Developmental machine learning
Machines that learn like children ... and help children learn better

Pierre-Yves Oudeyer
Project-Team INRIA-ENSTA-ParisTech FLOWERS
http://www.pyoudeyer.com
https://flowers.inria.fr
Twitter: @pyoudeyer
Deep RL system

Task specific biases + rules of the game

Code to compute External reward

The engineer

Lee Sedol
Autonomous Learning and Development in Human Infants

- How do developmental structures form?
- What is their role?
Cognitive sciences
models to understand better human development

Lifelong autonomous learning in robotics and AI

Many collaborations with researchers in
- Developmental psychology
- Neuroscience
- Robotics and AI
- Educational sciences

Flowers lab
Inria and Ensta ParisTech
France

Applications in educational technologies
Families of developmental «forces»

Body morphology and growth:
- Morphology, body growth and maturation
- Motor and perceptual primitives

Cognitive biases:
- Affordances
- Perceptual/linguistic categories grounded in action
- Hierarchies of actions, states, objectives

(McGeer, 1991)

-Intrinsic motivation, curiosity, active learning
- Autonomous/unsupervised collection of data
- Efficient learning of world models
- Self-organization of developmental trajectories

(Ly and Oudeyer, 2010)
Spontaneous active exploration

(Francis Vachon)
With Lauriane Rat-Fischer, S. Forestier and Alex Kacelnik
With Lauriane Rat-Fischer, S. Forestier and Alex Kacelnik
With Lauriane Rat-Fischer, S. Forestier and Alex Kacelnik
With Lauriane Rat-Fischer, S. Forestier and Alex Kacelnik
Intrinsic motivation, curiosity and active learning

- Intrinsic drive to reduce uncertainty, and to experiencing novelty, surprise, cognitive dissonance, challenge, incongruences, ...
- Optimal interest = optimal difficulty = neither trivial nor too difficult challenges

Berlyne (1960), White (1960), Kagan (1972), Csikszentmihalyi (1996), (Kidd et al., 2012), ...)
Towards a neuroscience of active sampling and curiosity

Jacqueline Gottlieb\textsuperscript{1,2,3} * and Pierre-Yves Oudeyer\textsuperscript{4,5}

Development of a unified formal and theoretical framework in psychology and neuroscience

The child as a sense-making organism: 
*Exploring to make good predictive models of the world and control it!*

- Measure of interestingness (intrinsic reward)
- Experiment selection
- Meta-model (model of novelty/uncertainty/improvement of M)
- Model learning M
- Prediction or goal achievement errors
- Prediction experiments
- Goal achievement experiments
- Sensori results

World
Robotic Playgrounds (Oudeyer et al., 2004; 2007)

Discovery of sensorimotor affordances

Discovery of speech communication

Discovery of nested tool use

Essential ingredients:

• Dynamic movement primitives (Schaal, Ijspeert et al., 2003, 2007)
• Object-based perceptual primitives (like infants, builds on prior perceptual learning)
• Self-supervised learning forward/inverse models with hindsight learning and episodic memory
• Curiosity-driven self-organization of learning curriculum through goal exploration

(Forestier et al., 2016, 2017)
What is an « interesting » learning experiment?

(Verbal) hypotheses from psychology and/or developmental biology:

- Cognitive homeostasis/auto-poiesis, high predictability (Varela and Maturana, explo. due to external perturbations)
- High novelty/high uncertainty? (many)
- Knowledge gap, cognitive dissonance? (Kagan, Festinger, Lowenstein)
- Intermediate novelty, intermediate complexity? (Berlyne, Kidd)
- Intermediate challenge? (White, Csikszentmihalyi)

Technical ideas from cognitive modeling or ML:

- High novelty/high uncertainty? (many)
- Surprise? (Itti and Baldi)
- Free energy? (Friston)
- Different forms of information gain/learning progress, e.g.:
  - KL-divergence between prior and posterior probabilistic model
  - Predictive information (Martius), predictive information gain (Little & Sommer)
  - Compression progress (Schmidhuber)
  - Empirical improvement of prediction or control (Oudeyer et al.)
The (absolute) Learning Progress hypothesis

Interestingness = \textit{proportional to empirical absolute learning progress} (absolute value of derivative)

→ Automated Curriculum Learning

(\textit{Oudeyer and Kaplan, 2003; 2007; Gottlieb et al., 2013; Oudeyer et al., 2016})

→ Few assumptions on underlying learning machinery and on match between biases and real world (as opposed to measures of learning progress based on KL-divergence measures)
Intrinsically Motivated Goal Exploration Processes

Context $s(t)$

Parameters of motor program $(DMP, RNN)$

Forward model(s) $F_i : \theta \rightarrow \hat{s}$

Inverse model(s) $I_i : \theta \rightarrow \hat{s}$

Mean speed of object $C_{mean}$

Vector of params of Bezié curve fitting traj. of obj. $A$

Trajectory: $\tau = \{s(t), a(t), \ldots, s(t + \Delta t)\}$

Behavioural descriptors over full trajectory (can be cost function measuring achievement of a complex property)

$\varphi = [\varphi_1(\tau), \varphi_2(\tau), \ldots, \varphi_i(\tau)]$

Prediction progress

Learned RNN embedding

Competition progress $\text{Intrinsically Motivated Goal Exploration (IMGEPs)}$
Goal sampling with Hierarchical Multi-Armed Bandits

Goal parameters space (continuous, high dim)

Automated Curriculum Learning

Learner (Inverse/forward model, Goal-conditioned Deep RL)

Utility = Absolute Learning progress

Empirical goal achievement errors

<table>
<thead>
<tr>
<th>Δe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δt</td>
</tr>
</tbody>
</table>

time

Probability to sample (e.g. with EXP3 alg.)

(Oudeyer and Kaplan, IEEE TEC 2007; Baranes and Oudeyer, RAS 2013; Colas et al., ICML 2019)
Exploring omni-directional locomotion

Policies: \( \pi \theta \) : oscillators in 8 motors, \( \theta \in [-1, 1]^{24} \)

Behavioral descriptors: \( \varphi \in \mathcal{R}^3 \) : translation and rotation over 3s

(Baranes and Oudeyer, IROS 2010, RAS 2013)
positions explorées dans l'intervalle axes selon l'axe relativement à l'espace atteignable). Nous pouvons aussi noter l'aspect asymétrique de cette répartition comprise dans délimite l'espace contenant les objectifs atteignables. 

\[ \text{Figure 6.5. Expérience 2 : Marche d'un Quadrupède via l'Utilisation de Synergies Motrices} \]

\[ \text{Figure 6.30 – exploration of policy:} \]

\[ \text{Figure 6.6. Expérience 3 : Contrôle d'une Canne à Pêche} \]

(Baranes and Oudeyer, Robotics and Autonomous Systems, 2013)
The Poppy Ergo robot has 6 motors, and moves with hardwired synergies that allow control of several other objects included in the environment, with which the agent cannot interact. Two analogical joysticks (Ultrastick 360) can be reached by the left arm and moved in any direction. The left arm has 20 degrees of freedom (DOF) and each JOINT controls (above a threshold) the intensity of the light of a LED circle around the arena. Finally, when the ball touches the border of the arena, a sound is produced and varied in pitch depending on ball rotational speed and radial extension. A tennis ball is freely moving in the blue arena which is slightly sloped so that the ball comes close to the center at the end of a movement. The speed of the ball controls (above a threshold) the intensity of the light of a LED circle around the arena. The outcome of an episode is a vector composed of the end position of the agent (rotation angle of the Ergo and of the ball around the center of the arena), and of the lamp (D), cart (1), shovel (5), pickaxe (2), and some other objects. The context is disabled in this experiment, the camera recording the ball trajectory (2) and the ball that controls some lights and sounds. Poppy is a robust and accessible open-source 3D printed robotic platform [Lapeyre et al., 2014].

### Behavioral descriptors

\[ \varphi = \begin{bmatrix} \varphi_i \end{bmatrix} \in \mathcal{R}^{310} \]

- Traj. Params. of object positions/sound/light
- \( \varphi_{hand} \)
- \( \varphi_{JoystickR} \)
- \( \varphi_{JoystickL} \)
- \( \varphi_{WhiteToy} \)
- \( \varphi_{Ball} \)
- \( \vdots \)
- \( \varphi_{distractor1} \)
- \( \varphi_{distractor2} \)
- \( \vdots \)
- \( \varphi_{distractor8} \)

(MACOB: Modular population-based IMGEPs)

(Forestier et al., 2017)
Curiosity-driven discovery of tool use

https://www.youtube.com/watch?v=NOLAwD4ZTWo

(Forestier et al., 2017)
Project Malmo (Minecraft) with neural net controllers

Rémy Portelas (Microsoft-Inria grant)
CURIOUS: intrinsically motivated modular multi-goal Deep RL

Modular UVFA (extended-UVFA)
Goal types and goal values:
- Move gripper to (x,y,z)
- Pick and place cube2 at (x,y,z)
- Push(cube1) at position (x,y)
- Stack cube1 over cube3 ...

(Colas et al., ICML19)
Recovery following a sensory failure.

CURIOUS recovers 95% of its original performance twice as fast as M-UVFA+HER.

Forgetting due to interferences among modules/goals

Mitigated thanks to LP-based re-exploration

Deep RL based IMGEPs (Curious) vs. Population-based IMGEPs:
+ better generalization
- Slower initial discoveries
How to learn (modular) representation of goals?

2 approaches:
1) Unsupervised learning (beta-VAEs) (Laversanne-Finot et al, CoRL 2018)
2) Leveraging language and its compositionality (Lair et al., Vigil workshop at Neurips 2019)
MUGL: Unsupervised goal learning and exploration

(Higgins et al., 2017)

Scene reconstruction/prediction

1): Unsupervised Goal Space Learning

β-VAE

2): Intrinsically Motivated Goal Exploration

3): Results

(Péré et al., ICLR 2018; Laversanne-Finot et al., CoRL 2018)
In each case, we ran 20 trials of 10,000 episodes each, for both the entangled and disentangled representations. In the entangled case, the interest is high only for the modules corresponding to latent variables encoding for the ball position (which is unknown by the agent, which does not distinguish the only ones that can be learned to control with motor commands. In the entangled case, the interest happens and those where it does not. It happens that modules that produce high learning progress correspond precisely to modules that can be controlled. As such, as a side benefit of using modular architectures, the monitoring of learning progress enables the agent to see which latent features can be controlled. This effect can be understood as follows: the entanglement introduces spurious correlations between the observations and the tasks in handcrafted and corresponds to the true degrees of freedom of the environment. In the other hand, the entanglement is defined in a way that allows to clearly identify the independent modules. When the representation possesses good disentanglement, the MGE algorithm is able to monitor the learnability of certain modules (possibly individual latent features, handcrafted, or automatically learned). In addition, the monitoring of learning progress enables the agent to discover which latent features can be controlled by its actions, and focus its exploration by setting the only ones that can be learned to control with motor commands. In the entangled case, the interest in the only ones that can be learned to control with motor commands. In the entangled case, the interest is high only for the modules corresponding to latent variables encoding for the ball position. In the disentangled case, one can see that the interest is high only for the modules corresponding to latent variables encoding for the position of the ball (which is unknown by the agent, which does not distinguish which are the independent controllable features of the environment). In the entangled case, the interest is high only for the modules corresponding to latent variables encoding for the position of the ball (which is unknown by the agent, which does not distinguish which are the independent controllable features of the environment).

**Benefits of disentanglement and modules**

The perspectives of this work are twofold. First, it would be interesting to show how the initial acquisition of abstract representations and skills can be leveraged by a curiosity-driven modular goal exploration architecture and its properties, they can be leveraged by a curiosity-driven modular goal exploration architecture and discovery controllable features during exploration, it would be interesting to re-use this knowledge to discover which latent features can be controlled. More specifically, we have shown that when the representation possesses good disentanglement, the modular architecture is actually detrimental to the performances, since each module encodes the position of the ball only partially and then, the representation is entangled, using a modular architecture is actually detrimental to the performances.

**Results**

The results are presented in Figure 4, which shows the evolution of the ratio of the number of cells visited with respect to all the cells in the discretized grid for different exploration noises. The figure includes two graphs: (a) Disentangled representation ($\beta$VAE) and (b) Entangled representation (VAE). The graphs show the discovery of independently controllable features.

**Hand-defined modular goal space**

- Hand-defined modular goal space
- Learned modular goal space
- Hand defined flat goal space
- Random parameter exploration
Using language as a cognitive tool to imagine new goals in curiosity-driven exploration.
Using language as a cognitive tool to imagine new goals in curiosity-driven exploration

Language Grounding through Social Interactions and Curiosity-Driven Multi-Goal Learning
Nicolas Lair, Cédric Colas, Rémy Portelas, Jean-Michel Dussoux, Peter Ford Dominey, Pierre-Yves Oudeyer
Vigil Workshop at Neurips 19
Understanding sentences by learning a reward function that predicts when it becomes true

State of body + objects + delta(state)

Sentence

Predicts whether sentence is true in that situation

Can be used as an internal reward function to measure whether an internally generated goal (= a sentence) is achieved by the goal-parameterized policy being learnt
Figure 3. **Reward function and policy learning**, a: F1-score of the reward function after convergence on $G^{\text{train}}$ (grey) and on the 5 types of goals from $G^{\text{test}}$ (colors). b: Average success rates on $G^{\text{train}}$ (plain) and $G^{\text{test}}$ (dashed) for various policy/critic architectures. c: Average success rates on $G^{\text{test}}$, split into the 5 generalization types. Mean ±1 std over 10 seeds for all figures.
Models of child development data
Self-organization of vocal development

DIVA Vocal tract model (Guenther et al.)

Two-layers of LP—based intrinsically motivated learning:
1) Active choice self-exploration vs. imitation
2) If self-exploration: active goal selection

Emergent developmental stages

(Oller, 2000)

0-3 mo: squeals, growls, yeals ...

3-7 mo: quasi-vowels

7-10 mo: language-independent proto-syllables

10 mo: influence by ambient language

12 mo: first words

Approximate age
Regularities and diversity of individual developmental trajectories

<table>
<thead>
<tr>
<th></th>
<th>No Phonation</th>
<th>Unarticulated</th>
<th>Articulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1</td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
</tr>
<tr>
<td>Simulation 2</td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td>Simulation 3</td>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
<td><img src="image9.png" alt="Graph" /></td>
</tr>
<tr>
<td>Simulation 4</td>
<td><img src="image10.png" alt="Graph" /></td>
<td><img src="image11.png" alt="Graph" /></td>
<td><img src="image12.png" alt="Graph" /></td>
</tr>
<tr>
<td>Simulation 5</td>
<td><img src="image13.png" alt="Graph" /></td>
<td><img src="image14.png" alt="Graph" /></td>
<td><img src="image15.png" alt="Graph" /></td>
</tr>
<tr>
<td>Simulation 6</td>
<td><img src="image16.png" alt="Graph" /></td>
<td><img src="image17.png" alt="Graph" /></td>
<td><img src="image18.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

FIGURE A2 | Developmental sequences emerging from the 9 simulations for the experiment described in section 3.1. Each subplot follows the same convention as in Figure 7. The simulations have been ordered, also in a subjective manner, from those which display a clear developmental sequence of the type No phonation → Unarticulated → Articulated to those less organized (from left to right, then top to bottom).
Curiosity-driven discovery of language as a tool to manipulate the environment

(Forestier and Oudeyer, CogSci 2017)
Teacher algorithms for curriculum learning of Deep RL in continuously parameterized environments

Rémy Portelas¹, Cédric Colas¹, Katja Hofmann², Pierre-Yves Oudeyer¹

¹Inria (FR)  ²Microsoft Research (UK)

CoRL 2019
Learners

Versus

Continuous set of tasks/envs
(through procedural generation)

\( \tau_0 \)

\( \tau_1 \)

\( \vdots \)

\( \tau_\infty \)
Methods - The CTS Framework
(CTS: Continuous Teacher-Student)

- The teacher samples parameters mapping to *distributions* of tasks/envs
  ⇒ creates *a curriculum where tasks/envs distributions evolve*
- The Deep RL Student is a black-box
- The parameter space may contain:
  - unfeasible subspaces
  - irrelevant dimensions
  - non-linear difficulty
ALP-GMM: sample tasks/envs distributions that maximize absolute learning progress
Example of mastered tasks after training

ALP-GMM + short DRL student

ALP-GMM + default DRL student

ALP-GMM + quadrupedal DRL student
Performance analysis on Hexagon Tracks
ALP-GMM
Good *generalization* to *diverse* obstacles

Random
Poor learning and generalization

[GitHub Repository](https://github.com/flowersteam/teachDeepRL)
Applications in educational technologies
Technologies for fostering efficient learning and intrinsic motivation

KidLearn project:
Personalization of teaching sequences (curriculum) in Intelligent Tutoring Systems
(Clement et al., Journal of Educational Data Mining, 2015; in prep.)

• Experiments with > 1000 children in more than 30 schools in Aquitaine
ZPDES-CO algorithm:
ALP + warm-start graph + final choice by child

(Clement et al., Journal of Educational Data Mining, 2015; in prep.)
Inside the zone of proximal development choose exercises stochastically according to the learning progress.

Exercise Type:
• aiming at different KC;
• or presented in a different modality;

(Clement, Roy, Oudeyer and Lopes, 2015, Journal of Educational Data Mining)
ZPDES-CO algorithm

Inside the zone of proximal development choose exercises stochastically according to the learning progress.
Inside the zone of proximal development choose exercises stochastically according to the learning progress.

After being able to solve A1, extend the ZPD.
Inside the zone of proximal development choose exercises stochastically according to the learning progress.

Always a probability of choosing other exercise types due to:
- individual characteristics
- problems in the knowledge graph
Learning impact

ZPDES-CO algorithm

Oracle algorithm (Pedagogical expert)
Motivational impact during learning sessions

Intrinsic motivation score session S3 (IMI questionnaire)

Oracle algorithm (Pedagogical expert)

ZPDES-CO algorithm
Take away

**Fundamental role of spontaneous developmental exploration**

- Autonomous goal exploration
- Driven by empirical learning progress measured at various scales of time and space

Scales to real world (high-dimensions, limited time, distractors)

- Organizes developmental trajectories
- Enables discoveries (tool use, language)

Can be used to guide human edTech design

(Gottlieb and Oudeyer, Nature Rev. Neurosc., 2018.; Oudeyer et al., 2016; )
Developmental autonomous learning

Thanks to:


Funding/Sponsors:
Why is (curiosity-driven) exploration of forward model less efficient than goal exploration?

- Redundancies
- Inhomogeneities

Forward model exploration: Knowing many ways to produce a few effects
Goal exploration: Knowing a few ways to produce many effects
Combining population-based and Deep-RL based IMGEPs

Population-based IMGEP (fast discoveries, episodic memory; Forestier et al., 2016)

Unsupervised modular goal space learning (Laversanne-Finot et al, 2018)

Monolithic Deep RL IMGEP (good generalization; Colas et al., 2019)

(Extension of Colas et al., GEP-PG: Decoupling exploration and exploitation in Deep RL, ICML 2018)
Curiosity applications beyond video games and robots:

Automated scientific discovery
Oil-in-Water Droplets Self-Organization

Automatized robot experiments

- 8 experiments running in parallel
- Specialized and stationary working stations
- Oils and surfactant handled separately
Intrinsically motivated goal exploration in a continuous game of life

Discrete Game of Life

Continuous Game of Life
Lenia, Bert Chan (2018)

Reinke, C., Etcheverry, M., Oudeyer, P-Y. (in prep) Intrinsically Motivated Exploration for Automated Discovery of Patterns in Morphogenetic Systems
Random Experiments
(mostly dead or uniform patterns)
Intrinsically Motivated Goal Exploration

32% spatially localized patterns (« animals »)
Goal exploration process

Intrinsically Motivated Exploration for Automated Discovery of Patterns in Morphogenetic Systems
Take away

Fundamental role of spontaneous developmental exploration

- Autonomous goal exploration
- Driven by empirical learning progress measured at various scales of time and space

Scales to real world (high-dimensions, limited time, distractors)

Can be used to guide human edTech design

Organizes developmental trajectories +

Enables discoveries (tool use, language)

(Gottlieb and Oudeyer, Nature Rev. Neurosc., 2018.; Oudeyer et al., 2016; )
Developmental autonomous learning

Thanks to:


**Funding/Sponsors:**

![Inria](link) ![ERC](link) ![European Research Council](link) ![ENSTA ParisTech](link) ![HFSP](link) ![Microsoft Research - Inria Joint Centre](link) ![Microsoft Research](link)
CURIOUS: intrinsically motivated modular multi-goal Deep RL

Controllable objects

Distractors

External world

LP-based sampling of modules and goals

Modular replay buffer with hindsight learning (module and goal substitution)

Modular UVFA
E.g. of modular goals:
Move gripper to (x,y,z)
Pick and place cube2 at (x,y,z)
Push(cube1) at position (x,y)
Stack cube1 over cube3 ...

(Colas et al., to appear at ICML19)
Recovery following a sensory failure.

CURIOUS recovers 95% of its original performance twice as fast as M-UVFA+HER.

Deep RL based IMGEPs (Curious) vs. Population-based IMGEPs:
+ better generalization
- Slower initial discoveries
Modeling overlapping waves of tool use development

(Siegler et al., 1996)

The overlapping waves model proposes that at any one age, children use multiple strategies; that with age and experience, they rely increasingly on more strategies (the ones with the higher numbers); and that development involves changes in use of existing strategies as well as discovery of new approaches.

(Forestier and Oudeyer, CogSci, 2016; ICDL-Epirob, 2016)
The Ergo-Robots (2012)

with
Mikhail Gromov
Mathematician

and
David Lynch
Film maker, artist

http://flowers.inria.fr/ergo-robots.php
Self-organization of culturally shared speech sounds

Most frequent vowel systems in human languages and emergent systems

<table>
<thead>
<tr>
<th>Vowels</th>
<th>3 vowels</th>
<th>4 vowels</th>
<th>5 vowels</th>
<th>6 vowels</th>
<th>7 vowels</th>
<th>8 vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>0.6%</td>
<td>1.9%</td>
<td>1.9%</td>
<td>3.1%</td>
<td>1.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td>a</td>
<td>0.1%</td>
<td>0.3%</td>
<td>2%</td>
<td>9%</td>
<td>1.7%</td>
<td>0%</td>
</tr>
<tr>
<td>i</td>
<td>3.2%</td>
<td>3.7%</td>
<td>28%</td>
<td>6.4%</td>
<td>6.3%</td>
<td>1.9%</td>
</tr>
<tr>
<td></td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>1.9%</td>
<td>1.2%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.1%</td>
<td>6%</td>
<td></td>
<td></td>
<td>1.6%</td>
<td></td>
</tr>
</tbody>
</table>

12% : frequency in human languages
19% : frequency in emergent systems

Models of the formation of speech sound systems in populations of individuals (de Boer, 2001; Oudeyer, 2006/19; Moulin-Frier et al., 2011)
Future research: learning to represent experiments
Studying the structure of free exploration in humans and monkeys

Intrinsically motivated exploration

Curiosity is a key element of human development, driving us to explore spontaneously novel objects, activities and environments. Many computational models of curiosity have been elaborated \[1\], \[5\]. Some of these models specifically targeted the relation between complexity, motivation and learning \[2\]. Another early example is McReynolds et al. who created the "curiosity box" where identical boxes, with different toys inside, are presented to young children \[3\]. Recently, an experiment was made with infants to study action selection guided by intrinsic motivations with a mechatronic board. In this study, subjects had to learn the relation between actions on pushbuttons and the opening of boxes \[4\]. They showed significant differences in the relation between pushbutton actions and the opening of the corresponding boxes.

Many psychology and neuroscience experiments in humans and monkeys, we still know little about the precise mechanisms that could be at play to drive spontaneous exploration \[8\].

The experimental setup is designed as a game setting human subjects into an intrinsically motivated activity \[11\]. Participants are exploring how to control an ellipsis displayed on the screen in front of them by moving their body joints tracked by a Kinect device. The interface shown to the user on a screen in particular, you can see the controlled ellipsis (in red) and the target one (in brown).

Subjects are exploring how to control an ellipsis displayed on the screen in front of them by moving their body joints tracked by a Kinect device.

We focus here on the intrinsic motivation mechanisms driving exploration, i.e. the processes allowing an agent to choose itself goals when freely involved in a task. Novel hypotheses have been formulated such that curiosity-driven sensorimotor exploration could be organized as to maximize uncertainty of novelty. Yet, experimental setups designed so far in the literature do not allow to separate between intrinsic motivation mechanisms and extrinsic goals.

We are interested in the properties of curiosity-driven exploration of a priori unknown sensorimotor spaces. More specifically, we are interested in the individual hypotheses. Furthermore, it is important to note that intrinsic motivation mechanisms could be influenced when subjects believe they could improve if they had five more tries. In each session, the subjects were asked to play each game once more and then asked him/her to believe they could improve if they had five more tries. In each session, the subjects were asked to play each game once more and then asked him/her to believe they could improve if they had five more tries. In each session, the subjects were asked to play each game once more and then asked him/her to believe they could improve if they had five more tries.

At the end of the sessions testing the 64-game version we conducted an additional procedure, administered without warning, requiring the subjects to play a minimum of 70 games and a minimum of 20 min. This dual requirement was meant to prevent a strategy of simply minimising the time on the task by selecting only the shortest games. Beyond these basic requirements, there were no additional constraints, and shows the maximum, minimum and average dot speed selected within a task group.

For the analyses in this study, the data was pooled across the sample. To generate the colormaps in the figure, we obtained the appropriate measure for one subject and then pooled across the sample. To generate the colormaps in the figure, we obtained the appropriate measure for one subject and then pooled across the sample. The values show the mean and s.e.m. across subjects. To examine the performance-dependent choice of the number of games in each of 6 performance bins, we computed the subject's distribution of selected speeds. For the analyses in this study, the data was pooled across the sample. To generate the colormaps in the figure, we obtained the appropriate measure for one subject and then pooled across the sample. The values show the mean and s.e.m. across subjects.
Combining population-based and Deep-RL based IMGEPs

Population-based IMGEP (fast discoveries, episodic memory; Forestier et al., 2016)

Unsupervised modular goal space learning (Laversanne-Finot et al., 2018)

Monolithic Deep RL IMGEP (good generalization; Colas et al., 2019)

(Extension of Colas et al., GEP-PG: Decoupling exploration and exploitation in Deep RL, ICML 2018)
Focus on modeling spontaneous curiosity-driven exploration in humans

- Understanding how it can be made to work for acquisition of motor skills in high-dimensional real world (robotic) bodies (Developmental robotics)
- Understanding how it links with developmental organization

**Psychology (1940-60)**
(Berlyne, White, Kaga, Festinger, ...)

**Theoretical biology and cognitive modeling**
Varela, Maturana (autopoïesis, 1974)
Oudeyer, Kaplan et al. (2003)

**Theoretical machine learning and RL**
Fedorov et al. (active learning, Optimal exp. Design, 1972)
Schmidhuber (LP based RL, 1991)
Barto, Singh et al. (IMRL, 2004)
Andreae et al. (novelty search with RL, 1978)

**Evolutionary computing**
Stanley et al., 2008; Mouret, Doncieux et al. (novelty search with GA/ES)

Related to various research lines

---

- Varela, Maturana (1974)
- Oudeyer, Kaplan et al. (2003)
- Fedorov et al. (active learning, Optimal exp. Design, 1972)
- Schmidhuber (LP based RL, 1991)
- Barto, Singh et al. (IMRL, 2004)
- Andreae et al. (novelty search with RL, 1978)
- Stanley et al., 2008; Mouret, Doncieux et al. (novelty search with GA/ES)

---

- Psychology (1940-60)
- Theoretical biology and cognitive modeling
- Theoretical machine learning and RL
- Evolutionary computing

---

- Varela, Maturana (autopoïesis, 1974)
- Oudeyer, Kaplan et al. (2003)
- Fedorov et al. (active learning, Optimal exp. Design, 1972)
- Schmidhuber (LP based RL, 1991)
- Barto, Singh et al. (IMRL, 2004)
- Andreae et al. (novelty search with RL, 1978)
- Stanley et al., 2008; Mouret, Doncieux et al. (novelty search with GA/ES)
Back to human experiments
How spontaneous exploration is structured during free play

- Exploration follows a growth in complexity actively controlled as predicted by models
- Factors influence exploration patterns: task difficulty, novelty, size of the choice space

(Baranes, Oudeyer and Gottlieb, 2014 Frontiers in Neuroscience)
(motor 1, motor2) → intensity (1 pix)
Learning and exploration dynamics
TARL: combining population-based and multi-goal Deep RL based IMGEPs (GEP-PG)

Population-based IMGEP → Replay buffer

- Fast exploration
- DDPG/TD3/SAC, ...
- CURIOUS, UVFA+HER

Good generalization
(1) Robots are useful to better conceptualize the impact of the body.

The example of walking
Morphology and self-organization of biped locomotion

Tad McGeer (McGeer, 1990), Nagoya Univ. (2005)
Morphological computation

(Ceccato et Cazalets, 2009)

• Collaboration with Labri/Univ. Bordeaux I

• Collaboration with J-R. Cazalets, Integrative Neuroscience Institute, Bordeaux

The Acroban humanoid (Ly, Lapeyre, Oudeyer, 2011, IROS)
Body: morphology, synergies and self-organization

A human-like bended leg shape reduces the motion amplitude on the upper body by 45% and increases the head stability by 30% (Humanoids 2013; IROS 2013)

Study of properties of various feet, including passive spring loaded articulations (Humanoids 2014)

(Ceccato et Cazalets, 2009)
Neuroscience, Univ. Bordeaux
From affordances to vocal interaction

Playground Experiments

- Autonomous learning of novel affordances and skills, e.g. object manipulation
- Self-organization of developmental trajectories, bootstrapping of communication
- Automatic formation of internal distinctive concepts for « self » vs « objects » vs « others »
- Regularities/diversity

→ New hypotheses for understanding information seeking and curiosity in infant development

(Oudeyer et al., 2007 IEEE TEC)
(Kaplan and Oudeyer, Front. Neuroscience, 2007)
Development of sensorimotor skills

Context $s(t)$

Parameters of motor program $(DMP, RNN)$

Trajectory:

$\tau = \{ s(t), a(t), \ldots, s(t + \Delta t) \}$

Behavioural descriptors over full trajectory (can be cost function measuring achievement of a complex property)

$\varphi = [\varphi_1(\tau), \varphi_2(\tau), \ldots, \varphi_i(\tau)]$

Mean speed of object $C$

Vector of params of Bezier curve fitting traj. of obj. $A$

Classifier counts of events encountered over traj.

Learned RNN embedding

(Oudeyer and Kaplan, 2007)
Experiments: Exploration and Learning in a Robotic Tool Use Setup

In order to benchmark different learning algorithms in a complex realistic environment with continuous policy and outcome spaces, we designed a real robotic setup composed of a humanoid arm in front of joysticks that can be used as tools to act on other objects. We show the running experimental setup in this video.

The code is available open-source together with the 3D shapes of printed objects.

4.1 Robotic Setup

The robotic setup has two platforms: in the first one, a Poppy Torso robot (the learning agent) is mounted in front of two joysticks (see Fig. 2). In the second platform, a Poppy Ergo robot (seen as a robotic toy) is controlled by the right joystick and can push a ball that controls some lights and sounds. Poppy is a robust and accessible open-source 3D printed robotic platform [Lapeyre et al., 2014].

Figure 2: Robotic setup. Left: a Poppy Torso robot (the learning agent) is mounted in front of two joysticks. Right: full setup: a Poppy Ergo robot (seen as a robotic toy) is controlled by the right joystick and can hit a tennis ball in the arena which changes some lights and sounds.

Robotic Arm

The left arm has 4 joints. The position of those joints at time $t$ is defined by the action $a_t$. Their bounds are defined so that the arm has a low probability to self-collide but can still reach a large volume, even on the left, top and behind the left shoulder to some extent. We use the framework of Dynamical Movement Primitives [Ijspeert et al., 2013] to generate smooth joint trajectories given a set of motor parameters. Each of the 4 joints is controlled by a DMP starting at the rest position of the joint (position $0$) and parameterized by $8$ weights: one weight on each of $7$ basis functions and one weight representing the end position of the joint trajectory (see Appendix B). Given $\theta$ (32 parameters between $-1$ and $1$) provided by the agent, the DMPs generates a policy roll-out by outputting a smooth 30-steps trajectory $a_{t0},...,a_{t_{\text{end}}}$ for the joints of the arm that once executed will translate into a 3D trajectory of the robotic hand for $5$ s. After producing each roll-out, the arm goes back in a rest position.

Tools and Toys

Two analogical joysticks (Ultrastick 360) can be reached by the left arm and moved in any direction. The 2D position of the joysticks (left-right and backward-forward axes) controls the Poppy Ergo robotic toy as follows. The left joystick does not control any variable. The Ergo robot has 6 motors, and moves with hardwired synergies that allow control of rotational speed and extension. The right joystick left-right axis controls in speed the rotation of the Ergo robot around the center of the second platform, which means that pushing the right joystick to the right with a small angle will move the Ergo towards the right with a small speed, and pushing the joystick with a higher angle will increase Ergo's rotational speed. The Ergo rotation angle is bounded in $\pi/\sqrt{2} \approx 1.178$ radians, and is reset to $0$ every $40$ iterations. The right joystick backward-forward axis controls the extension of the Ergo: if the joystick is in rest position, the Ergo stays in rest position. When the right joystick is moved forward, then the Ergo extends away from the center, using 3 of the 6 motors, and comes back when the joystick is released. A yellow tennis ball is freely moving in the blue arena which is slightly sloped so that the ball always comes close to the center at the end of a movement. The ball is 2D.

Action = 32D (32 continuous parameters of a DMP)

Perception = Trajectory of all objects: 310 continuous feature dimensions

MACOB: Modular population-based IMGEPs

(Forestier et al., 2017)