

# Intrinsic Motivation Systems for Autonomous Mental Development

Authors: Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner

Exploratory activities seem to be intrinsically rewarding for children and crucial for their cognitive development. **Can machine be endowed with such an intrinsic motivation system?**

Inspirations from human infant development:

- Development is progressive and incremental.
  - Challenge for robots: develop in an open-ended manner.
- Development is autonomous and active.
  - An intrinsic motivation system providing internal rewards during play experiences.

Exploratory activities seem to be intrinsically rewarding for children and crucial for their cognitive development. **Can machine be endowed with such an intrinsic motivation system?**

Inspirations from human infant development:

- Development is progressive and incremental.
  - Challenge for robots: develop in an open-ended manner.
- Development is autonomous and active.
  - An intrinsic motivation system providing internal rewards during play experiences.

This paper presents **Intelligent Adaptive Curiosity**, an intrinsic motivation system which pushes a robot towards situations in which it maximizes its learning progress.

Authors:

- Pierre-Yves Oudeyer, with Sony CSL Paris
- Frédéric Kaplan, with EPFL
- Verena V. Hafner, with Sony CSL Paris and TU Berlin

# Contents

- Existing intrinsic motivation systems
  - Learning machine, meta-learning machine, knowledge gain assessor
  - Different types of action selection
- Intelligent Adaptive Curiosity
  - Sensorimotor apparatus, regions and experts
  - Evaluation of learning progress and action selection
- Experiments
  - A simulated robot
  - The Playground Experiment
- Summary
- Discussion

# Intrinsic motivation systems

Existing computational approaches typically have the following two modules:

- A learning machine **M**:
  - Learns to predict the sensorimotor consequences when an action is executed in a sensorimotor context.
- A meta-learning machine **metaM**:
  - Learns to predict the errors in M's predictions.
  - The meta-predictions are basis for evaluating the potential interestingness of a situation.

The robot's actions are actively chosen according to some internal measures related to the novelty or predictability of the anticipated situation.

# Existing intrinsic motivation systems

Modules:

- **M**: learns to predict the sensorimotor consequences of actions.
- **metaM**: learns to predict the errors in M's predictions.

The robot's actions are actively chosen according to some internal measures related to the novelty or predictability of the anticipated situation.

Action-selection can be made depending on the predictions of M and metaM in order to:

- Maximize the prediction error of M. (Group 1)
- Maximize the decrease of M's mean error rate.
  - Compare the error rates in the close past. (Group 2)
  - Compare the error rates in situations which are similar, but not necessarily close in time. (Group 3)

The decrease of M's mean error rate can be monitored by a third module **KGA** (knowledge gain assessor).

# Existing intrinsic motivation systems

## Group 1: Error maximization

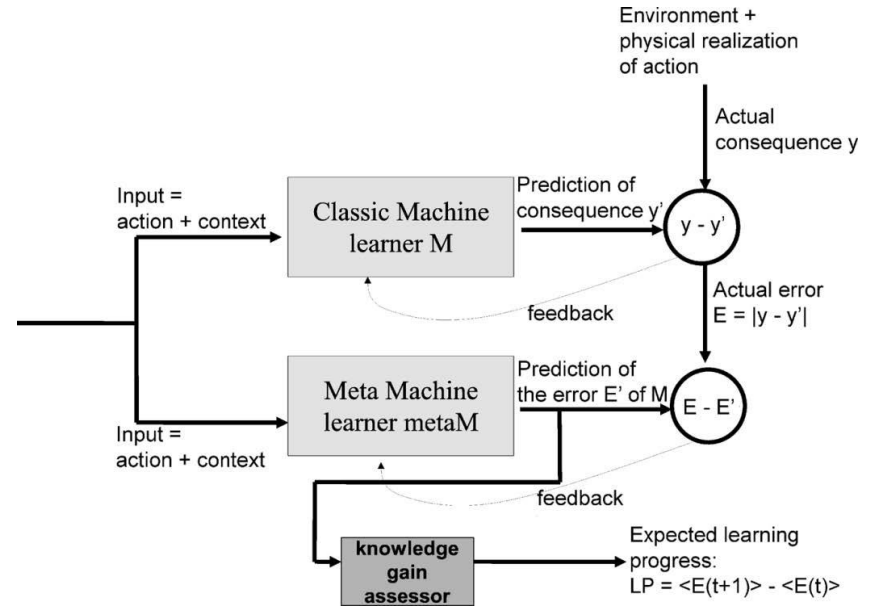
- Choose at each step the action that metaM predicts the largest error in M's prediction.

## Group 2: Progress maximization

- KGA evaluates the decrease of M's mean error rate in situations that are close in time.
- Choose the action that will lead to the greatest decrease of M's mean error rate.

## Group 3: Similarity-based progress maximization

- Builds a measure of similarity of situations
  - and ultimately an organization of the infinite continuous space of particular situations into higher level categories of situations.
- KGA evaluates the decrease of mean error rate in situations that are similar.



The architecture used in various models of Group 2 (progress maximization) and Group 3 (similarity-based progress maximization).

# Intelligent Adaptive Curiosity (IAC)

IAC is a drive to keep **the learning progress** maximal. It is an intrinsic motivation.

As a side effect, it pushes the robot toward novel situations in which things can be learned (**curiosity**),

and keeps the robot away from situations that are too predictable or too unpredictable (**intelligent**).

The situations that are attractive changes over time. Once something is learned, it will not provide learning progress anymore (**adaptive**).



# Intelligent Adaptive Curiosity (IAC)

IAC is a drive to keep **the learning progress** maximal. It is an intrinsic motivation.

As a side effect, it pushes the robot toward novel situations in which things can be learned (**curiosity**), and keeps the robot away from situations that are too predictable or too unpredictable (**intelligent**).

The situations that are attractive changes over time. Once something is learned, it will not provide learning progress anymore (**adaptive**).

## Implementation

- A memory which stores all the experiences encountered by the robot in the form of vector exemplars.
- The **sensorimotor space** is incrementally split into **regions**. Each region is characterized by an exclusive set of exemplars.
- Each region is associated with an **expert** (learning machine) making predictions for situations in the region.
- The prediction errors (associated with the region) are used for **evaluation of the potential learning progress** that can be gained by going in to situations in the region.
- In a situation, the robot selects the **action** that leads to a new situation with maximal expected learning progress.

# IAC: Sensorimotor apparatus, regions and experts

Sensorimotor apparatus:

- $S(t)$  - sensors
- $M(t)$  - action/motor parameters
- $SM(t)$  - sensorimotor context, concatenation of  $S(t)$  and  $M(t)$

IAC equips the robot with a memory of all exemplars ( $SM(t)$ ,  $S(t + 1)$ ) it has encountered.

The sensorimotor space is incrementally splitted into **regions**. Each region has an exclusive set of exemplars.

# IAC: Sensorimotor apparatus, regions and experts

Sensorimotor apparatus:

- $S(t)$  - sensors
- $M(t)$  - action/motor parameters
- $SM(t)$  - sensorimotor context, concatenation of  $S(t)$  and  $M(t)$

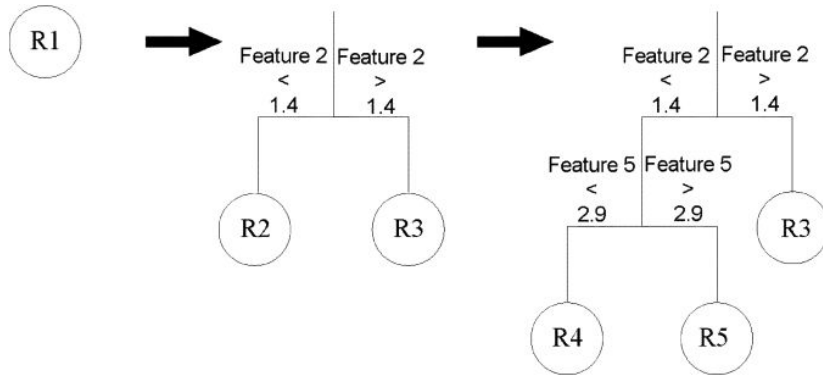
IAC equips the robot with a memory of all exemplars ( $SM(t)$ ,  $S(t + 1)$ ) it has encountered.

The sensorimotor space is incrementally splitted into **regions**. Each region has an exclusive set of exemplars.

Method: recursively for each region, split the region into two when criterion  $C_1$  is met, by separating the sensorimotor space with criterion  $C_2$ .

- When to split a region into two? (criterion  $C_1$ )
  - The number of exemplars exceeds  $T = 250$ .
- How to make the split? (criterion  $C_2$ )
  - Separate the sensorimotor space according to feature  $j$  in  $SM(t)$  and the cutting value  $v_j$ .
  - The cutting dimension  $j$  and value  $v_j$  are chosen such that the resulting two splits have the smallest sum of variances for  $S(t + 1)$ .

# IAC: Sensorimotor apparatus, regions and experts



Example of how the sensorimotor space is split into subspaces, called “regions”.

Method: recursively for each region, split the region into two when criterion  $C_1$  is met, by separating the sensorimotor space with criterion  $C_2$ .

- When to split a region into two? (criterion  $C_1$ )
  - The number of exemplars exceeds  $T = 250$ .
- How to make the split? (criterion  $C_2$ )
  - Separate the sensorimotor space according to feature  $j$  in  $SM(t)$  and the cutting value  $v_j$ .
  - The cutting dimension  $j$  and value  $v_j$  are chosen such that the resulting two splits have the smallest sum of variances for  $S(t + 1)$ .

# IAC: Evaluation of learning progress

Each region  $R_n$  is responsible for monitoring the evolution of error rate in the anticipations of the consequences of an action when the associated context  $SM(t)$  is covered by the region.

- The error rate  $e_n(t + 1)$  is the squared error of expert  $E_n$ 's prediction for the sensory state  $S(t + 1)$ .
- Past error rates  $e_n(t)$ ,  $e_n(t - 1)$ , ...,  $e_n(0)$  are stored with  $R_n$ .
- The **learning progress** that has been achieved through the transition from  $SM(t)$  to the context with a perceptual vector  $S(t + 1)$  is computed as the **smoothed derivative of the error curve** of  $E_n$  corresponding to the recent exemplars.

Mean error rates:

$$\langle e_n(t + 1) \rangle = \frac{\sum_{i=0}^{\theta} e_n(t + 1 - i)}{\theta + 1}$$

$$\langle e_n(t + 1 - \tau) \rangle = \frac{\sum_{i=0}^{\theta} e_n(t + 1 - \tau - i)}{\theta + 1}$$

Decrease in mean error rate:

$$D(t + 1) = \langle e_n(t + 1) \rangle - \langle e_n(t + 1 - \tau) \rangle$$

Learning progress:

$$L(t + 1) = -D(t + 1)$$

# IAC: Action selection

- Each time an action is performed in a context, the **internal reward** depends on how much learning progress has been achieved:  $r(t) = L(t)$ .
- The intrinsically motivated robot aims to maximize the internal reward, which is formulated as **maximization of the return** (future expected rewards).
- The problem is **simplified** to only maximize the expected reward at  $t + 1$ , evaluated by the learning progress achieved in  $R_n$  the last time  $R_n$  and  $E_n$  processed a new exemplar:  $E\{r(t + 1)\} \approx L(t - \theta_{Rn})$ .
  - The paper focuses on the study of the learning progress. Using a complex reinforcement machinery brings complexity and biases.

Learning progress:

$$L(t + 1) = -D(t + 1)$$

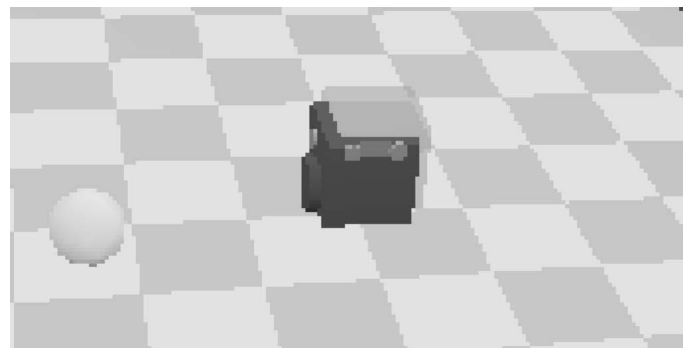
$$D(t + 1) = \langle e_n(t + 1) \rangle - \langle e_n(t + 1 - \tau) \rangle$$

# Experiment with a simple simulated robot

- The robot's movement and sound can be controlled by setting the speed of the left motor ( $l$ ), speed of the right motor ( $r$ ), and frequency of the sound ( $f$ ).
  - Motor vector  $M(t) = (l, r, f)$
- A toy moves according to the sound:
  - $f_1 = [0; 0.33]$ , moves randomly
  - $f_2 = [0.34; 0.66]$ , stops moving
  - $f_3 = [0.67; 1]$ , jumps into the robot
- The robot perceives distance between the toy and itself with infrared sensors.
  - Sensory vector  $S(t) = (d)$
- It tries to learn the mapping  $f : \mathbf{SM}(t) = (l, r, f, d) \mapsto \mathbf{S}(t+1) = (\tilde{d})$ .
- The intrinsically motivated robot will act to maximize its learning progress in terms of predicting the next toy distance.

How does the IAC system work in a continuous sensorimotor environment with inhomogeneous parts?

- unlearnable part
- easy to learn part
- complex and learnable part



A robotic simulation implemented by Webots. A two-wheeled robot moves in a room and emits a sound. An intelligent toy (represented by a sphere) moves according to the sound the robot produces.

# Simulated robot

Motor control, perception, and action perception loop:

- The robot moves in a room and emits a sound:  $M(t) = (l, r, f)$ .
- The toy moves randomly when the sound is in the zone  $f_1$ , stops moving if the sound is in  $f_2$ , and jumps into the robot if the sound is in  $f_3$ .
- The robot perceives the distance between the toy and itself:  $S(t) = (d)$ .
- The robot tries to learn the mapping:

$$f : \mathbf{SM}(t) = (l, r, f, d) \mapsto \mathbf{S}(t+1) = (\tilde{d}).$$

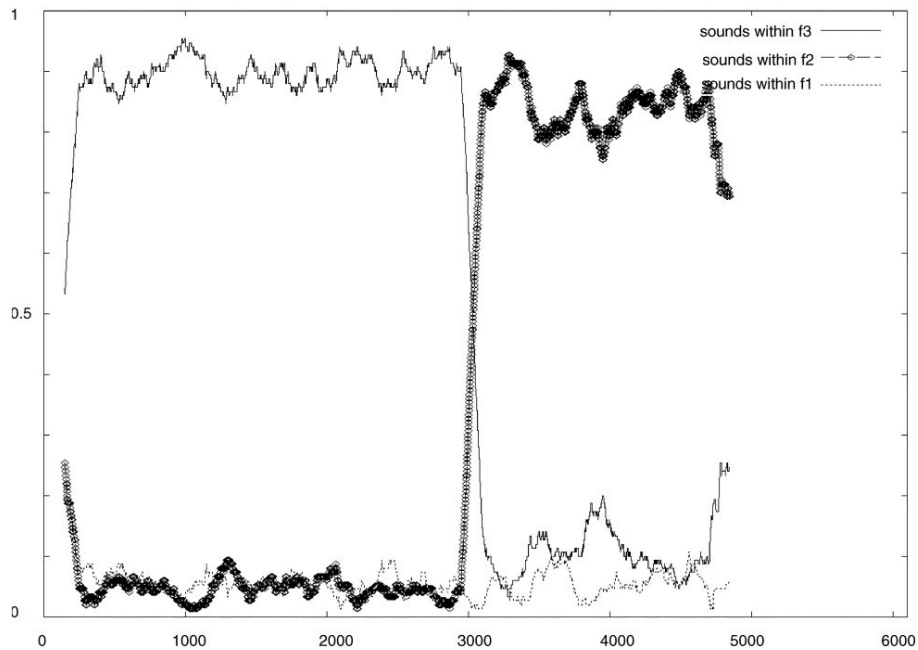
The sensorimotor space has **three zones with different complexities** (initially unknown to the robot):

- $f_1$  - unlearnable.
- $f_2$  - predictable and learnable, complex and dependent on the wheel speeds.
- $f_3$  - easy to learn and predict, the distance is always zero plus a noise.

The results will show the robot manages to autonomously discover the three zones, evaluate their complexity, and exploit this information to organize its behavior.



# Results



Percentage of time the robot spent in the zones  $f_1$ ,  $f_2$ , and  $f_3$ , measured for 5,000 time steps.

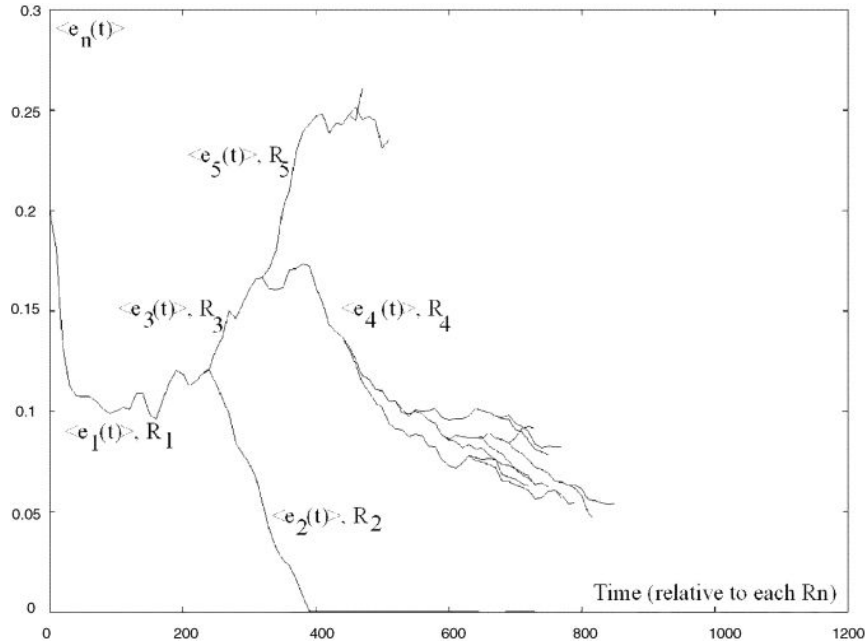
From an external point of view, there are three zones in the sensorimotor space:  $f_1$  ( $[0, 0.33]$ , unlearnable),  $f_2$  ( $[0.34, 0.66]$ , complex and learnable), and  $f_3$  ( $[0.67, 1]$ , predictable).

The percentage of time the robot spent in  $f_1$ ,  $f_2$  and  $f_3$  evolves over time.

- Stage 1: Initially the robot produces sounds with frequencies in  $[0, 1]$  uniformly.
- Stage 2: The robot concentrates on  $f_3$ .
- Stage 3: The robot concentrates on  $f_2$ .

The robot consistently avoids the situations where nothing can be learned, begins by easy situations, and then shifts to a more complex situation.

# Results



Evolution of mean error rates for all regions over time.

From the robot's point of view, the mean error rates in each region evolve over time.

- The first split occurred after the first 250 time steps.
  - $R_1$  is split into  $R_2$  and  $R_3$  using the feature  $f_2$  and the cutting value 0.35.  $R_2$ :  $f_3$  and part of  $f_2$ ;  $R_3$ :  $f_1$  and part of  $f_2$ .
- In  $R_2$  the error rate has a sharp decrease, while in  $R_3$  it has a sharp increase.
  - During this period, the robot concentrates on emitting sounds in  $f_3$ .
- The error rate in  $R_2$  plateaued at some point.
  - The robot switched from emitting sounds in  $f_3$  to emitting sounds in  $f_2$ .
- The robot learns to predict the consequences of varying the speeds, in addition to the sound frequency.
  - $R_4$  corresponds to  $f_2$  and  $R_5$  corresponds to  $f_3$ .
  - $R_4$  is split into more regions using the speed dimensions.

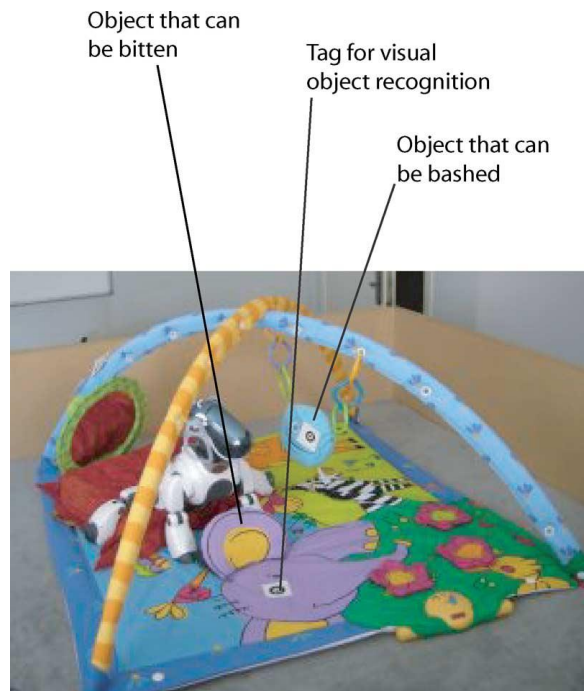
# The Playground Experiment



# The Playground Experiment

- The robot is equipped with three basic motor primitives:
  - *turning head* - controlled by pan ( $p$ ) and tilt ( $t$ ) of the head.
  - *bashing* - controlled by strength ( $b_s$ ) and angle ( $b_a$ ) of the leg movement.
  - *crouch biting* - controlled by depth ( $d$ ) of the crouching.
  - Motor vector  $\mathbf{M}(t) = (p, t, b_s, b_a, d)$
- It has an object visual detection sensor, a biting sensor, and an oscillation sensor, that take binary values  $O_v$ ,  $B_i$ ,  $O_s$  respectively.
  - Sensory vector  $\mathbf{S}(t) = (O_v, B_i, O_s)$
- It tries to learn the mapping:

$$f : \mathbf{SM}(t) = (p, t, b_s, b_a, d, O_v, B_i, O_s) \\ \mapsto \mathbf{S}(t + 1) = (\widetilde{O}_v, \widetilde{B}_i, \widetilde{O}_s).$$

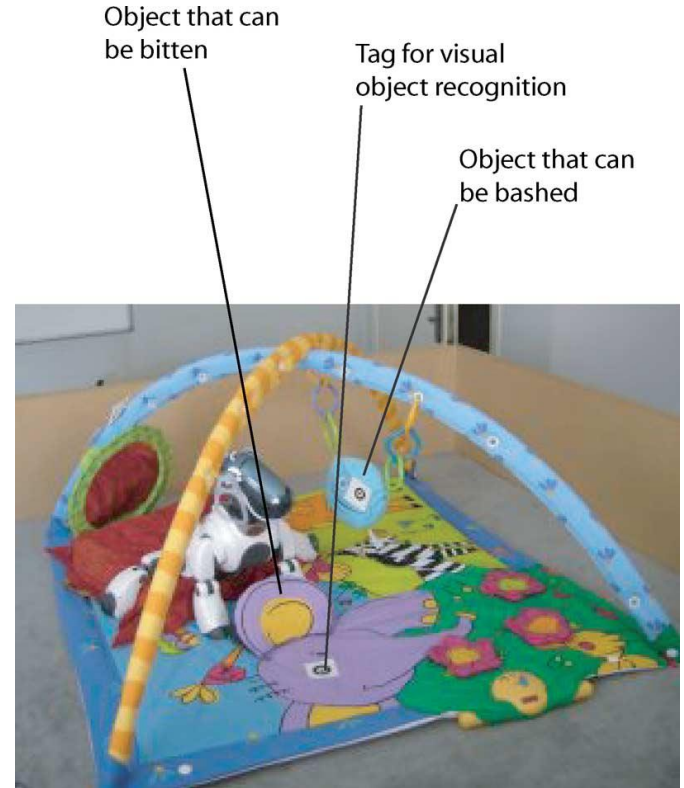


A Sony AIBO robot put on a play mat with toys that can be bitten, bashed, or visually detected. The environment is similar to the ones in which two- or three-month old children learn their first sensorimotor skills.

# The Playground Experiment

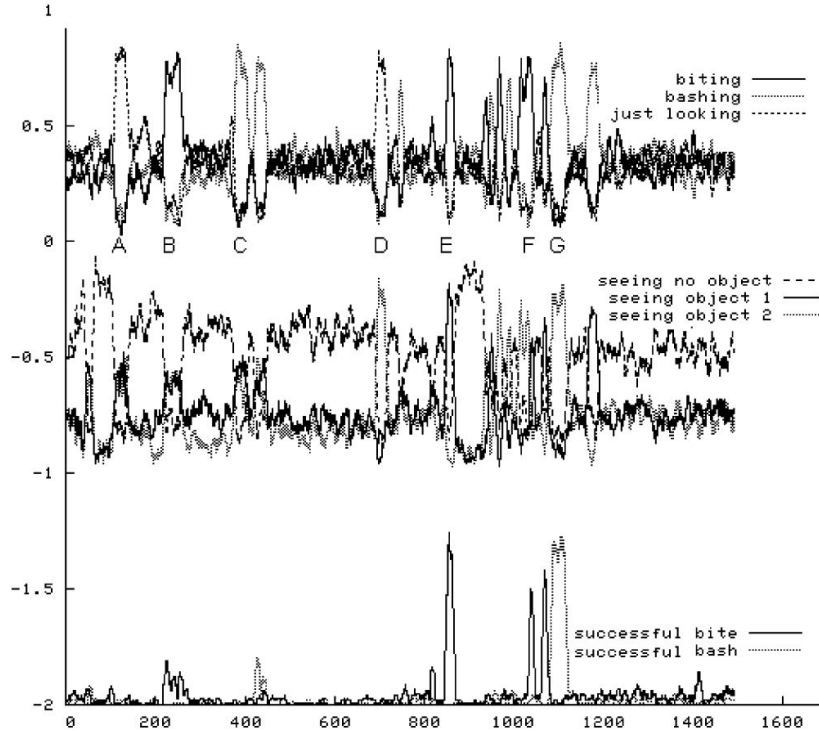
**Sensorimotor affordances** (initially unknown to the robot):

- The values of object visual detection sensor ( $O_v$ ) is correlated with the values of pan and tilt.
- The values of the biting or oscillation sensors ( $B_i$  and  $O_s$ ) can become 1 only when biting or bashing actions are performed toward an object.
- Some objects are more prone to provoke changes in  $B_i$  and  $O_s$  with certain kinds of actions.
- To get a change in  $O_s$ , bashing in the correct direction is not enough, and it also needs to look in the right direction.



A Sony AIBO robot put on a play mat with toys that can be bitten, bashed, or visually detected. The environment is similar to the ones in which two- or three-month old children learn their first sensorimotor skills.

# Results



Evolution of the percentage of several kinds of actions over time. The horizontal axis represents time and three vertical axes representing the percentage are stacked. Object 1: bitable. Object 2: bashable.

Top 3: Percentage of actions that

- 1) involves biting
- 2) involves bashing
- 3) is just looking

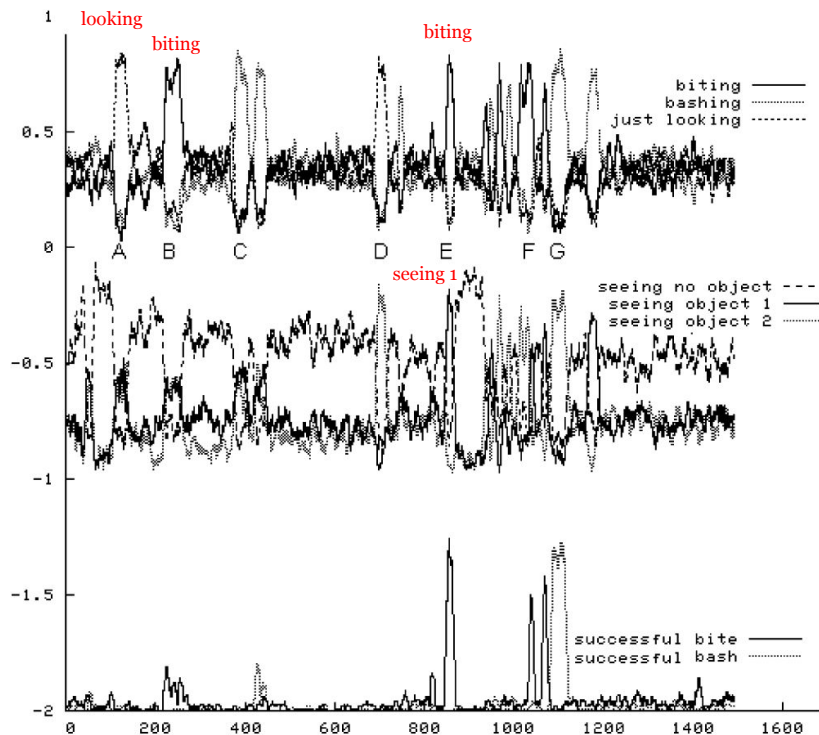
Mid 3: Gaze of the robot. Percentage of actions that is looking towards

- 1) no object
- 2) the bitable object ("object 1")
- 3) the bashable object ("object 2")

Bottom 2: Percentage of

- 1) successful biting actions (provoke a "1" value on the biting sensor)
- 2) successful bashing actions (provoke a "1" value on the oscillation sensor)

# Results



Evolution of the percentage of several kinds of actions over time. The horizontal axis represents time and three vertical axes representing the percentage are stacked. Object 1: bitable. Object 2: bashable.

A peak means that at the moment the robot focuses its activity and attention on a small subset of the sensorimotor space.

Observations:

- Before the first peak, successful bite or bash and seeing any object are rare, and all actions are produced equally often.
- Peak (A): the robot focuses on moving its head around and stops biting and bashing.
  - Several regions have been created. The region that corresponds to the sensorimotor loop of “just looking” has the highest source of learning progress.
- Peak (B): a focus on biting, that doesn’t co-occur with looking towards the bitable object.
- Peak (E): a focus on biting, coupled with a peak of looking towards the bitable object, and a peak of successful biting.
  - The robot uses the action primitive with the right affordances.


# Summary


- An intrinsic motivation system IAC that drives the development of a robot in continuous noisy inhomogeneous spaces.
- Experiments:
  - Simulated robotic setup: how IAC works and provokes behavioral and cognitive development.
  - Physical robotic setup: IAC allows the robot to autonomously generate a developmental sequence.
- Limits of the system:
  - Simplification to optimize only the immediate reward.
  - Rather simple sensorimotor space.




# Questions from Piazza discussion

Can the intrinsic reward in the IAC system help improve reinforcement learning methods?  
Does it combines well with goal-directed objective function? @45\_f3

 **Yuhang Mei** 8 hours ago  
Designing an appropriate reward function is a hard part of reinforcement learning. I wonder whether the IAC system proposed by this paper can be embedded in this part to help boost the effectiveness of reinforcement learning.  
helpful! | 0

 **Oliver A Wang (oliveraw)** 6 hours ago  
I found the IAC system very interesting as an implementation of intrinsic reward (reward for pure curiosity). I guess the next step might be to explore if this combines well with goal-directed objective functions? For example, if the goal is for a robot to navigate across the room to some target, does the IAC help to create a robust representation? Is it as simple as combining the reward functions? I'm wondering if anyone has seen research in this area or has any additional resources to look at.  
helpful! | 0

For models that perform best when learned in a task-aware manner, how to balance the intrinsic reward and the extrinsic reward? @45\_f8

 **Zixuan** 2 hours ago  
I agree that the curiosity-based intrinsic reward is definitely an important component for the self-directed exploration of an intelligent agent. But I wonder how to balance the intrinsic reward and extrinsic reward (task reward). For a lot of the papers that I've seen, it is shown that the model will perform best when it is learned in a task-aware manner, while intrinsic reward is task-independent.  
  
I wonder if there's a good way to reconcile the conflicts.  
helpful! | 0

Thank you

# Discussion summary

# 1. Intuition behind the modules in intrinsic motivation systems

As illustrated in Fig. 1 in the paper, intrinsic motivation systems typically comprise a learning machine, a meta-learning machine, and optionally a knowledge gain assessor.

How do babies learn to crawl, walk, or run? Rather than extrinsic reward like how far they have walked without falling, their motivations are to keep the play and exploration going. To develop an intelligent robot capable of a variety of skills, it needs to know when to learn each skill. This motivates the second module, the meta-learning machine. The robot can make predictions to decide if it has mastered a skill and needs to move on to something novel.

Novelty by itself is not enough since it can result in exploring novel situations without making progress. Like in infant development some skills can only be learned when the associated cognitive and morphological structures are ready. This motivates the third module, knowledge gain assessor. When the robot encounters situations where it cannot gain knowledge, it goes for other options.

(Source: In-class discussion, Piazza @45\_f7)

## 2. Measuring similarity

There are many ways to capture the knowledge gain intuition. The paper chooses similarity-based progress maximization. For example, walking is more similar to crawling, than running is to crawling. And babies can base on the skills they have already learned and extend a little bit to make progress. How to measure this similarity?

Computationally, we can try to group situations based on what the goal is, like moving forward (walking and crawling) or moving forward quickly (running). We can also consider other aspects, like the center of gravity, perception of depth, and other sensory signals.

(Source: In-class discussion)

### 3. Intrinsic reward in reinforcement learning

Designing an appropriate reward function is a hard part of reinforcement learning. Recent research has been incorporating “curiosity” into RL systems, e.g. 1) [\*Curiosity-Driven Exploration by Self-Supervised Prediction\*](#) 2) [\*Exploration by Random Network Distillation\*](#) 3) [\*Learning Latent Dynamics for Planning from Pixels\*](#). The idea is to use reinforcement learning to guide information gathering for neural networks, whose results are provided to the reinforcement learning agent as reward.

(Source: In-class discussion, Piazza @45\_f3 @45\_f8 @45\_f10)

## 4. Evaluating behavior complexity

Using a simple reward maximizing learning progress, the system in the paper developed a steady emergence of more complex behaviors. Despite that evaluating complexity objectively can be difficult, the paper illustrated in their experiments task-independent evaluation methods for the evolution of the system's complexity, such as identifying stages in developmental sequences, and measuring the internal variables of the robot.

(Source: Piazza @45\_f2)

## 5. Modeling the sensorimotor space

The sensory space and the motor space are not treated separately, reflecting the theories in previous readings that perception and motor functions to a certain degree are tied together. Mathematically, the sensory and motor vector are concatenated as the sensorimotor context. As a result, the system can develop regions according to both features.

(Source: Piazza @45\_f12)



## 6. Limitations of imitation learning

Imitation learning in robotic settings is about trying to copy the motor from the teacher, and it only works if the robot has the same morphology as the teacher. One example is designing airplanes. Mimicking the living things that can fly would lead to flapping wings. Without a developmental body but with a fixed morphology, we need to consider the computational principle behind it.

It can be helpful to characterize the sensory information. Like when driving in a rental car, knowing how the actions will lead to changes in your sensory systems can help you transfer driving skills to this new environment.

(Source: In-class discussion)