fournier et al., 2019, Blaes et al., 2019, Li et al., 2019, Lanier et al., 2019, Ding et al., 2019]. In navigation tasks, z_g can be target agent positions (e.g. in mazes Schaul et al. [2015], Florensa et al. [2018]). Agent can also target image-based goals. In that case, the state representation function φ is usually implemented by a generative model trained on experienced image-based states and goal embeddings can be sampled from the generative model or encoded from real images [Zhu et al., 2017, Pong et al., 2019, Nair et al., 2018, Warde-Farley et al., 2018, Codevilla et al., 2018, Florensa et al., 2019, Venkattaramanujam et al., 2019, Lynch et al., 2020, Lynch and Sermanet, 2020, Nair et al., 2020, Kovač et al., 2020].

Reward function. For this type of goals, the reward function $R_\mathcal{G}$ is based on a distance metric D. One can define a dense reward as inversely proportional to the distance between features of the current state and the target goal embedding: $R_g = R_\mathcal{G}(s|z_g) = -\alpha \times$

$D(\varphi(s),\ z_g)$ (e.g. Nair et al. [2018]). The reward can also be sparse: positive whenever that distance falls below a pre-defined threshold: $R_{\mathcal{G}}(s|z_g) = 1$ if $D(\varphi(s),\ z_g) < \epsilon$, 0 otherwise.

4.3 Goals as abstract binary problems

Some goals cannot be expressed as target state features but can be represented by *binary problems*, where each goal expresses as set of constraint on the state (or trajectory) such that these constraints are either verified or not (binary goal achievement).

Goal embeddings. In binary problems, goal embeddings can be any expression of the set of constraints that the state should respect. Akakzia et al. [2020], Ecoffet et al. [2020] both propose a pre-defined discrete state representation. These representations lie in a finite embedding space so that goal completion can be asserted when the current embedding $\varphi(s)$ equals the goal embedding z_g. Another way to express sets of constraints is via language-based predicates. A sentence describes the constraints expressed by the goal and the state or trajectory either verifies them, or does not [Hermann et al., 2017, Chan et al., 2019, Jiang et al., 2019, Bahdanau et al., 2019, 2018, Hill et al., 2019, Cideron et al., 2019, Colas et al., 2020c, Lynch and Sermanet, 2020], see [Luketina et al., 2019] for a recent review. Language can easily characterize *generic goals* such as "*grow any blue object*" [Colas et al., 2020c], *relational goals* like "*sort objects by size*" [Jiang et al., 2019], "*put the cylinder in the drawer*" [Lynch and Sermanet, 2020] or even *sequential goals* "*Open the yellow door after you open a purple door*" [Chevalier-Boisvert et al., 2019]. When goals can be expressed by language sentences, goal embeddings z_g are usually language embeddings learned jointly with either the policy or the reward function. Note that, although RL goals always express constraints on the state, we can imagine *time-extended goals* where constraints are expressed on the trajectory (see a discussion in Section 7.1).

Reward function. The reward function of a binary problem can be viewed as a binary classifier that evaluates whether state s (or trajectory τ) verifies the constraints expressed by the goal semantics (positive reward) or not (null reward). This binary classification setting has directly been implemented as a way to learn language-based goal-conditioned reward functions $R_g(s \mid z_g)$ in Bahdanau et al. [2019] and Colas et al. [2020c]. Alternatively, the setup described in Colas et al. [2020a] proposes to turn binary problems expressed by language-based goals into goals as specific target features. To this end, they train a language-conditioned goal generator that produces specific target features verifying constraints expressed by the binary problem. As a result, this setup can use a distance-based metric to evaluate the fulfillment of a binary goal.

4.4 Goals as a Multi-Objective Balance

Some goals can be expressed, not as desired regions of the state or trajectory space but as more general objectives that the agent should maximize. In that case, goals can parameterize a particular mixture of multiple objectives that the agent should maximize.

Goal embeddings. Here, goal embeddings are simply sets of weights balancing the different objectives $z_g = (\beta_i)_{i=[1..N]}$ where β_i is the weights applied to objective i and N is the number of objectives. Note that, when $\beta_j = 1$ and $\beta_i = 0,\ \forall i \neq j$, the agent can

decide to pursue any of the objective alone. In *Never Give Up*, for example, RL agents are trained to maximize a mixture of extrinsic and intrinsic rewards [Badia et al., 2020b]. The agent can select the mixing parameter β that can be viewed as a goal. Building on this approach, *Agent57* adds a control of the discount factor, effectively controlling the rate at which rewards are discounted as time goes by [Badia et al., 2020a].

Reward function. When goals are represented as a balance between multiple objectives, the associated reward function cannot be represented neither as a distance metric, nor as a binary classifier. Instead, the agent needs to maximize a convex combination of the objectives: $R_g(s) = \sum_{i=1}^{N} \beta_g^i R^i(s)$ where R^i is the i^{th} of N objectives and $z_g = \beta = \beta_i^g \mid_{i \in [1..N]}$ is the set of weights.

4.5 Goal-Conditioning

Now that we described the different types of goal embeddings found in the literature, remains the question of how to condition the agent's behavior – i.e. the policy – on them. Originally, the UVFA framework proposed to concatenate the goal embedding to the state representation to form the policy input. Recently, other mechanisms have emerged. When language-based goals were introduced, Chaplot et al. [2017] proposed the *gated-attention* mechanism where the state features are linearly scaled by attention coefficients computed from the goal representation $\varphi(z_g)$: input $= s \odot \varphi(z_g)$, where $\odot$ is the Hadamard product. Later, the Feature-wise Linear Modulation (FILM) approach [Perez et al., 2018] generalized this principle to affine transformations: input $= s \odot \varphi(z_g) + \psi(z_g)$. Alternatively, Andreas et al. [2016] came up with Neural Module Networks (NMN), a mechanism that leverages the linguistic structure of goals to derive a symbolic program that defines how states should be processed [Bahdanau et al., 2019].

4.6 Conclusion

This section presented a diversity of goal representations, corresponding to a diversity of reward functions architectures. However, we believe this represents only a small fraction of the diversity of goal types that humans pursue. Section 7 discusses other goal representations that RL algorithms could target.

5. How to Learn Goal Representations?

The previous section discussed various types of goal representations. Intrinsically motivated agents actually need to learn these goal representations. While individual goals are represented by their embeddings and associated reward functions, representing multiple goals also requires the representation of the *support* of the goal space, i.e. how to represent the collection of *valid goals* that the agent can sample from, see Figure 2. This section reviews different approaches from the literature.

5.1 Assuming Pre-Defined Goal Representation

Most approaches tackle the multi-goal RL problem, where goal spaces and associated rewards are pre-defined by the engineer and are part of the task definition. Navigation and manipu-

lation tasks, for example, pre-define goal spaces (e.g. target agent position and target block positions respectively) and use the Euclidean distance to compute rewards [Andrychowicz et al., 2017, Nair et al., 2017, Plappert et al., 2018, Colas et al., 2019, Li et al., 2019, Lanier et al., 2019, Ding et al., 2019, Schaul et al., 2015, Florensa et al., 2018]. Akakzia et al. [2020], Ecoffet et al. [2020] hand-define abstract state representation and provide positive rewards when these match target goal representations. This falls short of solving the intrinsically motivated goal exploration problem. The next sub-section investigates how goal representations can be learned.

5.2 Learning Goal Embeddings

Some approaches assume the pre-existence of a goal-conditioned reward function, but learn to represent goals by learning goal embeddings. This is the case of language-based approaches, which receive rewards from the environment (thus are GC-EMGEP), but learn goal embeddings jointly with the policy during policy learning [Hermann et al., 2017, Chan et al., 2019, Jiang et al., 2019, Bahdanau et al., 2018, Hill et al., 2019, Cideron et al., 2019, Lynch and Sermanet, 2020]. When goals are target images, goal embeddings can be learned via generative models of states, assuming the reward to be a fixed distance metric computed in the embedding space [Nair et al., 2018, Pong et al., 2019, Florensa et al., 2019, Nair et al., 2020].

5.3 Learning the Reward Function

A few approaches go even further and learn their own goal-conditioned reward function. Bahdanau et al. [2019], Colas et al. [2020c] learn language-conditioned reward functions from an expert dataset or from language descriptions of autonomous exploratory trajectories respectively. However, the AGILE approach from Bahdanau et al. [2019] does not generate its own goals. In the domain of image-based goals, Venkattaramanujam et al. [2019], Hartikainen et al. [2019] learn a distance metric estimating the square root of the number of steps required to move from any state s_1 to any s_2 and generates internal signals to reward agents for getting closer to their target goals. Warde-Farley et al. [2018] learn a similarity metric in the space of controllable aspects of the environment that is based on a mutual information objective between the state and the goal state s_g. [Gregor et al., 2016, Eysenbach et al., 2018] train agents to develop a set of skills leading to maximally different areas of the state space. Agents are rewarded for experiencing states that are easy to discriminate, while a discriminator is trained to better infer the skill z_g from the visited states. Wu et al. [2018] compute a distance metric representing the ability of the agent to reach one state from another using the Laplacian of the transition dynamics graph, where nodes are states and edges are actions. More precisely, they use the eigenvectors of the Laplacian matrix of the graph given by the states of the environment as basis to compute the L2 distance towards a goal configuration. All these methods set their own goals and learn their own goal-conditioned reward function. For these reasons, they can be considered as complete intrinsically motivated goal-conditioned RL algorithms.

5.4 Learning the Support of the Goal Distribution

The previous sections reviewed several approaches to learn goal embeddings and reward functions. To represent collections of goals, one also needs to represent the support of the goal distribution, which embeddings correspond to valid goals and which do not. Most approaches consider a pre-defined, bounded goal space in which any point is a valid goal (e.g. target positions within the boundaries of a maze, target block positions within the gripper's reach) [Andrychowicz et al., 2017, Nair et al., 2017, Plappert et al., 2018, Colas et al., 2019, Li et al., 2019, Lanier et al., 2019, Ding et al., 2019, Schaul et al., 2015]. However, not all approaches assume pre-defined goal spaces. Some approaches use the set of previously experienced representations to form the support of the goal distribution [Veeriah et al., 2018, Akakzia et al., 2020, Ecoffet et al., 2020]. In Florensa et al. [2018], a Generative Adversarial Network (GAN) is trained on past representations of states ($\varphi(s)$) to model a distribution of goals and thus its support. In the same vein, approaches handling image-based goals usually train a generative model of image states based on Variational Auto-Encoders (VAE) to model goal distributions and support [Nair et al., 2018, Pong et al., 2019, Nair et al., 2020]. In both cases, valid goals are the one generated by the generative model. We saw that the support of valid goals can be pre-defined, a simple set of past representations or approximated by a generative model trained on these. In all cases, the agent can only sample goals *within* the convex hull of previously encountered goals (in representation space). We say that goals are *within* training distribution. This drastically limits exploration and the discovery of new behaviors. Children, on the other hand, can imagine creative goals. Pursuing these goals is thought to be the main driver of exploratory play in children [Chu and Schulz, 2020]. This is made possible by the compositionality of language, where sentences can easily be combined to generate new ones. The IMAGINE algorithm leverages the creative power of language to generate such *out-of-distribution* goals [Colas et al., 2020c]. The support of valid goals is extended to any combination of language-based goals experienced during training. They show that this mechanism augments the generalization and exploration abilities of learning agents. In Section 6, we discuss how agents can learn to adapt the goal sampling distribution to maximize the learning progress of the agent.

5.5 Conclusion

This section presented how previous approaches tackled the problem of learning goal representations. While most approaches rely on pre-defined goal embeddings and/or reward functions, some approaches proposed to learn internal reward functions and goal embeddings jointly.

6. How to Prioritize Goal Selection?

Intrinsically motivated goal-conditioned agents also need to select their own goals. While goals can be generated by uninformed sampling of the goal space, agents can benefit from mechanisms optimizing goal selection. In practice, this boils down to the automatic adaptation of the goal sampling distribution as a function of the agent performance.

6.1 Automatic Curriculum Learning for Goal Selection

In real-world scenarios, goal spaces can be too large for the agent to master all goals in its lifetime. Some goals might be trivial, others impossible. Some goals might be reached by chance sometimes, although the agent cannot make any progress on them. Some goals might be reachable only after the agent mastered more basic skills. For all these reasons, it is important to endow intrinsically motivated agents learning in open-ended scenarios with the ability to optimize their goal selection mechanism. This ability is a particular case of Automatic Curriculum Learning applied for goal selection: mechanisms that organize goal sampling so as to maximize the long-term performance improvement (distal objective). As this objective is usually not directly differentiable, curriculum learning techniques usually rely on a proximal objective. In this section, we look at various proximal objectives used in automatic curriculum learning (ACL) strategies to organize goal selection. Interested readers can refer to Portelas et al. [2020], which present a broader review of ACL methods for RL.

Intermediate difficulty. Intermediate difficulty has been used as a proxy for long-term performance improvement, following the intuition that focusing on goals of intermediate difficulty results in short-term learning progress that will eventually turn into long-term performance increase. GOALGAN assigns feasibility scores to goals as the proportion of time the agents successfully reaches it [Florensa et al., 2018]. Based on this data, a GAN is trained to generate goals of intermediate difficulty, whose feasibility scores are contained within an intermediate range. Sukhbaatar et al. [2017] and Campero et al. [2020] train a goal policy with RL to propose challenging goals to the RL agent. The goal policy is rewarded for setting goals that are neither too easy nor impossible. Zhang et al. [2020] select goals that maximize the disagreement in an ensemble of value functions. Value functions agree when the goals are too easy (the agent is always successful) or too hard (the agent always fails) but disagree for goals of intermediate difficulty.

Uniform feasibility. Racanière et al. [2019] generalize the idea of intermediate difficulty and train a goal generator to sample goals of uniform feasibility. This approach seems to lead to better stability and improved performance on more complex tasks compared to GOALGAN [Florensa et al., 2018].

Novelty - diversity. Pong et al. [2019], Warde-Farley et al. [2018], Pitis et al. [2020] all bias the selection of goals towards sparse areas of the goal space. For this purpose, they train density models in the goal space. While Warde-Farley et al. [2018], Pong et al. [2019] aim at a uniform coverage of the goal space (diversity), Pitis et al. [2020] skew the distribution of selected goals even more, effectively maximizing novelty. Kovač et al. [2020] proposed to enhance these methods with a goal sampling prior focusing goal selection towards controllable areas of the goal space.

Short-term learning progress. Some approaches estimate the learning progress of the agent in different regions of the goal space and bias goal sampling towards areas of high absolute learning progress using bandit algorithms [Colas et al., 2019, Blaes et al., 2019, Fournier et al., 2018, 2019, Akakzia et al., 2020]. Colas et al. [2019] and Akakzia et al. [2020] organize goals into modules and compute average LP measures over modules. Fournier et al. [2018] defines goals as a discrete set of precision requirements in a reaching task and

computes LP for each requirement value. The use of absolute LP enables agents to focus back on goals for which performance decreases (due to perturbations or forgetting). Akakzia et al. [2020] introduces the success rate in the value optimized by the bandit: $v = (1 - \text{SR}) \times \text{LP}$, so that agents favor goals with high absolute LP and low competence.

6.2 Hierarchical Reinforcement Learning for Goal Sequencing.

Hierarchical reinforcement learning (HRL) can be used to guide the sequencing of goals [Dayan and Hinton, 1993, Sutton et al., 1998, 1999, Precup, 2000]. In HRL, a high-level policy is trained via RL or planning to generate sequence of goals for a lower level policy so as to maximize a higher-level reward. This allows to decompose tasks with long-term dependencies into simpler sub-tasks. Low-level policies are implemented by traditional goal-conditioned RL algorithms [Levy et al., 2018, Röder et al., 2020] and can be trained independently from the high-level policy [Kulkarni et al., 2016, Frans et al., 2017] or jointly [Levy et al., 2018, Nachum et al., 2018, Röder et al., 2020]. Most approaches consider hand-defined spaces for the sub-goals (e.g. positions in a maze). Recent approaches propose to use the state space directly [Nachum et al., 2018] or to learn the sub-goal space (e.g. Vezhnevets et al. [2017], or with generative model of image states in Nasiriany et al. [2019]).

7. Future Avenues

This section proposes several directions to improve over current GC-IMGEP approaches towards solving the intrinsically motivated skills acquisition problem.

7.1 Towards a Greater Diversity of Goal Representations

This section proposes new types of goal representation that RL agents could leverage to tackle the intrinsically motivated skills acquisition problem.

Time-extended goals. All RL approaches reviewed in this paper consider *time-specific* goals, that is, goals whose completion can be assessed from any state s. This is due to the Markov property requirement, where the next state and reward need to be a function of the previous state only. *Time-extended* goals – i.e. goals whose completion can be judged by observing a sequence of states (e.g. *jump twice*) – can however be considered by adding time-extended features to the state (e.g. the difference between the current state and the initial state Colas et al. [2020c]). To avoid such *ad-hoc* state representations, one could imagine using reward function architectures that incorporate forms of memory such as Recurrent Neural Network (RNN) architectures [Elman, 1993] or Transformers [Vaswani et al., 2017]. Although recurrent policies are often used in the literature [Chevalier-Boisvert et al., 2019, Hill et al., 2019, Loynd et al., 2019, Goyal et al., 2019], recurrent reward functions have not been investigated. Time-extended goals include goals expressed as repetitions of a given interaction (e.g. *knock three times*). Such goals call for novel inductive biases in reward function architectures (e.g. memory, counting ability etc.). Note that POP-IMGEP approaches are not limited by the Markov property and have used time-extended goals (e.g. Forestier and Oudeyer [2016]).

Learning goals. *Goal-driven learning* is the idea that humans use *learning goals*, goals about their own learning abilities as a way to simplify the realization of *task goals* [Ram et al., 1995]. Here, we refer to *task goals* as goals that express constraints on the physical state of the agent and/or environment. On the other hand, *learning goals* refer to goals that express constraints on the knowledge of the agent. Although most RL approaches target task goals, one could envision the use of *learning goals* for RL agents. In a way, learning-progress-based learning is a form of learning goal: as the agent favors regions of the goal space to sample its task goals, it formulates the goal of learning about this specific goal region [Baranes and Oudeyer, 2013b, Fournier et al., 2018, 2019, Colas et al., 2019, Blaes et al., 2019, Akakzia et al., 2020]. Embodied Question Answering problems can also be seen as using learning goals. The agent is asked a question (i.e. a learning goal) and needs to explore the environment to answer it (acquire new knowledge) [Das et al., 2018, Yuan et al., 2019].

Goals as optimization under selected constraints. We discussed the representations of goals as a balance between multiple objectives. An extension of this idea is to integrate the selection of constraints on states or trajectories. One might want to maximize a given metric (e.g. walking speed), while setting various constraints (e.g. maintaining the power consumption below a given threshold or controlling only half of the motors). The agent could explore in the space of constraints, setting constraints to itself, building a curriculum on these, etc. This is partially investigated in Colas et al. [2020b], where the agent samples constraint-based goals in the optimization of control strategies to mitigate the economic and health costs in simulated epidemics. This approach however, only considers constraints on minimal values for the objectives and requires the training of an additional Q-function per constraint.

7.2 Towards Goal Composition and Imagination

It is commonly admitted that children play supports learning [Piaget, 1955, Montessori, 2013, Sutton-Smith, 2009]. Recently, Chu and Schulz [2020] proposed to see play as an efficient behavior towards imagined goals. In a similar way, imagining creative out-of-distribution goals might help agents to explore more efficiently their environment. Humans are indeed very good at imagining and understanding out-of-distribution goals using compositional generalization (e.g. *put a hat on a goat*). If an agent knows how to perform atomic goals (e.g. *build a tower with the blue blocks* and *build a pyramid with the red blocks*), we would like to automatically generalize to composition of these goals, as well as their logical combinations (e.g. build a tower of red blocks, but do not build a pyramid of blue blocks). The logical combinations of atomic goals was investigated in Tasse et al. [2020], Chitnis et al. [2021], and Colas et al. [2020a], Akakzia et al. [2020]. The first approach represents the space of goals as a Boolean algebra, which allows immediate generalization to compositions of goals (AND, OR, NOT). The second approach considers using general symbolic and logic languages to express goals, but uses symbolic planning techniques that are not yet fully integrated in the goal-conditioned deep RL framework. The third and fourth train a generative model of goals conditioned on language inputs. Because it generates discrete goals, it can compose language instructions by composing the finite sets of discrete goals associated to each instruction (AND is the intersection, OR the union etc). However, these

works fall short of exploring the richness of goal compositionality and its various potential forms. Tasse et al. [2020] seem to be limited to specific goals as target features, while Akakzia et al. [2020] requires discrete goals.

7.3 Language as a Tool for Creative Goal Generation

Whether they are specific or abstract, time-specific or time-extended, whether they represent mixture of objectives, constraints, or logical combinations, all goals can be expressed easily by humans through language. Language, thus, seems like the ideal candidate to express goals in RL agents. So far, language was only used to formulate a few forms of goals (see Section 4). In the future, it might be used to express any type of goals. Recurrent Neural Networks (RNN) [Elman, 1993], Deep Sets [Zaheer et al., 2017], Graph Neural Networks (GNN) or Transformers are all architectures that benefits from inductive biases and could be leveraged to facilitate new forms of goal representations (time-extended, set-based, relational etc.). As a learning signal, Colas et al. [2020c] propose to leverage descriptive sentences from social partners as a way to train goal representations. Mixing self-supervised learning and language inputs from humans as in Lynch and Sermanet [2020] might be the way forward.

8. Discussion & Conclusion

This paper defined the intrinsically motivated skills acquisition problem and proposed to view intrinsically motivated goal-conditioned RL algorithms or GC-IMGEP as computational tools to tackle it. These methods belong to the new field of *developmental machine learning*, the intersection of the developmental robotics and RL fields. We reviewed current goal-conditioned RL approaches under the lens of intrinsically motivated agents that learn to represent and generate their own goals in addition of learning to achieve them.

We propose a new general definition of the *goal* construct: a pair of compact goal representation and an associated goal-achievement function. Interestingly, this viewpoint allowed us to categorize some RL approaches as goal-conditioned, even though the original papers did not explicitly acknowledge it. For instance, we view the Never Give Up [Badia et al., 2020b] and Agent 57 [Badia et al., 2020a] architectures as goal-conditioned, because agents actively select parameters affecting the task at hand (parameter mixing extrinsic and intrinsic objectives, discount factor) and see their behavior affected by this choice (goal-conditioned policies).

This point of view also offers a direction for future research. Intrinsically motivated agents need to learn to represent goals and to measure goal achievement. Future research could extend the diversity of considered goal representations, investigate novel reward function architectures and inductive biases to allow time-extended goals, goal composition and to improve generalization.

The general vision we convey in this paper is the vision of the learning agent as a curious scientist. A scientist that would formulate hypotheses about the world and explore it to find out whether they are true. A scientist that would ask questions, and setup intermediate goals to explore the world and find answers. A scientist that would set challenges to itself

to learn about the world, to discover new ways to interact with it and to grow its collection of skills and knowledge. Such a scientist could decide of its own agenda. It would not need to be instructed and could be guided only by its curiosity, by its desire to discover new information and to master new skills.

Approach	Goal Type	Goal Rep.	Reward Function	Goal sampling strategy
GC-IMGEPs that assume goal embeddings and reward functions				
Fournier et al. [2018]	Target features (+tolerance)	Pre-def	Pre-def	LP-Based
HAC Levy et al. [2018]	Target features	Pre-def	Pre-def	HRL
HIRO Nachum et al. [2018]	Target features	Pre-def	Pre-def	HRL
CURIOUS Colas et al. [2019]	Target features	Pre-def	Pre-def	LP-based
CLIC Fournier et al. [2019]	Target features	Pre-def	Pre-def	LP-based
CWYC Blaes et al. [2019]	Target features	Pre-def	Pre-def	LP-based
GO-EXPLORE Ecoffet et al. [2020]	Target features	Pre-def	Pre-def	Novelty
NGU Badia et al. [2020b]	Objective Balance	Pre-def	Pre-def	Uniform
AGENT 57 Badia et al. [2020a]	Objective Balance	Pre-def	Pre-def	Meta-learned
DECSTR Akakzia et al. [2020]	Binary problem	Pre-def	Pre-def	LP-based
GC-IMGEPs that learn their goal embedding and assume reward functions				
RIG Nair et al. [2018]	Target features (images)	Learned (VAE)	Pre-def	From VAE prior
GOALGAN Florensa et al. [2018]	Target features	Pre-def + GAN	Pre-def	Intermediate difficulty
Florensa et al. [2019]	Target features (images)	Learned (VAE)	Pre-def	From VAE prior
SKEW-FIT Pong et al. [2019]	Target features (images)	Learned (VAE)	Pre-def	Diversity
SETTER-SOLVER Racanière et al. [2019]	Target features (images)	Learned (Gen. model)	Pre-def	Uniform difficulty
MEGA Pitis et al. [2020]	Target features (images)	Learned (VAE)	Pre-def	Novelty
CC-RIG Nair et al. [2020]	Target features (images)	Learned (VAE)	Pre-def	From VAE prior
AMIGO Campero et al. [2020]	Target features (images)	Learned (with policy)	Pre-def	Adversarial
GRIMGEP Kovač et al. [2020]	Target features (images)	Learned (with policy)	Pre-def	Diversity and ALP
Full GC-IMGEPs				
DISCERN Warde-Farley et al. [2018]	Target features (images)	Learned (with policy)	Learned (similarity)	Diversity
DIAYN Eysenbach et al. [2018]	Discrete skills	Learned (with policy)	Learned (discriminability)	Uniform
Hartikainen et al. [2019]	Target features (images)	Learned (with policy)	Learned (distance)	Intermediate difficulty
Venkattaramanujam et al. [2019]	Target features (images)	Learned (with policy)	Learned (distance)	Intermediate difficulty
IMAGINE Colas et al. [2020c]	Binary problem (language)	Learned (with reward)	Learned	Uniform + Diversity

Table 1: **A classification of GC-IMGEP approaches.** The classification groups algorithms depending on their degree of autonomy: 1) GC-IMGEPs that rely on pre-defined goal representations (embeddings and reward functions); 2) GC-IMGEPs that rely on pre-defined reward functions but learn goal embeddings and 3) GC-IMGEPs that learn complete goal representations (embeddings and reward functions). For each algorithm, we report the type of goals being pursued (see Section 4), whether goal embeddings are learned (Section 5), whether reward functions are learned (Section 5.3) and how goals are sampled (Section 6). We mark in bold algorithms that use a developmental approaches and explicitly pursue the intrinsically motivated skills acquisition problem.

References

J. Achiam and S. Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

A. Akakzia, C. Colas, P.-Y. Oudeyer, M. Chetouani, and O. Sigaud. DECSTR: Learning goal-directed abstract behaviors using pre-verbal spatial predicates in intrinsically motivated agents. *arXiv preprint arXiv:2006.07185*, 2020.

J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.12. URL `http://dx.doi.org/10.1109/CVPR.2016.12`.

M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Advances in neural information processing systems*, pages 5048–5058, 2017.

M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *IEEE transactions on autonomous mental development*, 1(1):12–34, 2009.

A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, D. Guo, and C. Blundell. Agent57: Outperforming the atari human benchmark. *arXiv preprint arXiv:2003.13350*, 2020a.

A. P. Badia, P. Sprechmann, A. Vitvitskyi, D. Guo, B. Piot, S. Kapturowski, O. Tieleman, M. Arjovsky, A. Pritzel, A. Bolt, et al. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020b.

D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville. Systematic generalization: What is required and can it be learned?, 2018.

D. Bahdanau, F. Hill, J. Leike, E. Hughes, P. Kohli, and E. Grefenstette. Learning to Understand Goal Specifications by Modelling Reward. In *International Conference on Learning Representations*, jun 2019.

A. Baranes and P. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *CoRR*, abs/1301.4862, 2013a. URL `http://arxiv.org/abs/1301.4862`.

A. Baranes and P.-Y. Oudeyer. R-iac: Robust intrinsically motivated exploration and active learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169, 2009a.

A. Baranes and P.-Y. Oudeyer. Proximo-distal competence based curiosity-driven exploration. In *Learning, in" International Conference on Epigenetic Robotics, Italie*. Citeseer, 2009b.

A. Baranes and P.-Y. Oudeyer. Intrinsically motivated goal exploration for active motor learning in robots: A case study. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, 2010.

A. Baranes and P.-Y. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013b.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, pages 1471–1479, 2016.

M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

D. E. Berlyne. Curiosity and exploration. *Science*, 153(3731):25–33, 1966.

G. Berseth, D. Geng, C. Devin, C. Finn, D. Jayaraman, and S. Levine. Smirl: Surprise minimizing rl in dynamic environments. *arXiv preprint arXiv:1912.05510*, 2019.

S. Blaes, M. Vlastelica Pogančić, J. Zhu, and G. Martius. Control what you can: Intrinsically motivated task-planning agent. *Advances in Neural Information Processing Systems*, 32: 12541–12552, 2019.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

A. Campero, R. Raileanu, H. Küttler, J. B. Tenenbaum, T. Rocktäschel, and E. Grefenstette. Learning with amigo: Adversarially motivated intrinsic goals. *arXiv preprint arXiv:2006.12122*, 2020.

A. Cangelosi and M. Schlesinger. *Developmental robotics: From babies to robots*. MIT press, 2015.

R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

H. Chan, Y. Wu, J. Kiros, S. Fidler, and J. Ba. Actrce: Augmenting experience via teacher's advice for multi-goal reinforcement learning, 2019.

D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov. Gated-attention architectures for task-oriented language grounding, 2017.

H. Charlesworth and G. Montana. Plangan: Model-based planning with sparse rewards and multiple goals. *arXiv preprint arXiv:2006.00900*, 2020.

M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio. Baby{AI}: First Steps Towards Grounded Language Learning With a Human In the Loop. In *International Conference on Learning Representations*, 2019.

R. Chitnis, T. Silver, J. Tenenbaum, L. P. Kaelbling, and T. Lozano-Pérez. Glib: Efficient exploration for relational model-based reinforcement learning via goal-literal babbling. In *AAAI*, 2021.

J. Chu and L. Schulz. Exploratory play, rational action, and efficient search. *PsyArXiv*, 2020.

K. Chua, R. Calandra, R. McAllister, and S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4754–4765, 2018.

G. Cideron, M. Seurin, F. Strub, and O. Pietquin. Self-educated language agent with hindsight experience replay for instruction following. *arXiv preprint arXiv:1910.09451*, 2019.

F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

C. Colas, O. Sigaud, and P.-Y. Oudeyer. Gep-pg: Decoupling exploration and exploitation in deep reinforcement learning algorithms. *arXiv preprint arXiv:1802.05054*, 2018.

C. Colas, P. Oudeyer, O. Sigaud, P. Fournier, and M. Chetouani. CURIOUS: intrinsically motivated modular multi-goal reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 1331–1340, 2019.

C. Colas, A. Akakzia, P.-Y. Oudeyer, M. Chetouani, and O. Sigaud. Language-conditioned goal generation: a new approach to language grounding for rl. *arXiv preprint arXiv:2006.07043*, 2020a.

C. Colas, B. Hejblum, S. Rouillon, R. Thiébaut, P.-Y. Oudeyer, C. Moulin-Frier, and M. Prague. Epidemioptim: A toolbox for the optimization of control policies in epidemiological models. *arXiv preprint arXiv:2010.04452*, 2020b.

C. Colas, T. Karch, N. Lair, J.-M. Dussoux, C. Moulin-Frier, P. Ford Dominey, and P.-Y. Oudeyer. Language as a cognitive tool to imagine goals in curiosity-driven exploration. *arXiv*, 2020c.

C. Colas, V. Madhavan, J. Huizinga, and J. Clune. Scaling map-elites to deep neuroevolution. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 67–75, 2020d.

A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063, 2018.

P. Dayan and G. E. Hinton. Feudal reinforcement learning. In *Advances in neural information processing systems*, pages 271–278, 1993.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp. Goal-conditioned imitation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 15324–15335. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9667-goal-conditioned-imitation-learning.pdf`.

A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. First return then explore. *arXiv preprint arXiv:2004.12919*, 2020.

A. J. Elliot and J. W. Fryer. The goal construct in psychology. *Handbook of motivation science*, 18:235–250, 2008.

J. L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71 – 99, 1993. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(93)90058-4. URL `http://www.sciencedirect.com/science/article/pii/0010027793900584`.

B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv*, 2018.

B. Eysenbach, X. Geng, S. Levine, and R. Salakhutdinov. Rewriting history with inverse rl: Hindsight inference for policy improvement. *arXiv preprint arXiv:2002.11089*, 2020.

C. Florensa, D. Held, X. Geng, and P. Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528, 2018.

C. Florensa, J. Degrave, N. Heess, J. T. Springenberg, and M. Riedmiller. Self-supervised learning of image embedding for continuous control, 2019.

S. Forestier and P. Oudeyer. Modular active curiosity-driven discovery of tool use. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3965–3972, Oct 2016. doi: 10.1109/IROS.2016.7759584.

S. Forestier and P.-Y. Oudeyer. Modular active curiosity-driven discovery of tool use. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 3965–3972. IEEE, 2016.

S. Forestier, R. Portelas, Y. Mollard, and P.-Y. Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*, 2017.

P. Fournier, O. Sigaud, M. Chetouani, and P.-Y. Oudeyer. Accuracy-based curriculum learning in deep reinforcement learning. *arXiv preprint arXiv:1806.09614*, 2018.

P. Fournier, C. Colas, M. Chetouani, and O. Sigaud. Clic: Curriculum learning and imitation for object control in non-rewarding environments. *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

K. Frans, J. Ho, X. Chen, P. Abbeel, and J. Schulman. Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767*, 2017.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

A. Gopnik, A. N. Meltzoff, and P. K. Kuhl. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co, 1999.

J. Gottlieb and P.-Y. Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758–770, 2018.

A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf. Recurrent independent mechanisms, 2019.

K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control, 2016.

K. Hartikainen, X. Geng, T. Haarnoja, and S. Levine. Dynamical distance learning for semi-supervised and unsupervised skill discovery. In *International Conference on Learning Representations*, 2019.

K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, and P. Blunsom. Grounded Language Learning in a Simulated 3D World, jun 2017. URL `http://arxiv.org/abs/1706.06551`.

T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, et al. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro. Emergent systematic generalization in a situated agent, 2019.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Y. Jiang, S. Gu, K. Murphy, and C. Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. In *Workshop on "Structure & Priors in Reinforcement Learning"at ICLR 2019*, jun 2019. URL `http://arxiv.org/abs/1906.07343`.

L. P. Kaelbling. Learning to achieve goals. In *IJCAI*, pages 1094–1099. Citeseer, 1993.

F. Kaplan and P.-Y. Oudeyer. In search of the neural circuits of intrinsic motivation. *Frontiers in neuroscience*, 1:17, 2007.

C. Kidd and B. Y. Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3): 449–460, 2015.

K. Kim, M. Sano, J. De Freitas, N. Haber, and D. Yamins. Active world model learning with progress curiosity. In *International Conference on Machine Learning*, pages 5306–5315. PMLR, 2020.

G. Kovač, A. Laversanne-Finot, and P.-Y. Oudeyer. Grimgep: Learning progress for robust goal sampling in visual deep reinforcement learning, 2020.

T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683, 2016.

J. B. Lanier, S. McAleer, and P. Baldi. Curiosity-driven multi-criteria hindsight experience replay. *CoRR*, abs/1906.03710, 2019. URL `http://arxiv.org/abs/1906.03710`.

J. Lehman and K. O. Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 211–218, 2011.

A. Levy, R. Platt, and K. Saenko. Hierarchical reinforcement learning with hindsight. *arXiv preprint arXiv:1805.08180*, 2018.

R. Li, A. Jabri, T. Darrell, and P. Agrawal. Towards practical multi-object manipulation using relational reinforcement learning. *arXiv preprint arXiv:1912.11032*, 2019.

C. Linke, N. M. Ady, M. White, T. Degris, and A. White. Adapting behaviour via intrinsic reward: A survey and empirical study. *arXiv preprint arXiv:1906.07865*, 2019.

L. Lonini, S. Forestier, C. Teulière, Y. Zhao, B. E. Shi, and J. Triesch. Robust active binocular vision through intrinsically motivated learning. *Frontiers in neurorobotics*, 7: 20, 2013.

M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in neural information processing systems*, pages 206–214, 2012.

R. Loynd, R. Fernandez, A. Celikyilmaz, A. Swaminathan, and M. Hausknecht. Working memory graphs, 2019.

J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel. A Survey of Reinforcement Learning Informed by Natural Language. *IJCAI'19*, jun 2019. URL `http://arxiv.org/abs/1906.03926`.

C. Lynch and P. Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020.

C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. Learning latent plans from play. In L. P. Kaelbling, D. Kragic, and K. Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1113–1132. PMLR, 30 Oct–01 Nov 2020. URL `http://proceedings.mlr.press/v100/lynch20a.html`.

D. J. Mankowitz, A. Žídek, A. Barreto, D. Horgan, M. Hessel, J. Quan, J. Oh, H. van Hasselt, D. Silver, and T. Schaul. Unicorn: Continual learning with a universal, off-policy agent. *arXiv preprint arXiv:1802.08294*, 2018.

G. Martius, R. Der, and N. Ay. Information driven self-organization of complex robotic behaviors. *PloS one*, 8(5):e63400, 2013.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

M. Montessori. *The Montessori Method*. Transaction publishers, 2013.

J.-B. Mouret and J. Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

O. Nachum, S. S. Gu, H. Lee, and S. Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3303–3313, 2018.

A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel. Overcoming exploration in reinforcement learning with demonstrations. *arXiv preprint arXiv:1709.10089*, 2017.

A. Nair, S. Bahl, A. Khazatsky, V. Pong, G. Berseth, and S. Levine. Contextual imagined goals for self-supervised robotic learning. In *Conference on Robot Learning*, pages 530–539, 2020.

A. V. Nair, V. Pong, M. Dalal, S. Bahl, S. Lin, and S. Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pages 9191–9200, 2018.

S. Nasiriany, V. Pong, S. Lin, and S. Levine. Planning with goal-conditioned policies. In *Advances in Neural Information Processing Systems*, pages 14843–14854, 2019.

M. Nguyen and P.-Y. Oudeyer. Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots*, 36(3):273–294, 2014.

J. Oh, S. Singh, H. Lee, and P. Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. *arXiv preprint arXiv:1706.05064*, 2017.

P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2007.

P.-Y. Oudeyer and L. B. Smith. How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, 8(2):492–502, 2016.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Piaget. *The construction of reality in the child*, volume 82. Routledge, 1955.

S. Pitis, H. Chan, S. Zhao, B. Stadie, and J. Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. *arXiv preprint arXiv:2007.02832*, 2020.

M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

V. H. Pong, M. Dalal, S. Lin, A. Nair, S. Bahl, and S. Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

R. Portelas, C. Colas, L. Weng, K. Hofmann, and P.-Y. Oudeyer. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

D. Precup. *Temporal abstraction in reinforcement learning*. PhD thesis, The University of Massachussetts, 2000.

S. Racanière, A. Lampinen, A. Santoro, D. Reichert, V. Firoiu, and T. Lillicrap. Automated curricula through setter-solver interactions. *arXiv*, 2019.

A. Ram, D. B. Leake, and D. Leake. *Goal-driven learning*. MIT press, 1995.

M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Van de Wiele, V. Mnih, N. Heess, and J. T. Springenberg. Learning by playing-solving sparse reward tasks from scratch. *arXiv preprint arXiv:1802.10567*, 2018.

F. Röder, M. Eppe, P. D. Nguyen, and S. Wermter. Curious hierarchical actor-critic reinforcement learning. *arXiv preprint arXiv:2005.03420*, 2020.

M. Rolf and J. J. Steil. Efficient exploratory learning of inverse kinematics on a bionic elephant trunk. *IEEE transactions on neural networks and learning systems*, 25(6):1147–1160, 2013.

M. Rolf, J. J. Steil, and M. Gienger. Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, 2(3):216–229, 2010.

L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, and B. M. Lake. A benchmark for systematic generalization in grounded language understanding, 2020.

V. G. Santucci, G. Baldassarre, and M. Mirolli. Grail: a goal-discovering robotic architecture for intrinsically-motivated learning. *IEEE Transactions on Cognitive and Developmental Systems*, 8(3):214–231, 2016.

V. G. Santucci, P.-Y. Oudeyer, A. Barto, and G. Baldassarre. Intrinsically motivated open-ended learning in autonomous robots. *Frontiers in Neurorobotics*, 13:115, 2020.

T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320, 2015.

J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. *arXiv preprint arXiv:2005.05960*, 2020.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

S. Sukhbaatar, Z. Lin, I. Kostrikov, G. Synnaeve, A. Szlam, and R. Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

R. S. Sutton, D. Precup, and S. P. Singh. Intra-option learning about temporally abstract actions. In *ICML*, volume 98, pages 556–564, 1998.

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

R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768, 2011.

B. Sutton-Smith. *The ambiguity of play*. Harvard University Press, 2009.

G. N. Tasse, S. James, and B. Rosman. A boolean task algebra for reinforcement learning. *arXiv preprint arXiv:2001.01394*, 2020.

M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

V. Veeriah, J. Oh, and S. Singh. Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*, 2018.

S. Venkattaramanujam, E. Crawford, T. Doan, and D. Precup. Self-supervised learning of distance functions for goal-conditioned reinforcement learning. *arXiv preprint arXiv:1907.02998*, 2019.

A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1703.01161*, 2017.

D. Warde-Farley, T. Van de Wiele, T. Kulkarni, C. Ionescu, S. Hansen, and V. Mnih. Unsupervised control through non-parametric discriminative rewards. *arXiv preprint arXiv:1811.11359*, 2018.

Y. Wu, G. Tucker, and O. Nachum. The laplacian in rl: Learning representations with efficient approximations, 2018.

X. Yuan, M.-A. Côté, J. Fu, Z. Lin, C. Pal, Y. Bengio, and A. Trischler. Interactive language learning by question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/d19-1280. URL `http://dx.doi.org/10.18653/v1/D19-1280`.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

Y. Zhang, P. Abbeel, and L. Pinto. Automatic curriculum learning through value disagreement. *arXiv preprint arXiv:2006.09641*, 2020.

Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.