Goal-Conditioned Reinforcement Learning, Multitask learning and Autotelic agents

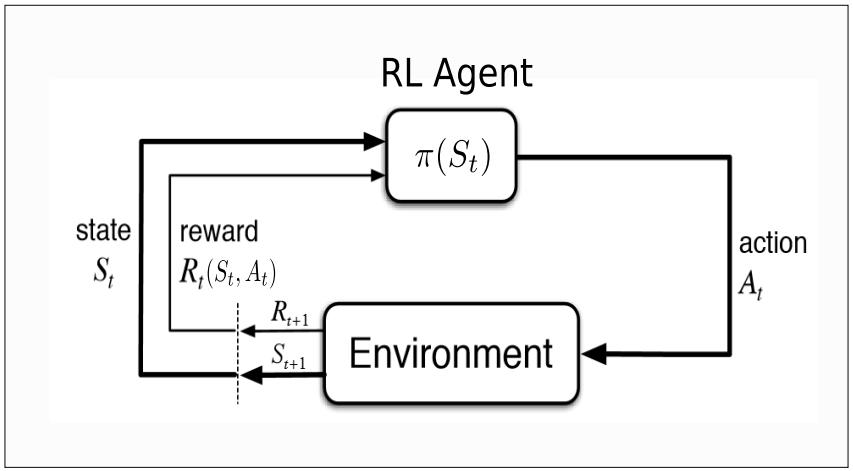
Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Outline

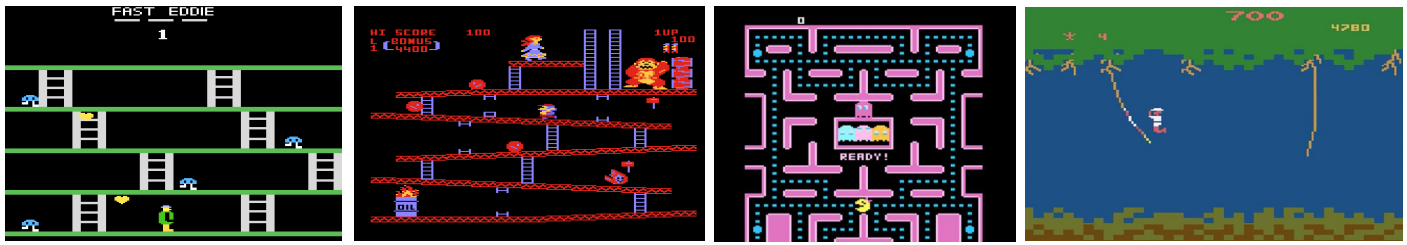
- Motivation
- General framework
- Basic concepts: desired and achieved goals, transfer
- Three core components:
  - Hindsight Experience Replay
  - Curriculum learning
  - Goal representation learning
- Several perspectives:
  - The hard exploration, single goal perspective
  - The extrinsic multigoal perspective
  - The continual learning / open-ended learning perspectives
  - The hierarchical RL perspective
  - The unsupervised RL perspective
- Role of the components in the perspectives
- Some key papers

Motivation I: standard RL is not rich enough

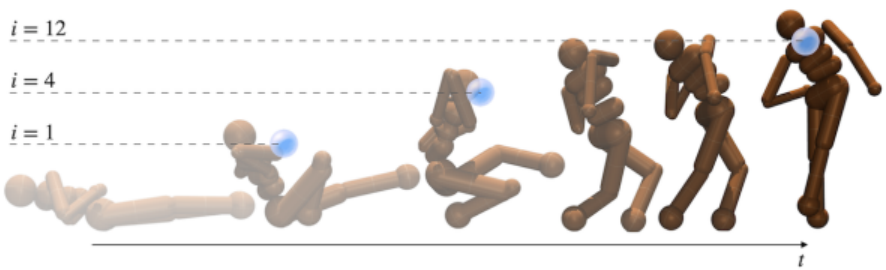


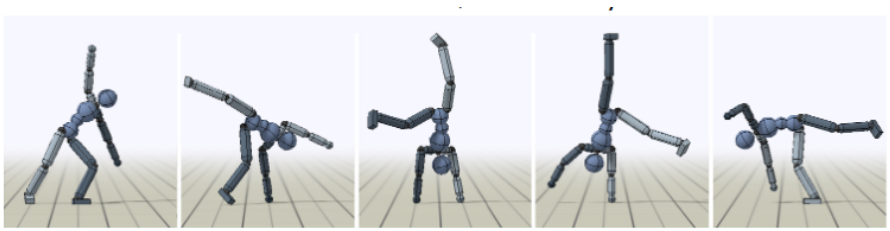
- The standard RL framework addresses a single task which is only specified through a reward function
- Not rich enough to account for many learning phenomena when we face multiple tasks/goals: transfer learning, curriculum learning, etc.
- And most importantly for the intuition that we have explicit goals
- Goal-conditioned RL (GCRL) is a framework to account for this richer context.

Motivation II: having a single controller, transfer learning



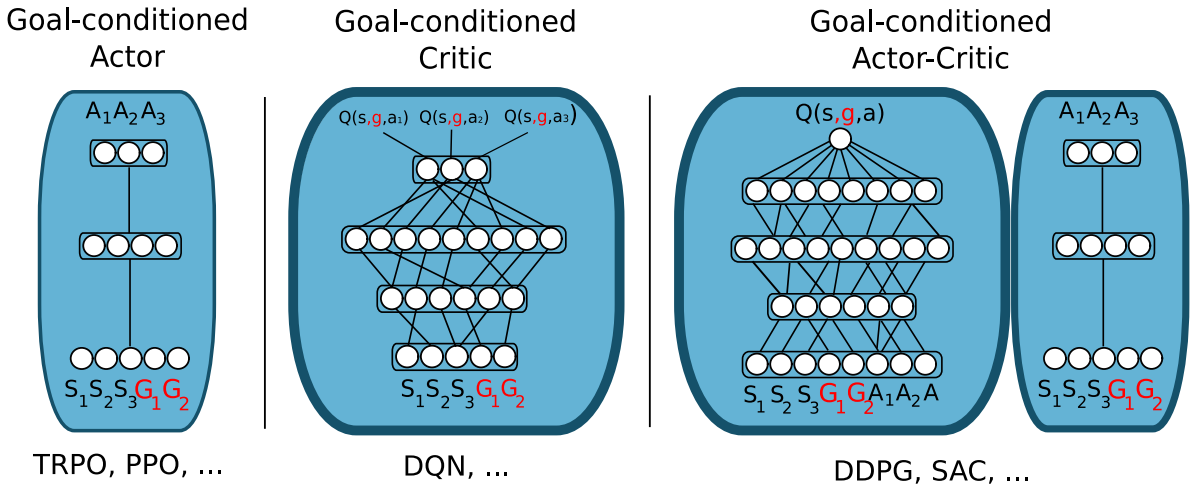
- In ATARI, DQN learns a different policy for each game
- The policies are learned independently
- Though there might be some common structure in different games





- For a humanoid, we would like having a single controller for all movements

GCRL



- Precursor: the Kaelbling paper (tabular RL)
- Universal Value Function Approximators (anterior to DQN)
- Learned with standard Q-LEARNING or ACTOR-CRITIC schemes
- Main advantage: generalization over the state space AND the goal space

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

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

Goal achievement

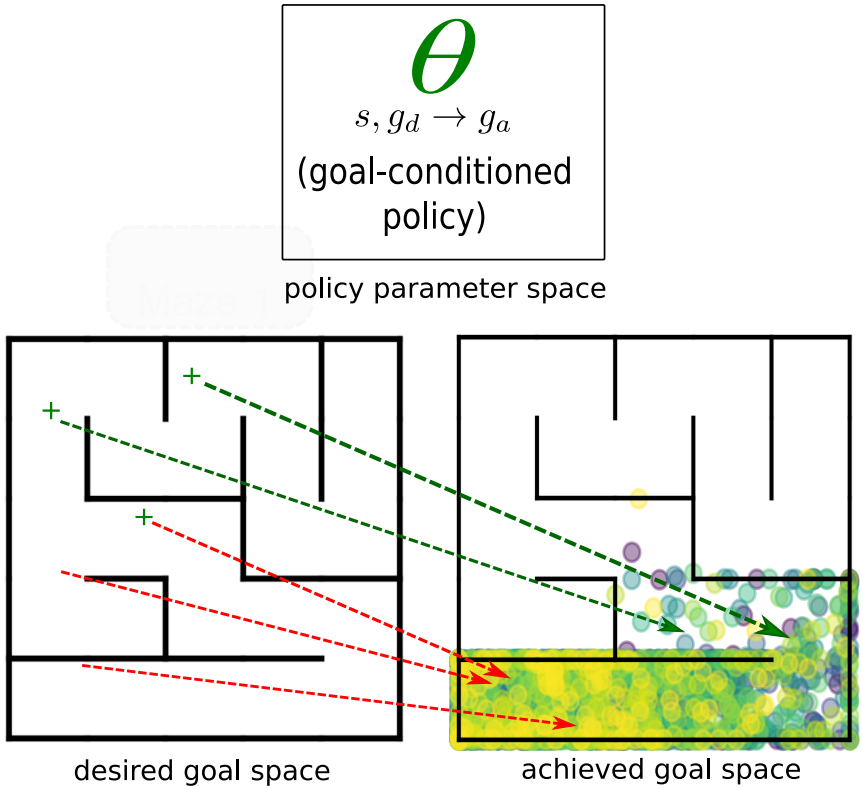
Generalized definition of the goal construct for RL:

- ***Goal***: a $g = (z_g, R_g)$ pair where z_g is a compact *goal parameterization* or *goal embedding* and R_g is a *goal-achievement* function.
- **Goal-achievement function:** $R_g(\cdot) = R_{\mathcal{G}}(\cdot \mid z_g)$ where $R_{\mathcal{G}}$ is a goal-conditioned reward function.

- A goal is a point in a goal space (embedding), or a member of a discrete goal set
- Goal achievement function:
  - Given a trajectory, tells which goal has been achieved (can be none).
  - Assumes $achieved = f(trajectory, goal)$, constraint satisfaction on the trajectory
  - Useful for HER
- Goal-conditioned reward function: given a desired goal
  - Dense reward functions: decreasing function of the distance between the state and the desired goal (assumes projecting in the same space)
  - Sparse reward functions: 1 if the goal is achieved, 0 otherwise (or 0/-1 to favor exploration)
  - Sparse reward more used in autotelic agents

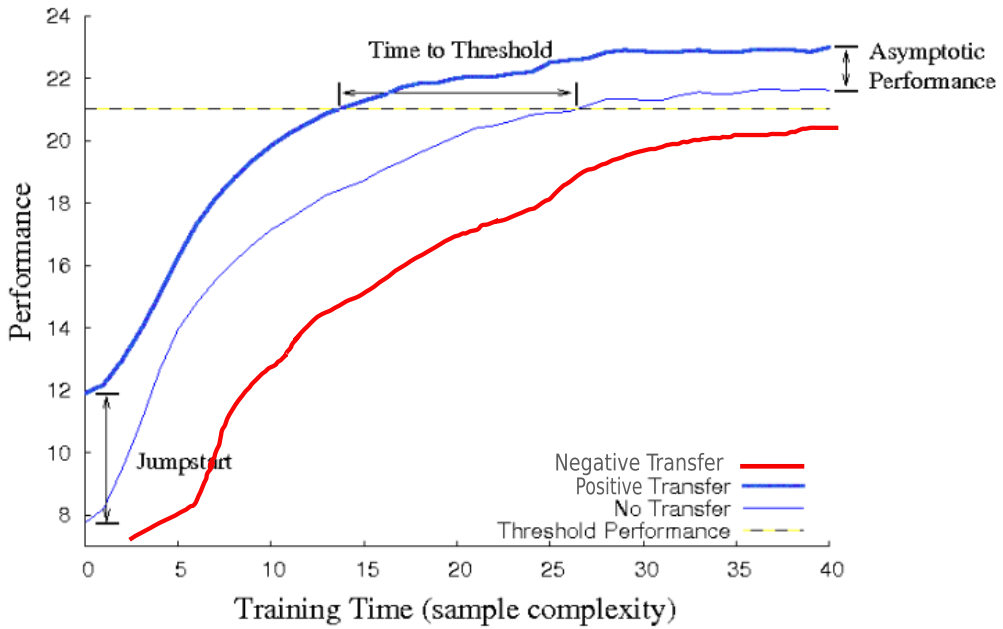
Colas, C., Karch, T., Sigaud, O., and Oudeyer, P.-Y. (2022) Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199

Desired goal and achieved goal



- The desired goal is the goal we input to the policy
- The achieved goal is the goal given by the goal achievement function from the obtained trajectory
- In general, they are not equal
- One perspective on GCRL is to try to get them equal

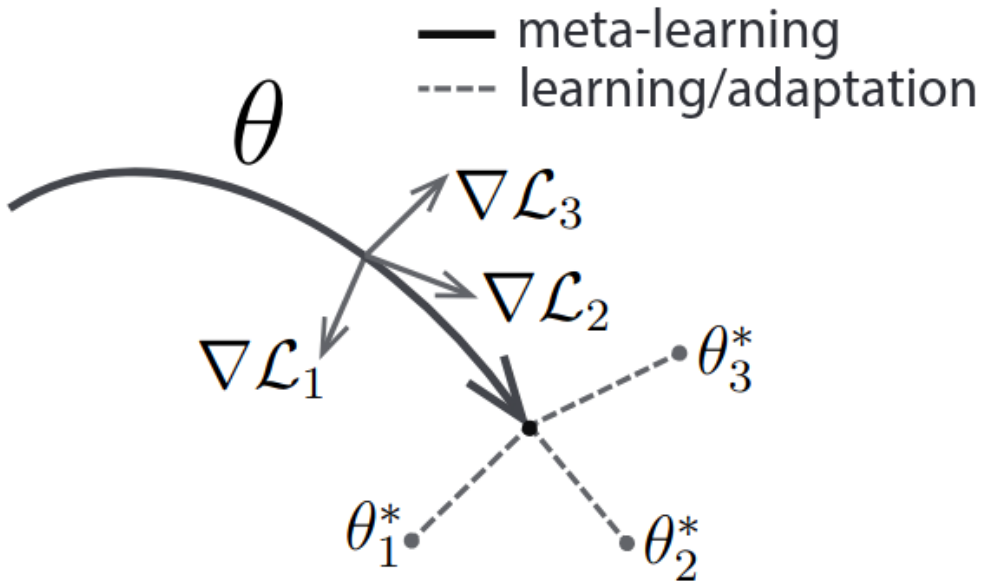
Transfer learning



- Need to be off-policy for efficient transfer

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

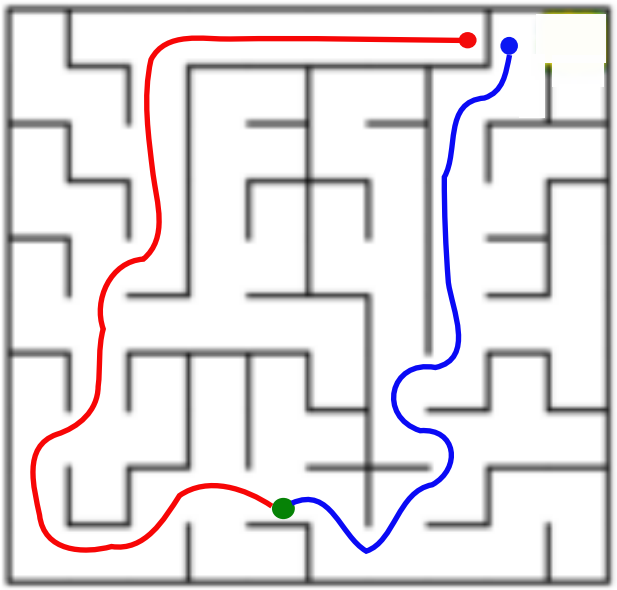
Model-Agnostic Meta-Learning (MAML)



- Trying to optimize transfer
- Gradient to find a policy that adapts to several tasks in few gradient steps (few shot)
- Gradient about a gradient
- Many follow-up works (REPTILE, ...): hot topic in 2019

Finn, C., Abbeel, P., and Levine, S. (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR

Intuition about generalization



- To generalize well, we need local continuity in the input-output space
- A small change in goal (or in starting state) should result in a small change of trajectory
- Not true of the above example
- Goal representation learning in a latent space might be useful

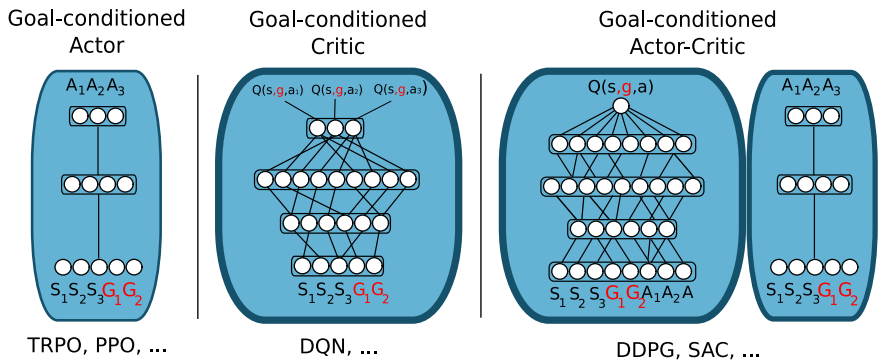
Learning from failures

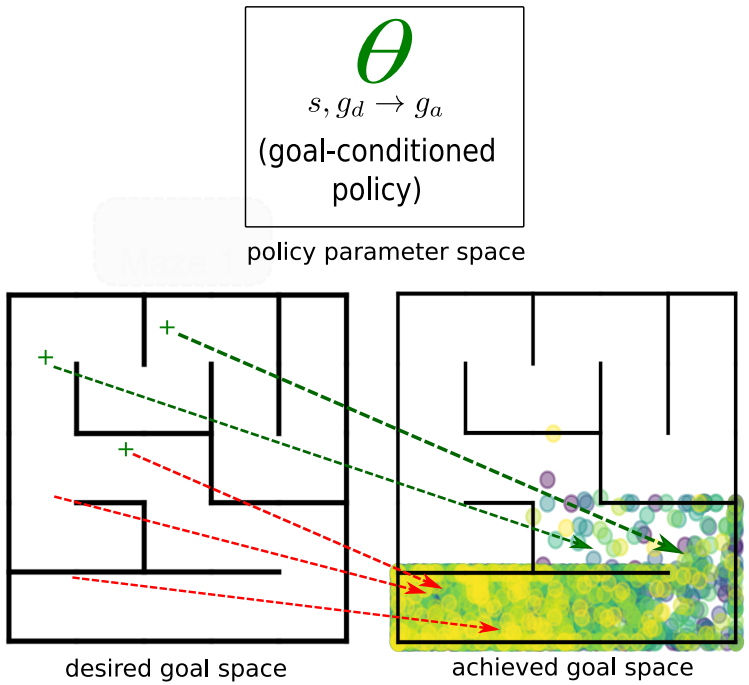
- Without finding reward, an RL agent learns nothing

- Consider a learning agent whose goal is to reach a particular outcome
- In the beginning, this agent may often fail
- The failed experiment produced another outcome than the expected one
- But this can be turned into useful knowledge
- This is the essence of Hindsight Experience Replay (HER)

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. (2017) Hindsight Experience Replay. *arXiv preprint arXiv:1707.01495*

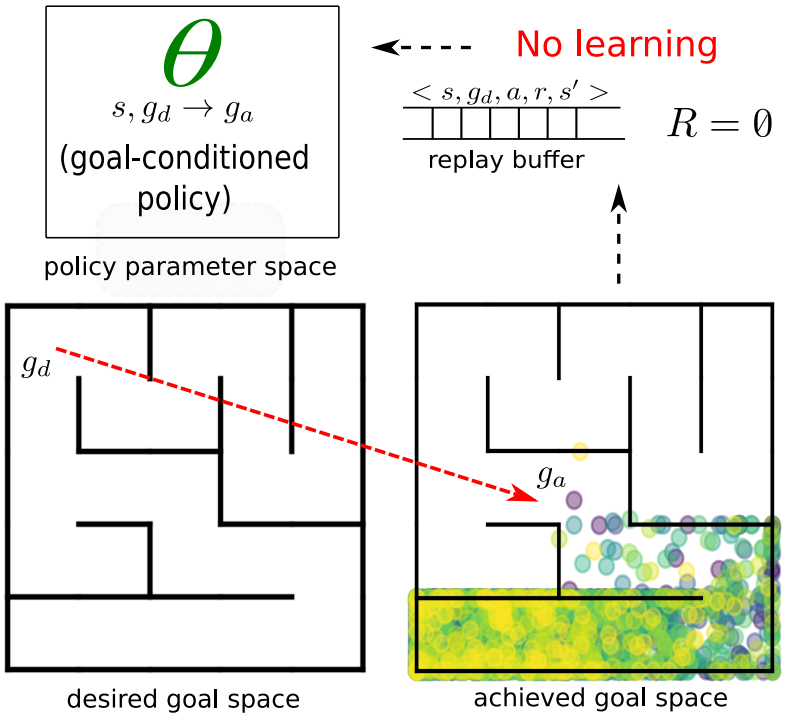
Four components





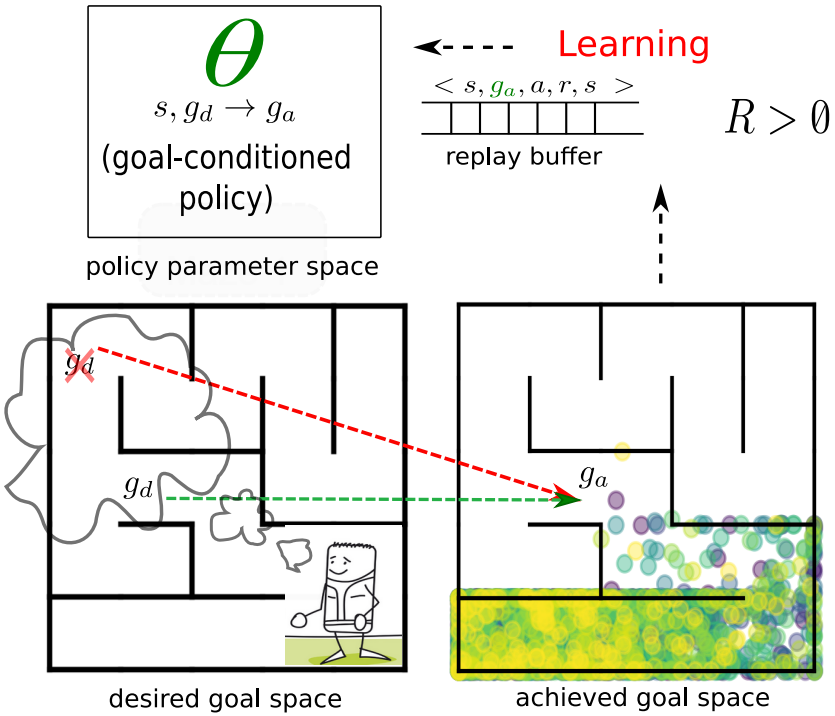
1. Goal conditioned policies
2. Mapping from policy parameter space and desired goal space to achieved goal space
3. Any (off-policy) RL algorithm (DQN, DDPG, TD3, PPO, SAC, ...) with a replay buffer
4. A special replay buffer with a goal relabelling mechanism

Goal relabelling mechanism (1)



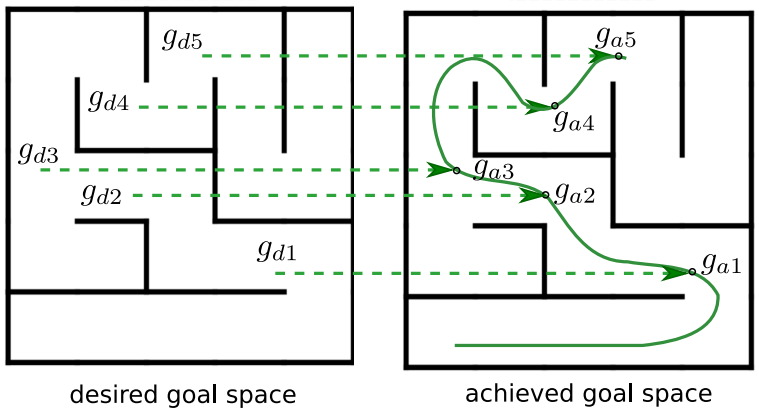
- The agent targets a desired goal g_d
- The policy π_θ produces an achieved goal g_a
- The trajectory is stored (it may produce no reward)

Goal relabelling mechanism (2)



- The agent pretends it was targetting g_a
- HER relabels the stored trajectory with g_a instead of g_d
- This propagates value in the (state, action) space through generalization
- And the agent competence increases over unseen goals
- Note: the policy achieved g_a when conditioned on g_d, it may not achieve g_a when conditioned on g_a...

When the goal is a state



- If goal space = state space, HER may set as goal any state along the trajectory
- Trade-off between replaying more and trying more new actions (risk of over-fitting to replay)
- Variants of HER: CHER (Curriculum + HER), DHER (dynamic goals), MCHER (multi-criteria)...

Fang, M., Zhou, C., Shi, B., Gong, B., Xu, J., and Zhang, T. (2018) DHER: Hindsight experience replay for dynamic goals. In *International Conference on Learning Representations*

HER vs Curriculum learning

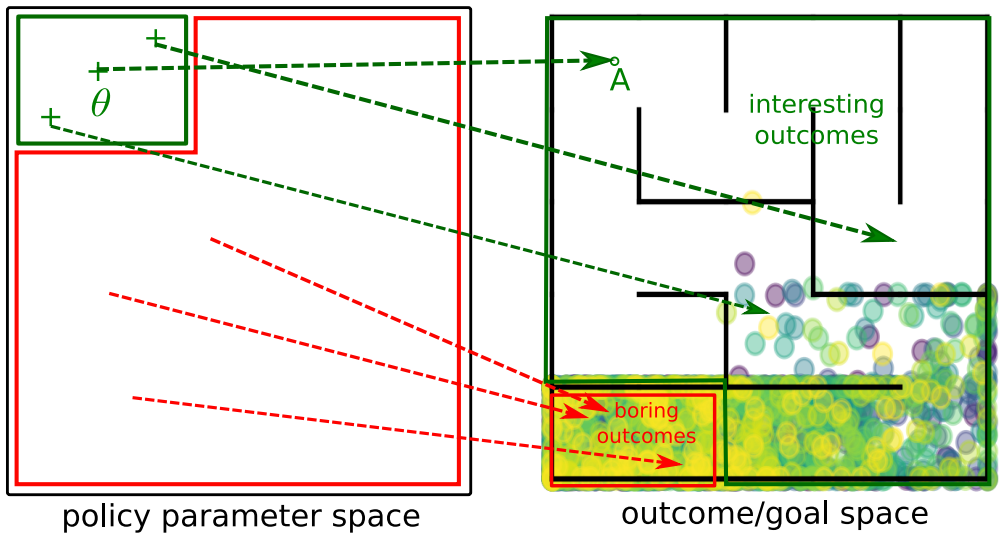
- HER acts by relabelling achieved goals as fake desired goals
- Does not produce new trajectories by selecting new desired goals
- Curriculum acts by selecting desired goals
- Both can be combined

Fang, M., Zhou, T., Du, Y., Han, L., and Zhang, Z. (2019) Curriculum-guided hindsight experience replay. *Advances in neural information processing systems*, 32

Desired Goals and achieved goals

- General curriculum question: how to set the desired goals to learn better?
- Question 1: coverage, exploration: how to set the desired goals to reach more achieved goals?
- Question 2: performance: how to better reach the already achieved goals?
- Questions 1 and 2 are overlapping: most algorithms contribute to both, but with a different perspective

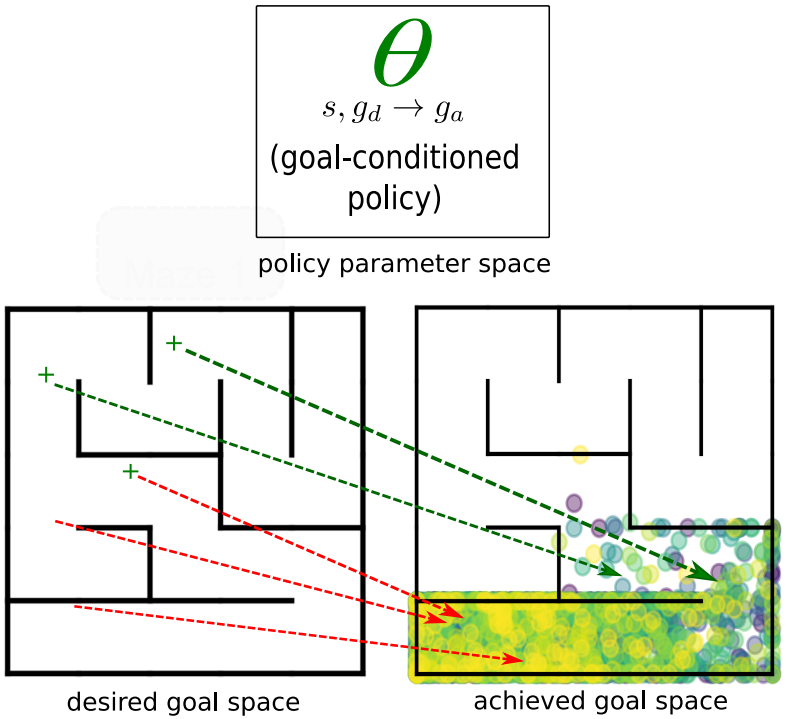
Lesson from Goal Exploration Processes



- Very often, few parameter vectors map to interesting achieved goals
- The GEP algorithm favors sampling these interesting achieved goals
  - Sample a random desired goal
  - Find the nearest achieved goal A' and select the corresponding θ
  - Perturb θ into θ' and get a new achieved goal A'
- Results in sampling "at the border" of currently achieved goals

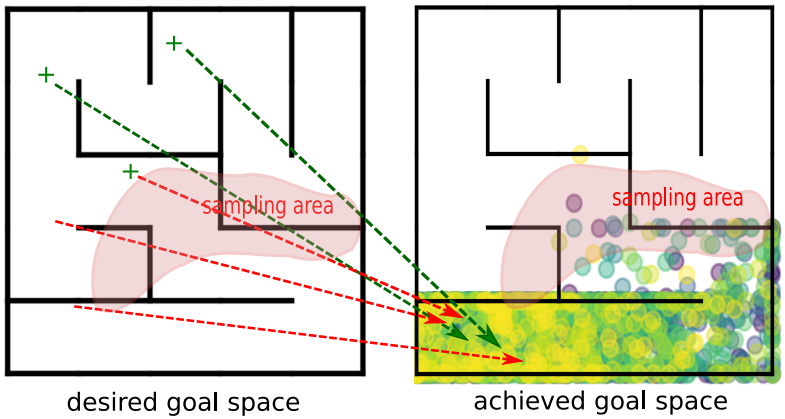
Chenu, A., Perrin-Gilbert, N., Doncieux, S., & Sigaud, O. (2021) Selection-expansion: A unifying framework for motion-planning and diversity search algorithms. *arXiv preprint arXiv:2104.04768*

Transposition to GCRL



- We consider the desired goal $\rightarrow$ achieved goal mapping
- The mapping depends on the GCP
- The perfect mapping would be the identity
- RL and HER contribute to improving the mapping (particularly HER)

GCRL dynamics

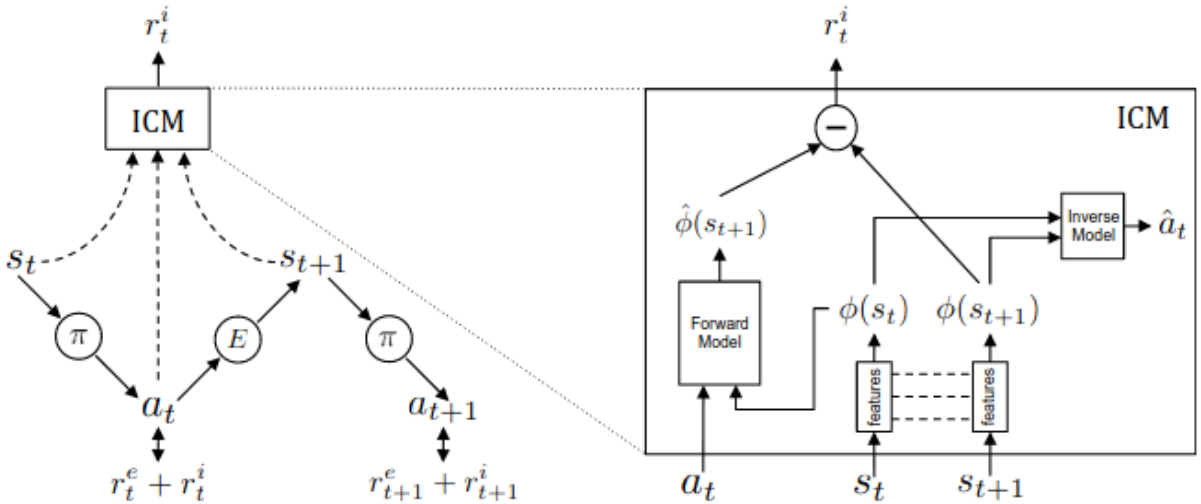


- In many problems, an initial policy will have a very 'injective' mapping
- GEP equivalent: sample a random desired goal, find the closest achieved goal, and take a close desired goal
- Similar to sampling at the border
- It would be better to sample in the useful corridor
- Many algorithms do so

Pitis, S., Chan, H., Zhao, S., Stadie, B., and Ba, J. (2020) Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*, pages 7750–7761. PMLR

Castanet, N., Lamprier, S., and Sigaud, O. (2022) Stein variational goal generation for reinforcement learning in hard exploration problems. *arXiv preprint arXiv:2206.06719*

Curriculum based on surprise/novelty



- Intrinsic motivation: reward states for which the forward model predicts poorly
- Target goals corresponding to rewarded states
- Results in visiting poorly visited states
- White noise problem: the agent may get stuck on what it cannot predict

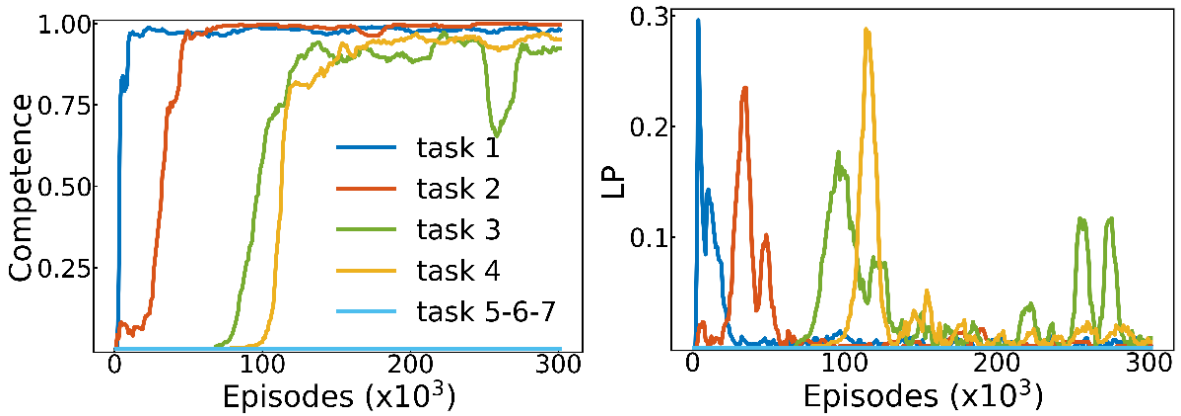
Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017) Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*

Curriculum based on learning progress: approach

- Too easy: no progress, should be avoided
- Too hard: no progress, should be avoided
- Try bandits where learning progress is faster
- Exploration may result in adding new bandits

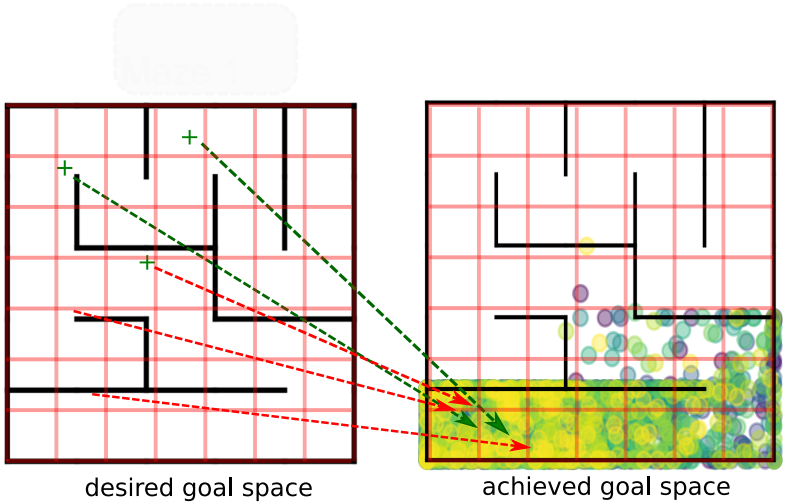
Lopes, M. and Oudeyer, P.-Y. (2012) The strategic student approach for life-long exploration and learning. In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–8. IEEE

Curriculum based on learning progress: dynamics



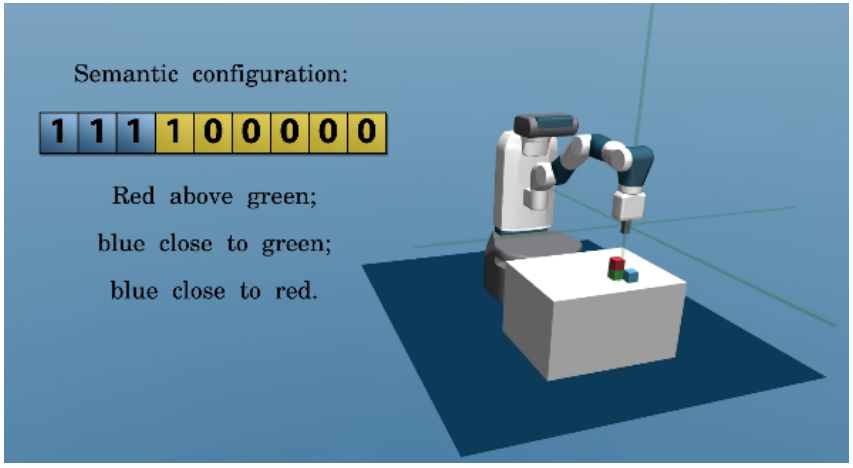
- Competence raises in order of growing difficulty
- LP generates more training in order of growing difficulty
- Catastrophic forgetting generates new training

LP applied to desired/achieved goals



- Cluster goals into regions (cells)
- Compute progress in terms of (some) distance between desired and achieved goals?
- Sample cells with highest progress
- These will probably be the border of the most visited regions
- Discovery of new goals is a side effect of exploration

Limits of the goal as state view



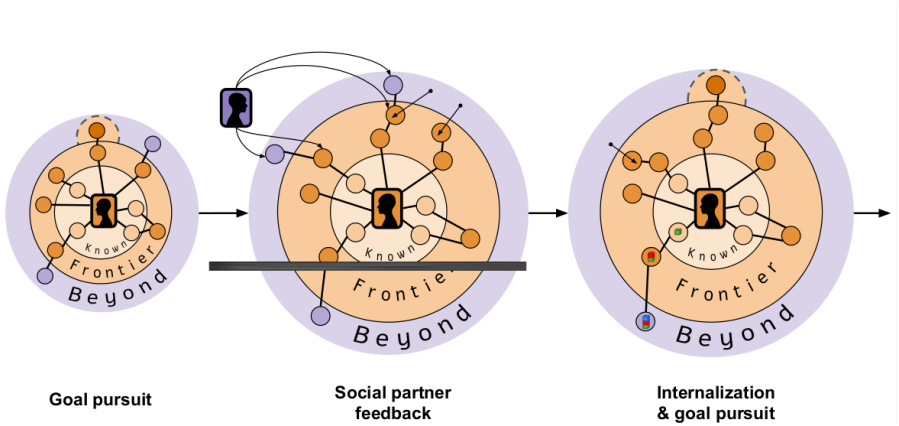
- The goal space can differ form the state space
- Various goal spaces: language, latent...
- Many goals in the goal space may not be reachable
- Just sample from already discovered goals
- Some achievable goals may never be discovered

Goal representation learning

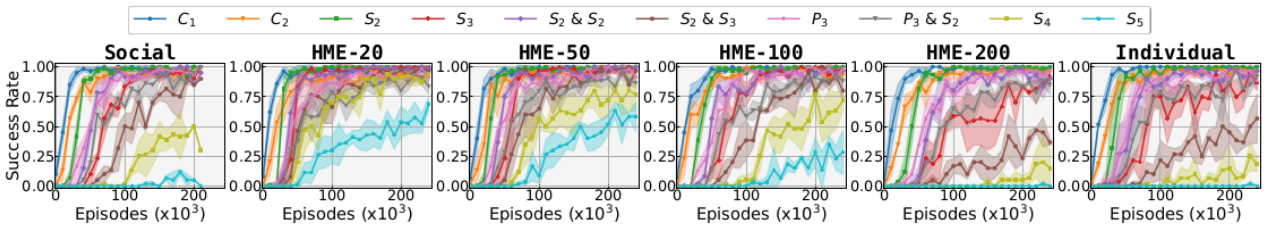
- In most GCRL problems, the goal space is the state space, or is given
- We would like the agent to discover it
- A lot based on VAEs or β-VAEs
- AURORA, TAXONS, UGL, MUGL
- Alternative: contrastive learning
- Goal generator versus goal space
- Unsupervised pre-training to learn a representation of the state space
- Specificity of Open-ended RL: learn a state and action space from lower level sensor/actuators

Doncieux, S., Filliat, D., Díaz-Rodríguez, N., Hospedales, T., Duro, R., Coninx, A., Roijers, D. M., Girard, B., Perrin, N., & Sigaud, O. (2018) Open-ended learning: a conceptual framework based on representational redescription. *Frontiers in Robotics and AI*, 12

Help Me Explore



- Key property: the tutor has a model of the learner's knowledge
- It proposes Frontier + Beyond goals (HME)
- The learner internalizes tutor's goals, it can train on them and on its own goals



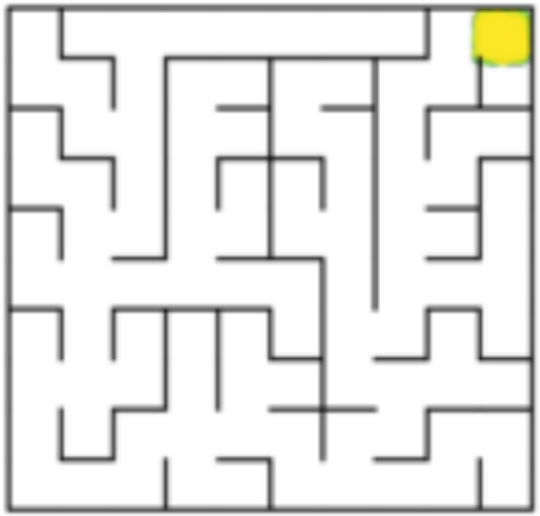
- Guided play is more efficient than learning on its own and full guidance

Akakzia, A., Serris, O., Sigaud, O., and Colas, C. (2022) Help me explore: Minimal social interventions for graph-based autotelic agents. *arXiv preprint arXiv:2202.05129* (submitted to ICLR 2023)

Synthesis

- HER accelerates training in the sparse reward context by replaying achieved trajectories
- Curriculum is about the choice of desired goals
- Desired/achieved goals coverage based curriculum: more for exploration
- LP-based curriculum: more for consolidation
- Goal discovery as a central issue
- A social partner may help goal discovery

Hard exploration, single goal perspective: Overview



- The agent targets a difficult goal, i.e. a sparse reward RL problem
- Without a reward signal, model-free RL produces inefficient random search
- HER provides additional reward signals
- No Curriculum learning is used: the desired goal stays fixed
- No goal representation learning: in general, the goal representation is given

Extrinsic multigoal perspective: Overview

- The agent targets many goals
- Learning to achieve each goal in isolation is sample inefficient
- The HER agent learns unexpected goals through its failures
- All goals are available at all times
- Curriculum learning helps sequencing goal achievement

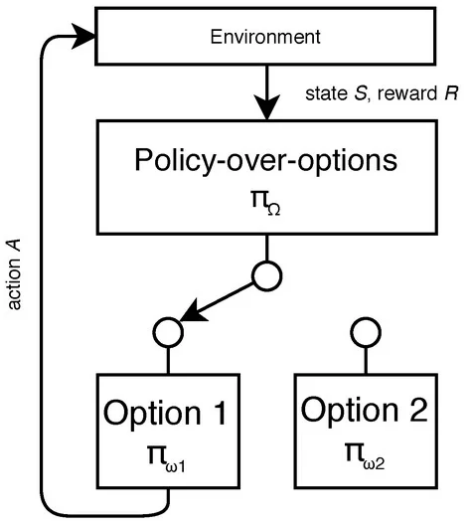
Continual / Open-ended learning perspective

- Continual learning: same as multigoal, but no control on goals: not available in advance or agani
- Focus on catastrophic forgetting
- Generalize to new coming tasks, few shot learning
- Meta-RL approaches
- Open-ended learning: unsufficiently established concept

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2018) Continual lifelong learning with neural networks: A review. *arXiv preprint arXiv:1802.07569*

Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. (2021) Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*

Hierarchical RL perspective: Overview



- The goal-conditioned policy is used by a higher level goal-selector policy
- The Feudal perspective and the option perspective
- The goal space can be discrete, in which case the higher level policy has discrete actions

Nachum, O., Gu, S., Lee, H., and Levine, S. (2018) Data-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:1805.08296*

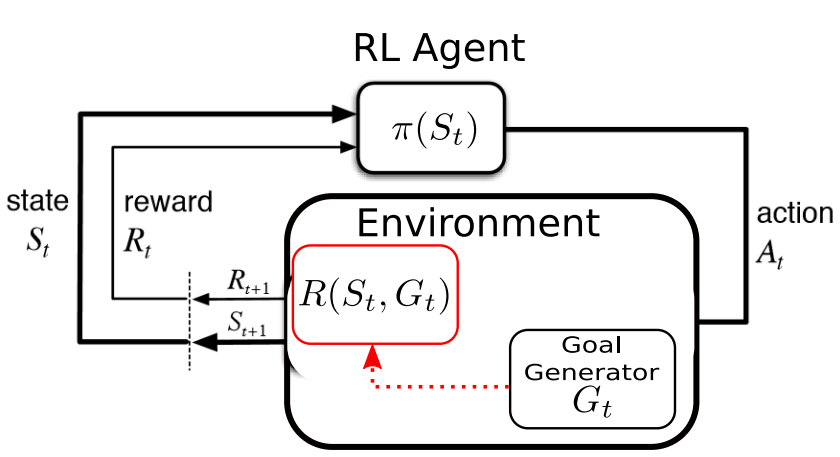
Bacon, P.-L., Harb, J., and Precup, D. (2017) The option-critic architecture. In *AAAI*, pages 1726–1734

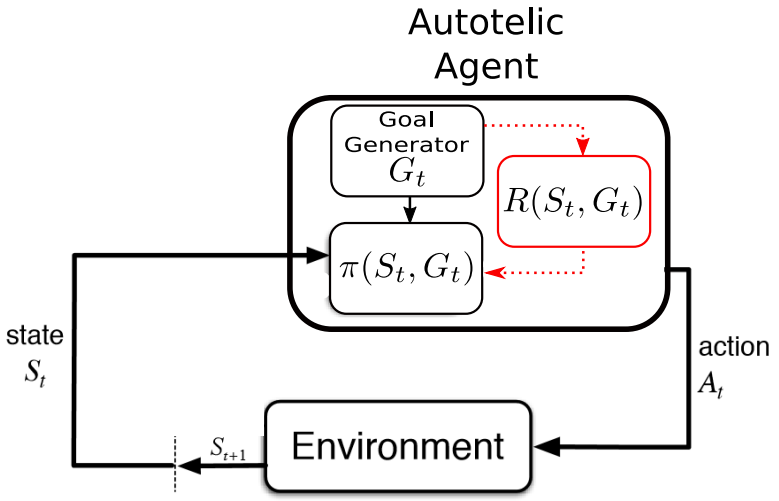
Pierrot, T., Perrin, N., Behbahani, F., Laterre, A., Sigaud, O., Beguir, K., and de Freitas, N. (2020) Learning compositional neural programs for continuous control. *arXiv preprint arXiv:2007.13363*

Unsupervised RL: outline

- RL agents are not autonomous: they depend on the design of an external reward function
- Reward engineering is a known challenge
- In GCRL, the desired goal can come from the environment or from the agent
- In standard multitask RL, goals come from the environment
- In autotelic learning, the agent generates its own goals
- The reward signal is internal $\rightarrow$ specific instance of unsupervised RL

GoalEnv vs Autotelic Agent view



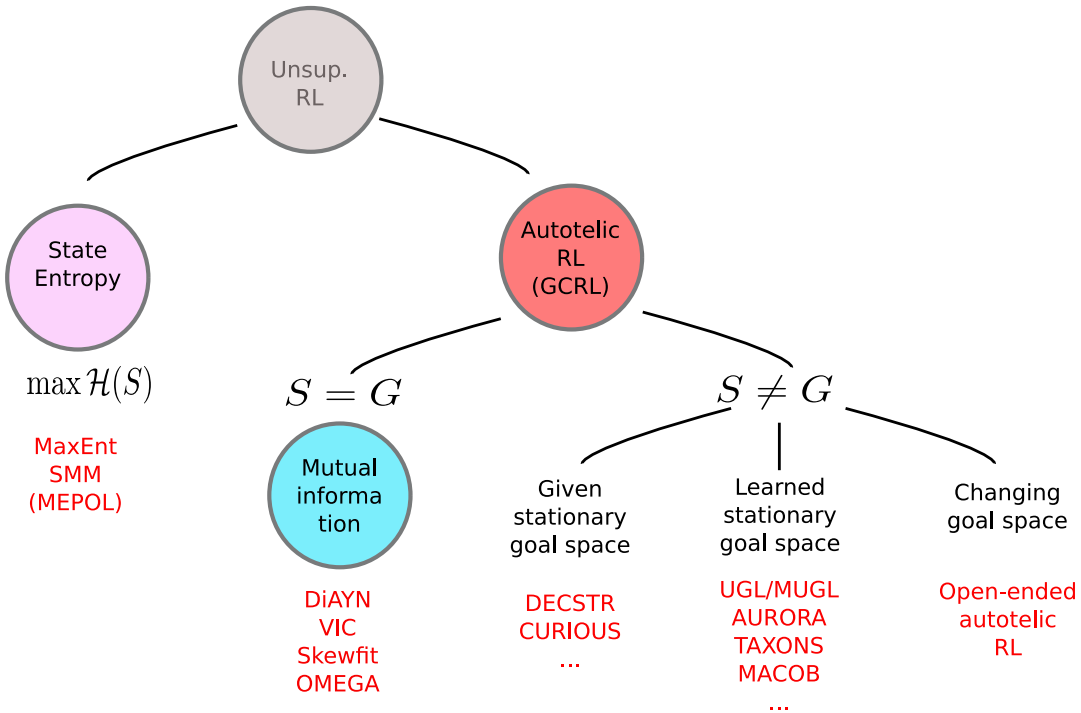


- In OpenAI gym, and SB3 (as most librairies?) the common view of GCRL
- Autotelic agents: agents equipped with forms of intrinsic motivations that enable them to represent, self-generate and pursue their own goals
- Goal generator based on: diversity, hierarchical RL, curriculum learning, social signals...

Colas, C., Oudeyer, P.-Y., Sigaud, O., Fournier, P., & Chetouani, M. (2019) CURIOUS: Intrinsically motivated multi-task, multi-goal reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 1331–1340

Colas, C., Karch, T., Sigaud, O., and Oudeyer, P.-Y. Intrinsically motivated goal-conditioned reinforcement learning: a short survey. *arXiv preprint arXiv:2012.09830*, 2020

Unsupervised RL and autotelic RL



- Open-ended autotelic RL: the agent defines its own goal spaces, its own state and actions spaces in a reward free environment (the ultimate framework!)

Any question?

Send mail to: Olivier.Sigaud@upmc.fr


Akakzia, A., Serris, O., Sigaud, O., and Colas, C. (2022).
Help me explore: Minimal social interventions for graph-based autotelic agents.
arXiv preprint arXiv:2202.05129.


Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. (2017).
Hindsight Experience Replay.
arXiv preprint arXiv:1707.01495.


Bacon, P.-L., Harb, J., and Precup, D. (2017).
The option-critic architecture.
In *AAAI*, pages 1726–1734.


Castanet, N., Lamprier, S., and Sigaud, O. (2022).
Stein variational goal generation for reinforcement learning in hard exploration problems.
arXiv preprint arXiv:2206.06719.


Chenu, A., Perrin-Gilbert, N., Doncieux, S., and Sigaud, O. (2021).
Selection-expansion: A unifying framework for motion-planning and diversity search algorithms.
arXiv preprint arXiv:2104.04768.


Colas, C., Karch, T., Sigaud, O., and Oudeyer, P.-Y. (2020).
Intrinsically motivated goal-conditioned reinforcement learning: a short survey.
arXiv preprint arXiv:2012.09830.


Colas, C., Karch, T., Sigaud, O., and Oudeyer, P.-Y. (2022).
Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey.
Journal of Artificial Intelligence Research, 74:1159–1199.


Colas, C., Oudeyer, P.-Y., Sigaud, O., Fournier, P., and Chetouani, M. (2019).
CURIOUS: Intrinsically motivated multi-task, multi-goal reinforcement learning.
In *International Conference on Machine Learning (ICML)*, pages 1331–1340.

Doncieux, S., Filliat, D., Díaz-Rodríguez, N., Hospedales, T., Duro, R., Coninx, A., Roijers, D. M., Girard, B., Perrin, N., and Sigaud, O. (2018).
Open-ended learning: a conceptual framework based on representational redescription.
Frontiers in Robotics and AI, 12.

Fang, M., Zhou, C., Shi, B., Gong, B., Xu, J., and Zhang, T. (2018).
DHER: Hindsight experience replay for dynamic goals.
In *International Conference on Learning Representations*.

Fang, M., Zhou, T., Du, Y., Han, L., and Zhang, Z. (2019).
Curriculum-guided hindsight experience replay.
Advances in neural information processing systems, 32.

Finn, C., Abbeel, P., and Levine, S. (2017).
Model-agnostic meta-learning for fast adaptation of deep networks.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

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

Lopes, M. and Oudeyer, P.-Y. (2012).
The strategic student approach for life-long exploration and learning.
In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–8. IEEE.

Nachum, O., Gu, S., Lee, H., and Levine, S. (2018).
Data-efficient hierarchical reinforcement learning.
arXiv preprint arXiv:1805.08296.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2018).
Continual lifelong learning with neural networks: A review.
arXiv preprint arXiv:1802.07569.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017).

Curiosity-driven exploration by self-supervised prediction.
arXiv preprint arXiv:1705.05363.

Pierrot, T., Perrin, N., Behbahani, F., Laterre, A., Sigaud, O., Beguir, K., and de Freitas, N. (2020).

Learning compositional neural programs for continuous control.
arXiv preprint arXiv:2007.13363.

Pitis, S., Chan, H., Zhao, S., Stadie, B., and Ba, J. (2020).

Maximum entropy gain exploration for long horizon multi-goal reinforcement learning.
In *International Conference on Machine Learning*, pages 7750–7761. PMLR.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015).

Universal value function approximators.
In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1312–1320. JMLR.org.

Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. (2021).

Open-ended learning leads to generally capable agents.
arXiv preprint arXiv:2107.12808.

Taylor, M. E. and Stone, P. (2009).

Transfer learning for reinforcement learning domains: A survey.
Journal of Machine Learning Research, 10(7).