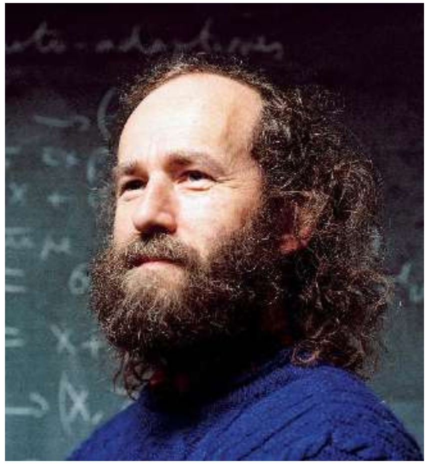
Explainable Reinforcement Learning through Genetic Programming

Alessandro Leite
TAU, INRIA Saclay, LISN
Île-de-France

July 29th, 2022

74ª Reunião Anual da SBPC
24 a 30 de julho de 2022

Credits



Mathurin Videau[1,2]

Olivier Teytaud[2]

Marc Schoenauer[1,2]

Videau, Mathurin, Alessandro Leite, Olivier Teytaud, and Marc Schoenauer. "*Multi-Objective Genetic Programming for Explainable Reinforcement Learning*." In Genetic Programming, edited by Eric Medvet, Gisele Pappa, and Bing Xue, 278–93. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022.

[1]TAU, INRIA Saclay, LISN
[2] Meta AI Research

AI techniques have been used across different domains

Play (and win) games

AlphaGo

DeepMind

Answer queries

IBM

Watson

Debate

Project Debater

IBM

Recognise speech

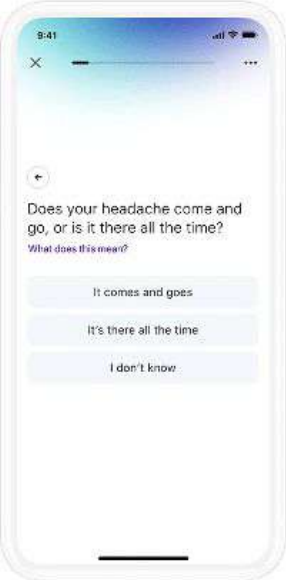
Hey Siri, call Mum

Recognise faces

facebook

Translate across languages

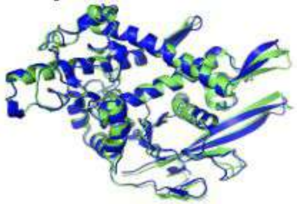
Google Translate



Detect & Diagnose Diseases

babylon

Predict protein structures



Drive vehicles

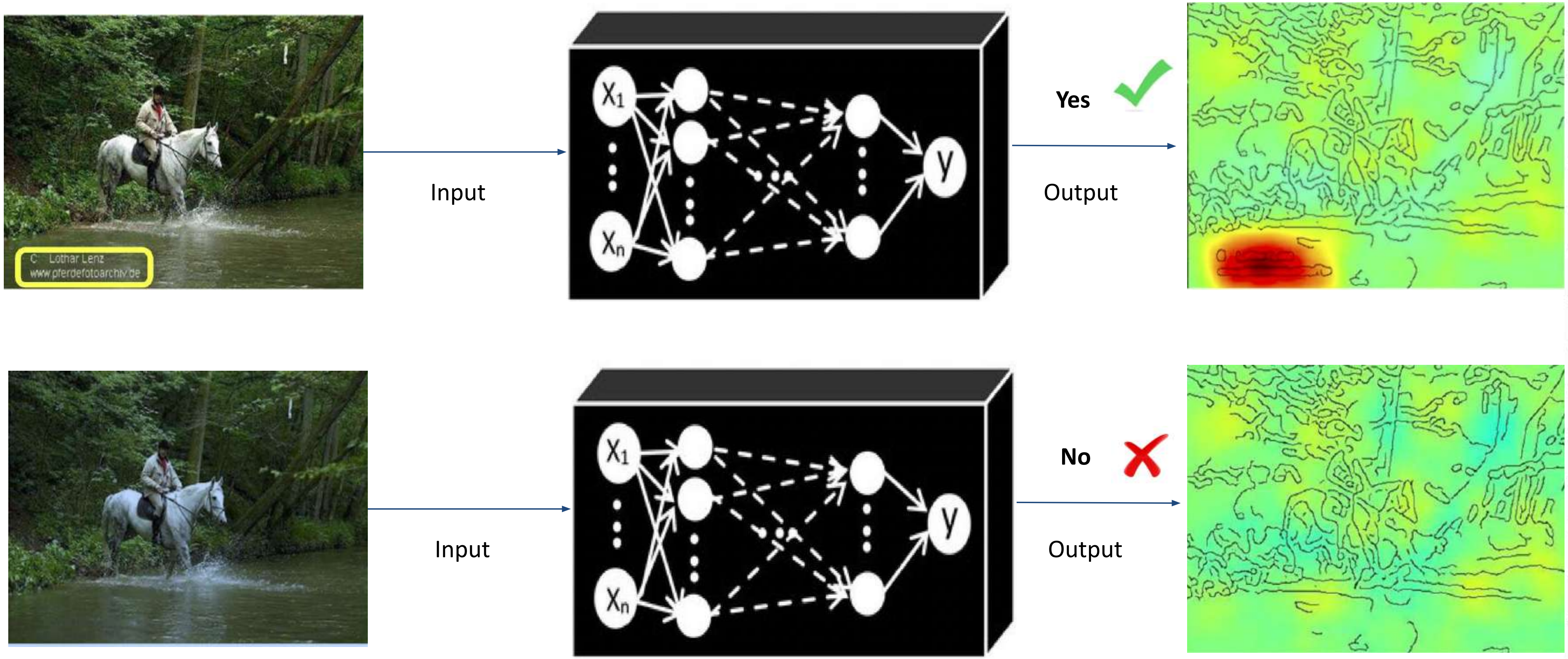
WAYMO

Vacuum clean

dyson

Model may be right for the wrong reasons (Clever Hans)

AI models may rely on proxies to produce an output



Sebastian Lapuschkin et al. "*Unmasking Clever Hans predictors and assessing what machines really learn*". In: Nature communications 10.1 (2019), pp. 1–8.

Why do we need explanations?

Explanations

- reflect an attempt to communicate an understanding[1]
- create trajectories, expanding individuals' understanding in real-time
- may highlight incompleteness
- relate the event being explained to principles, invoking causal relations[2]
- answer a “*why question*” justifying an event

[1]Frank C Keil. “Explanation and understanding”. In: Annu. Rev. Psychol. 57 (2006), pp. 227–254.
[2]Tania Lombrozo. “Explanation and abductive inference”. In: The Oxford Handbook of Thinking and Reasoning. Ed. by Keith J. Holyoak and Robert G. Morrison. Oxford University Press, 2012

eXplainable AI (XAI) provides tools to explain ML models

Interpretability (intrinsic property of a model)

- It describes the internals of a system in a way that is understandable to humans[1]
- It must employ a vocabulary that is meaningful for a human observer

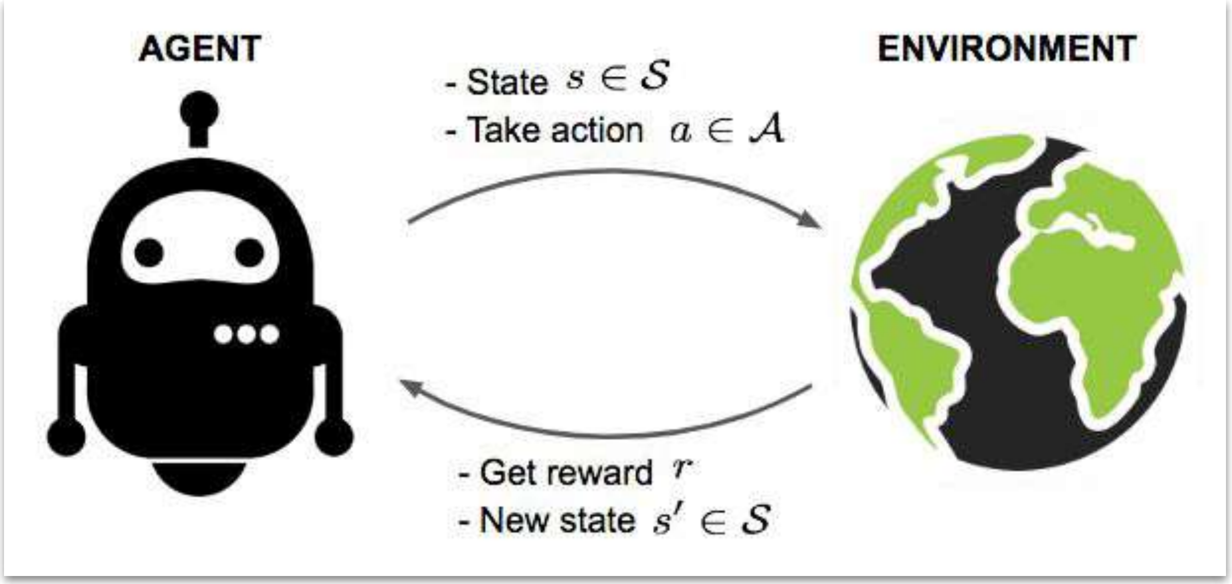
Explanation (post-hoc analysis)

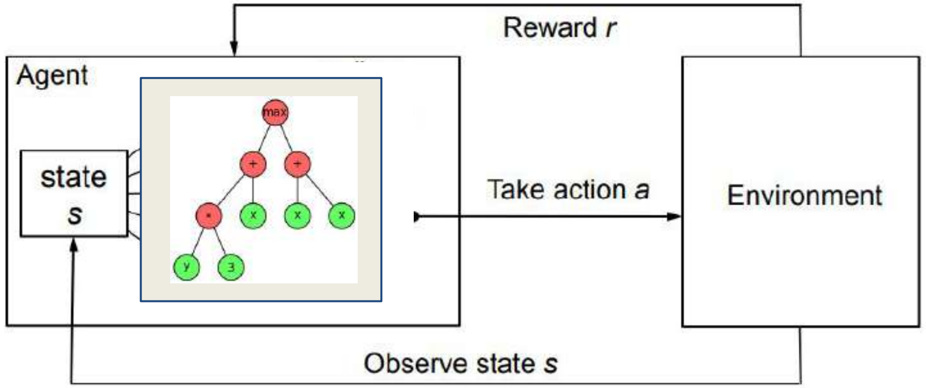
- Provide the reasons for the behavior of a given machine learning model[2]
- Any action taken with the intent of providing an explanation of a model to a human observer

[1]Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: arXiv:1702.08608 (2017).
[2]Alejandro Barredo Arrieta et al. "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI". In: Information Fusion 58 (2020), pp. 82–115

Deep reinforcement learning is behind various breakthroughs in reinforcement learning (RL)

Traditional reinforcement learning approach

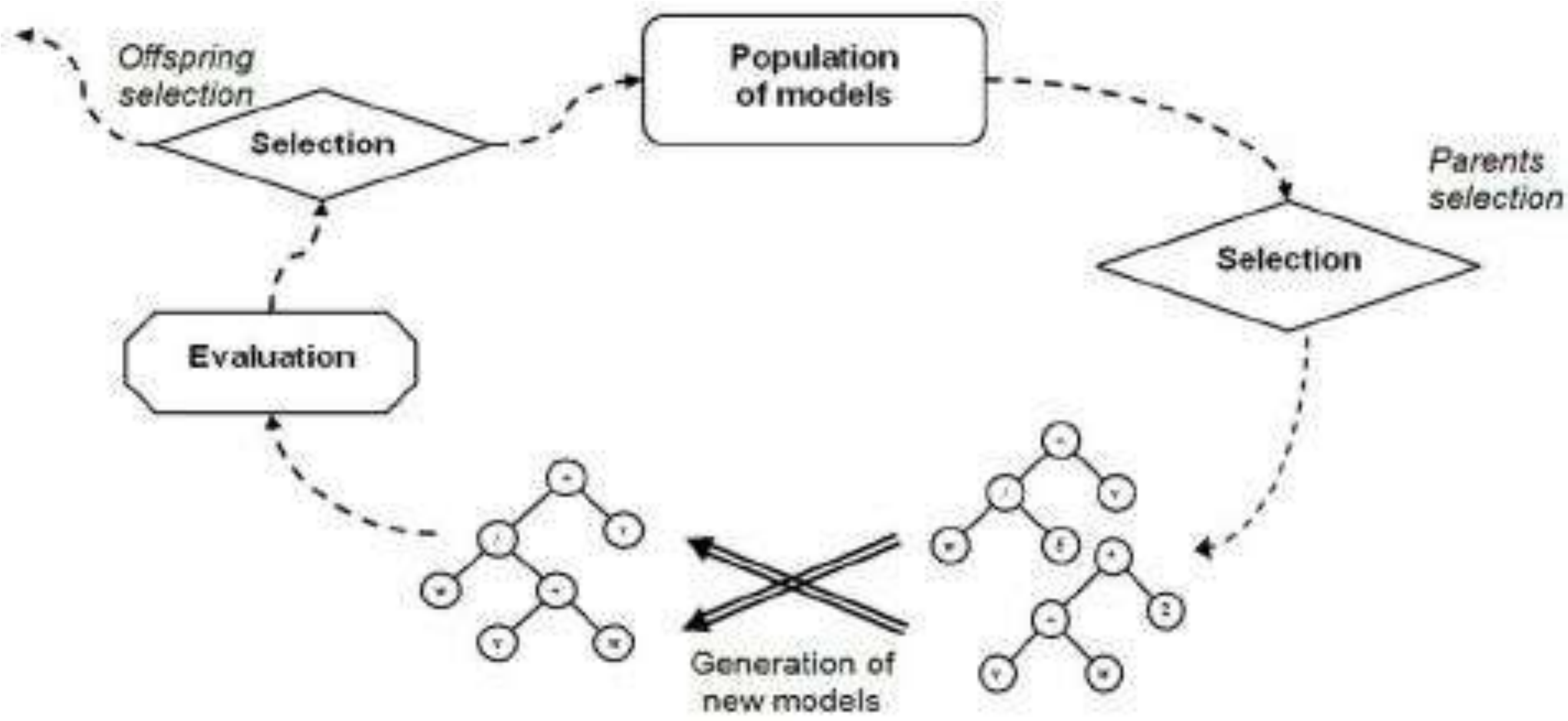




Classifical deep reinforcement learning (DRL)

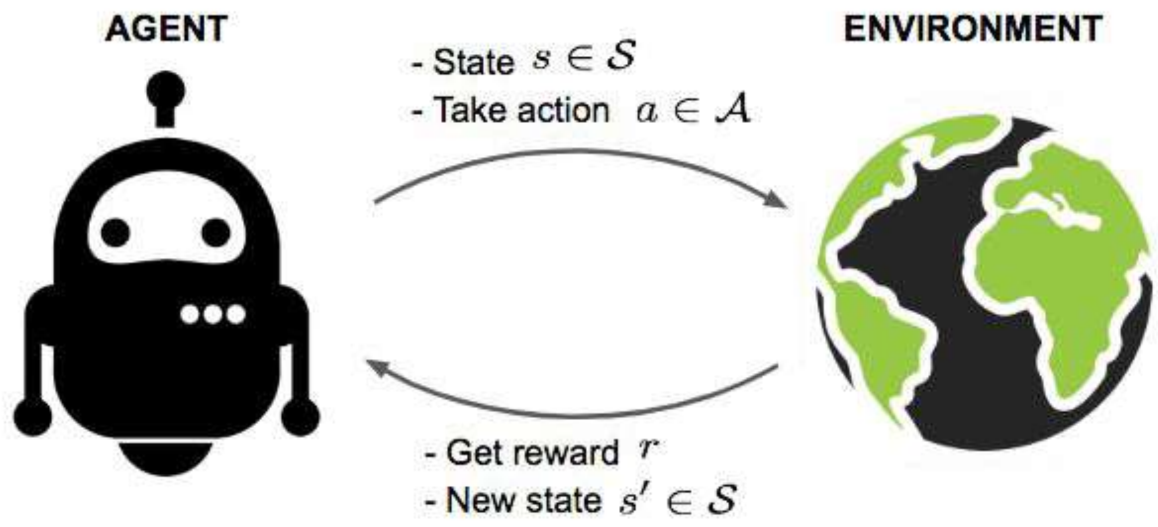
Goal: find policy π: *s*→*a* that maximizes a cumulative reward

Genetic programming (GP) for program synthesis



Using genetic programming (GP) for direct policy search

Operator	Complexity
Variables	1
Constants	1
$+, -, \times$	1
$/$	2
exp, log, ^	3
sin, cos	4
$<, >$	4
if	5



Fitness =

Mean cumulative rewards
Program's complexity

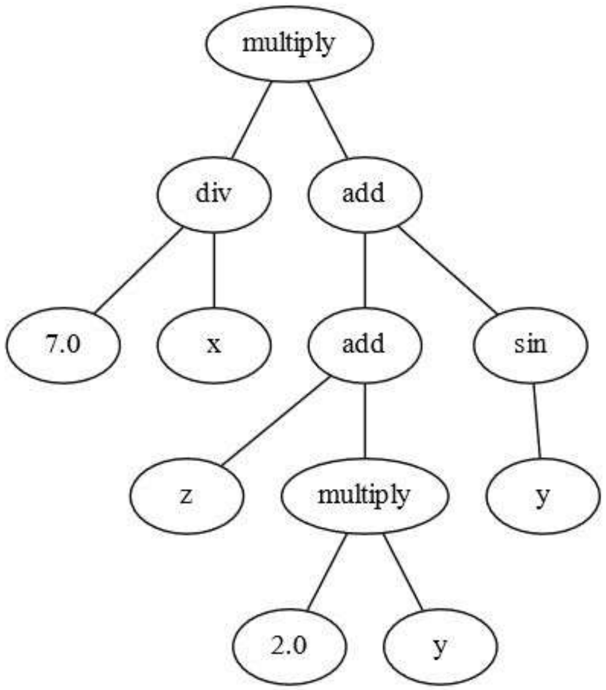
Evaluation strategy:

1. Gradual augmentation of the simulation budget
2. Distribution of the simulation budget toward the most promising individuals

Using Tree and Linear GP to represent the programs

Expression $\frac{7}{x} * (z + 2y + \sin(y))$

Tree GP



➢ **Crossover**: sub-trees' exchanging
➢ **Mutation**: randomly change a node value

Linear GP

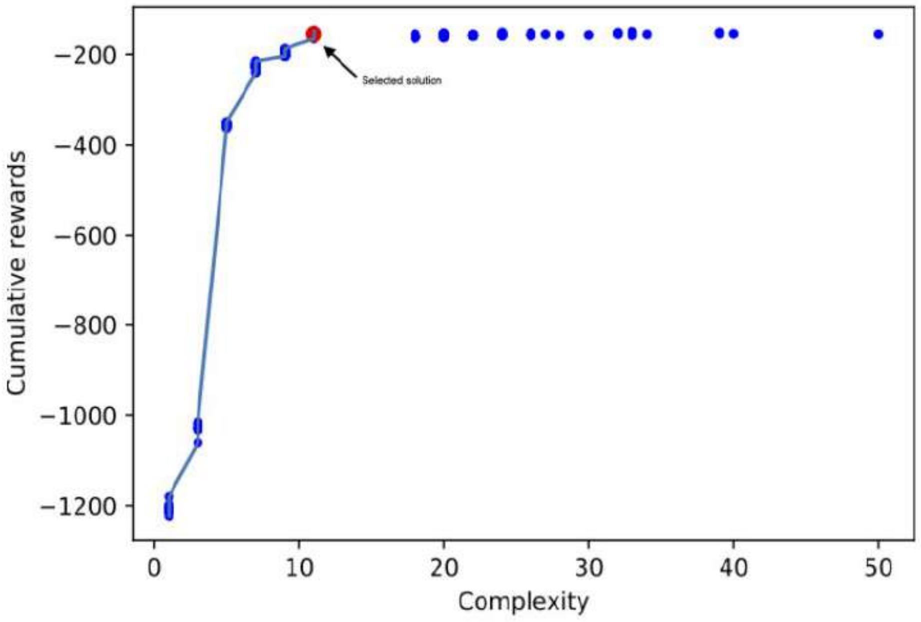
```
\\initialisation du registre
float registre[] = [
    0.0, 0.0, 0.0, \\registre de calcule
    x, y, z,       \\registre pour les entrées
    2.0, 7.0 ]     \\registre pour les constantes

void programme(r){
    r[0] = r[6] * r[4]\\ 2 * y
    r[0] = r[5] + r[0]\\ z + (2y) (1)

    r[1] = sin( r[4] )\\ sin(y) (2)
    r[0] = r[0] + r[1]\\  (1) + (2) (3)

    r[2] = r[7] / r[3]\\ 7 / x (4)
    r[0] = r[0] + r[2]\\ (3) + (4)
}
```

➢ **Crossover:** exchange blocks of instructions
➢ **Mutation:** add/remove instructions

Going towards explainable RL policies

Multi-objective optimization

- **NSGA-II** - multi-objective optimization algorithm
  - Cumulated Reward
  - Complexity (#operations)
- Return a knee point of the Pareto frontier, favoring performance
- Used with the tree-GP representation



Variation Operators

- Higher probability to remove than to add instructions
- used with linear-GP representation

Classical control tasks

Acrobot-v1
Swing up a two-link robot.

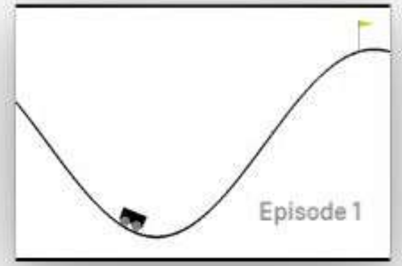
(6, 3)

CartPole-v1
Balance a pole on a cart.

(4, 2)

Discrete action space

Continuous action space



MountainCarContinuous-v0
Drive up a big hill with continuous control.

(2, 1)

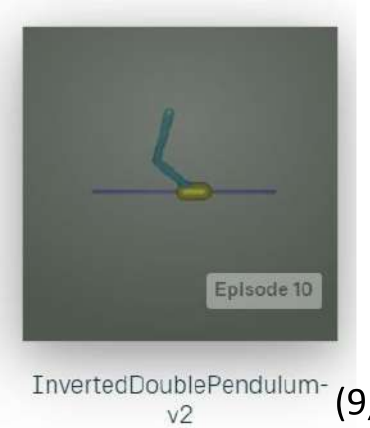
Pendulum-v0
Swing up a pendulum.

(3, 1)

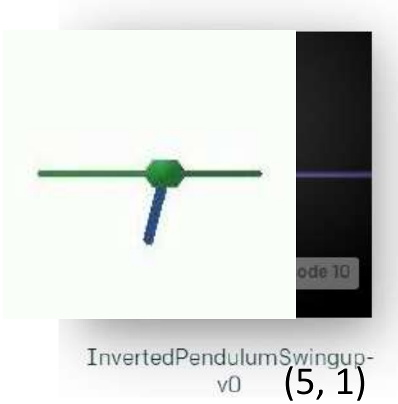
(# features/inputs, # outputs)

Complex locomotion tasks

Mujoco[1]



(9, 1)

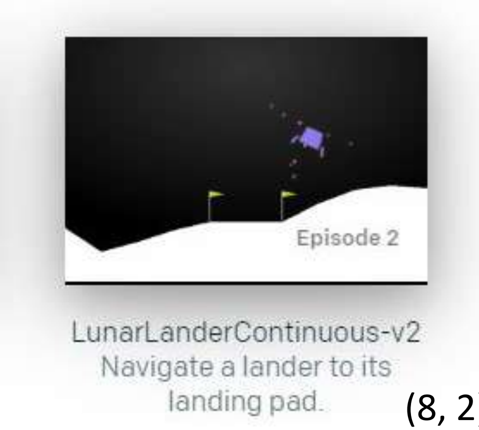


(5, 1)

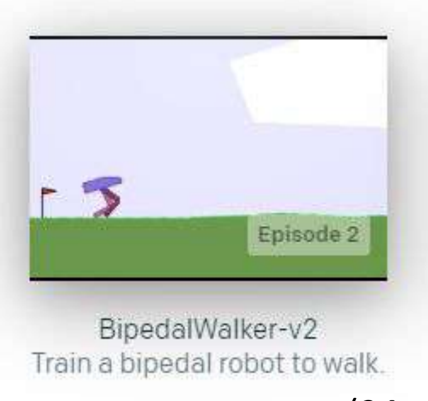
(15, 3)

(28, 8)

Box2D[2]



(8, 2)



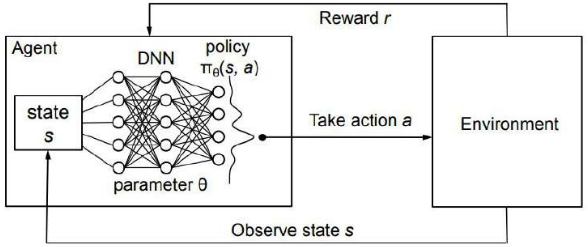
(24, 4)

(24, 4)

[1]"Bullet physics engine." [Online]. Available: https://github.com/benelot/pybullet-gym
[2]Brockman G. et al. (2016). OpenAI gym. *arxiv:1606.01540*.: https://gym.openai.com/envs/#box2d

Comparing the performance of the GP-based policy with the one of neural networks and direct policy search (DPS)

Deep Reinforcement Learning



- **Policy**: a deep neural network. **Complexity** = #weights
  - Use the reward at every time step to update the weights
- **DNN**: Best of PPO[1], SAC[2] and A2C[3], state-of-the-art algorithms
  Proximal Policy Optimization, Soft Actor Critic, and Advantage Actor Critic

Direct Policy Search

- **Policy**: a neural network. **Complexity** = #weights tuned by some HPO algorithm
  - Global optimization of the weights so as to maximize the cumulative reward
- **DPS**: Best of (almost) all global optimization algorithms in ***Nevergrad***[4]

[1]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. Proximal Policy Optimization algorithms. arXiv:1707.06347 (2017)
[2]Haarnoja, T., Zhou, A., Abbeel, P., Levine, S. Soft Actor-Critic: Off-policy maximum entropy DeepRL with a stochastic actor. Proc. ICML, pp. 1861–1870 (2018)
[3]Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for DeepRL. Proc. ICML pp. 1928–1937 (2016)
[4]Rapin, J., Teytaud, O. Nevergrad - A gradient-free optimization platform. https://github.com/FacebookResearch/Nevergrad (2018)

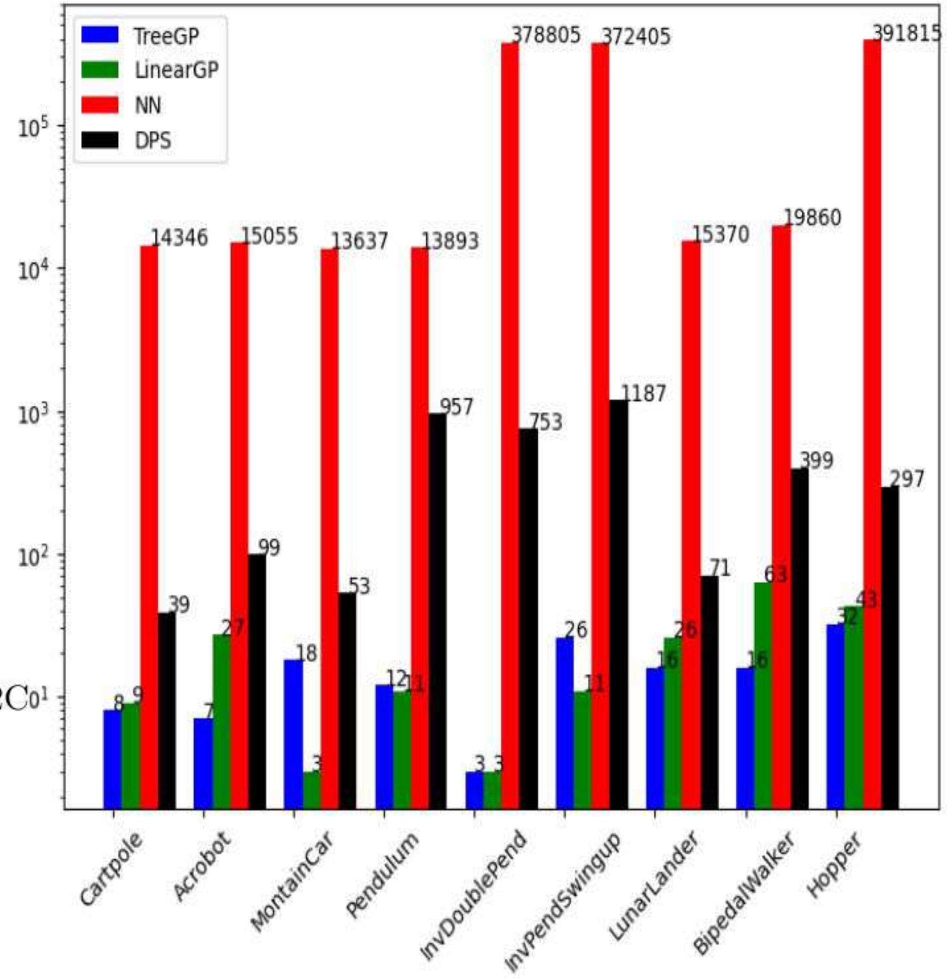
GP-based policy can outperform neural networks in different tasks

Cumulated reward (higher is better)

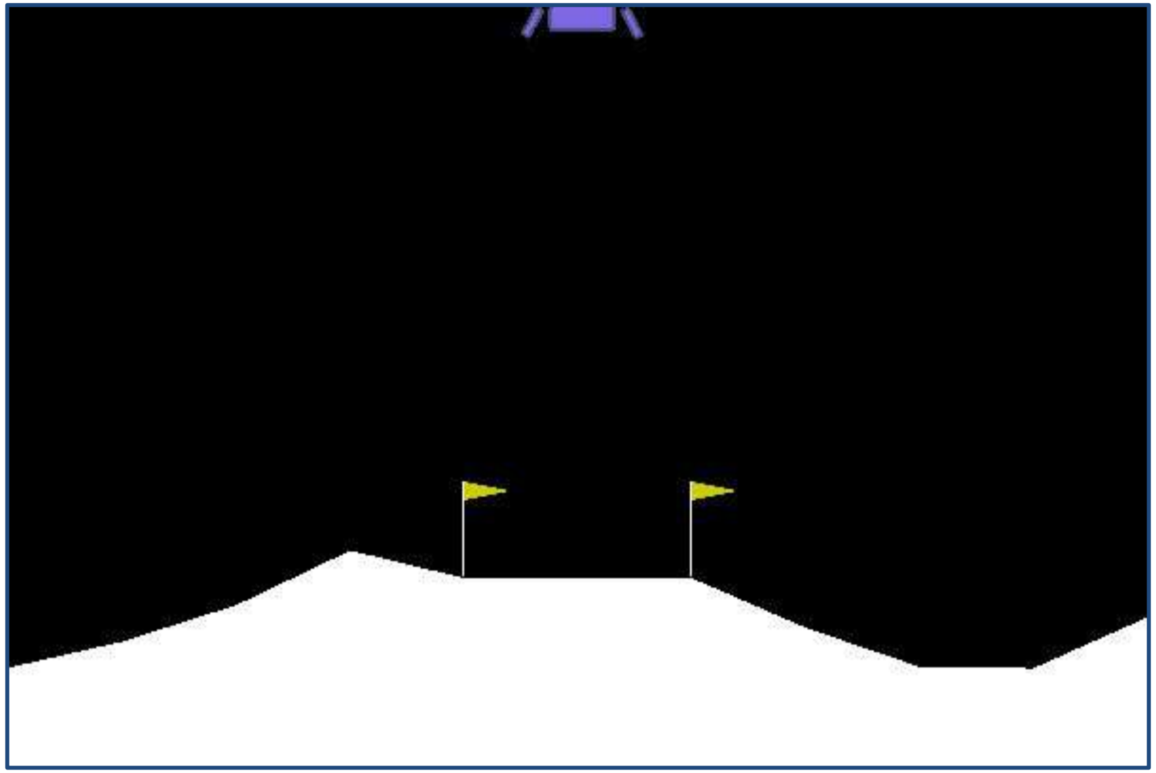
Environment	# in	# out	Tree GP	Linear GP	DPS	DNN
Control tasks						
Cartpole	4	2	500.0	500.0	500.0	500.0†
Acrobot	6	3	-83.17	-80.99	**-72.74**	−82.98†
MountainCarC0	2	1	99.31	88.16	**99.4**	94.56‡
Pendulum	3	1	-154.36	-164.66	**-141.9**	-154.69‡
Mujoco						
InvDoublePend	9	1	9092.17	9089.50	**9360**	9304.32‡
InvPendSwingUp	5	1	**893.35**	887.08	893.3	891.45‡
Hopper	15	3	999.19	949.27	2094	**2604.91**‡
Box2D						
LunarLanderC0	8	2	**287.58**	262.42	282.†PPO ‡SAC *A2C	
BipedalWalker	24	4	268.85	257.22	**310.1**	299.44*
BipedalWalkerHardcore	24	4	9.25	10.63	8.16	**246.79***

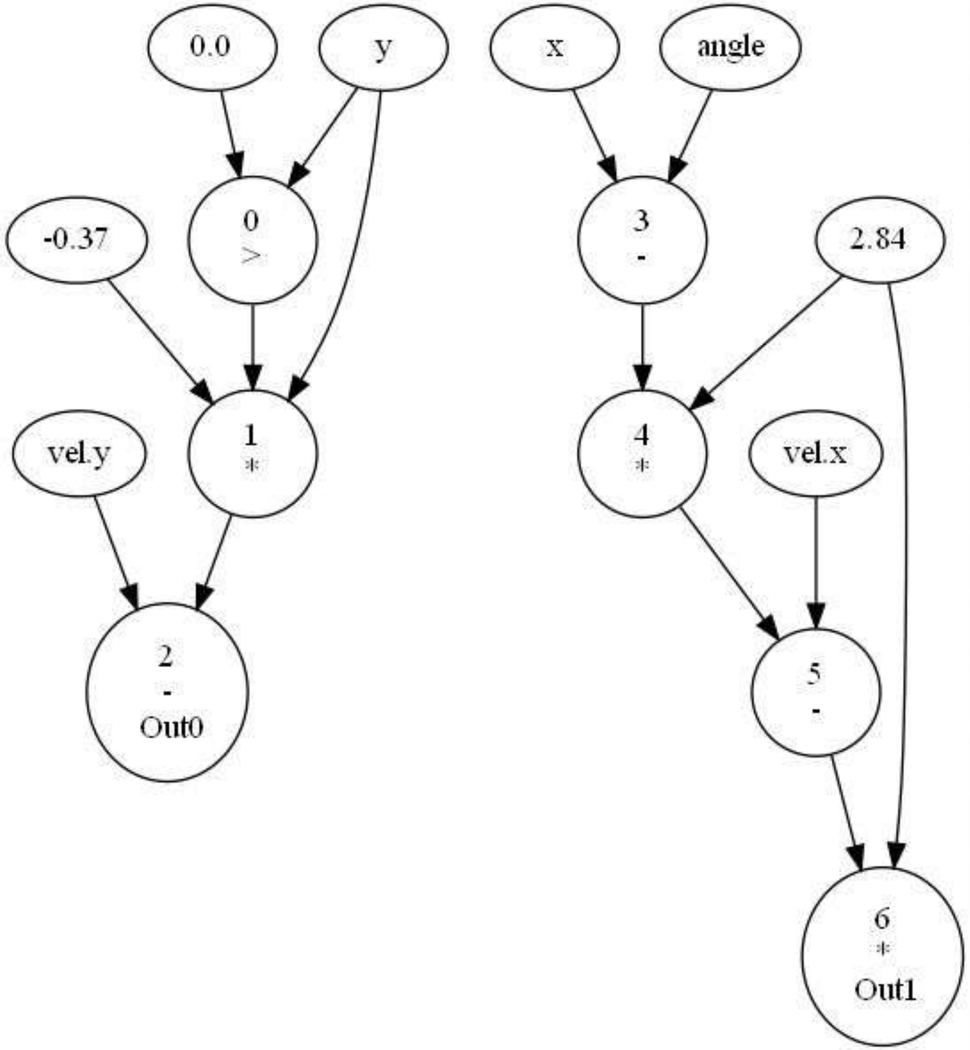
DPS: direct policy search (Nevergrad)

Complexity (lower is better)



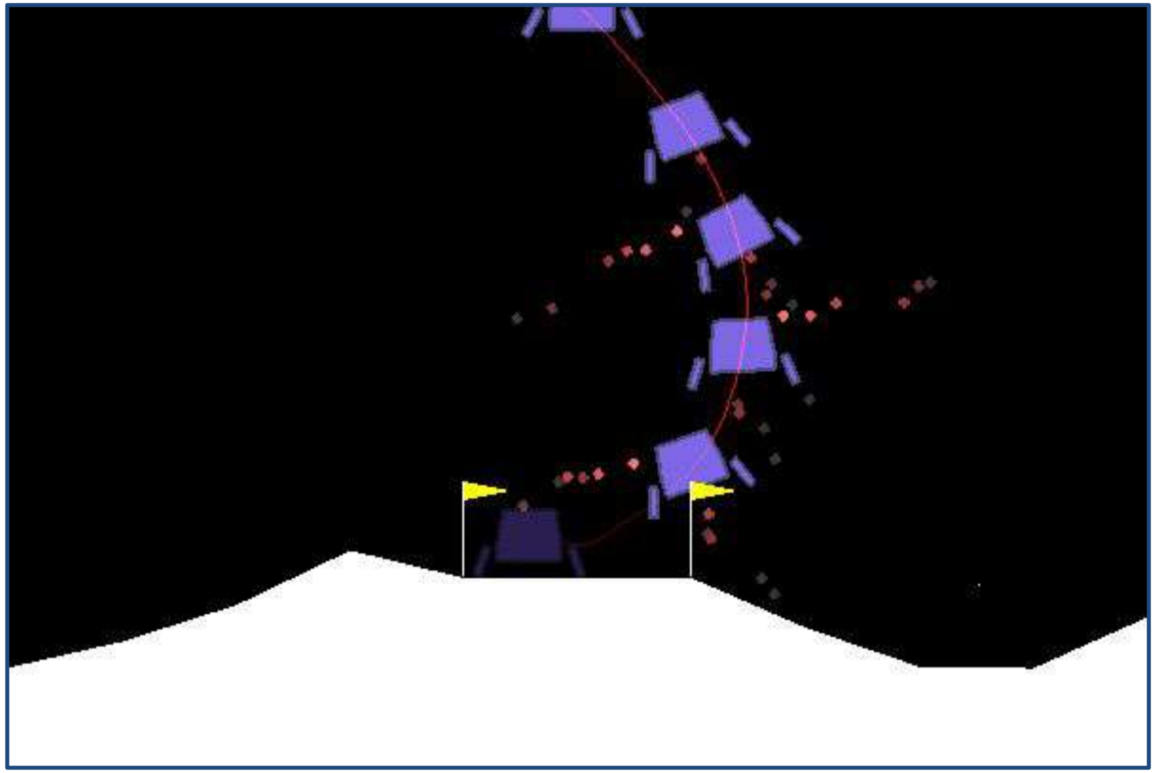
GP-based strategy leads to a simple Lunar Lander policy

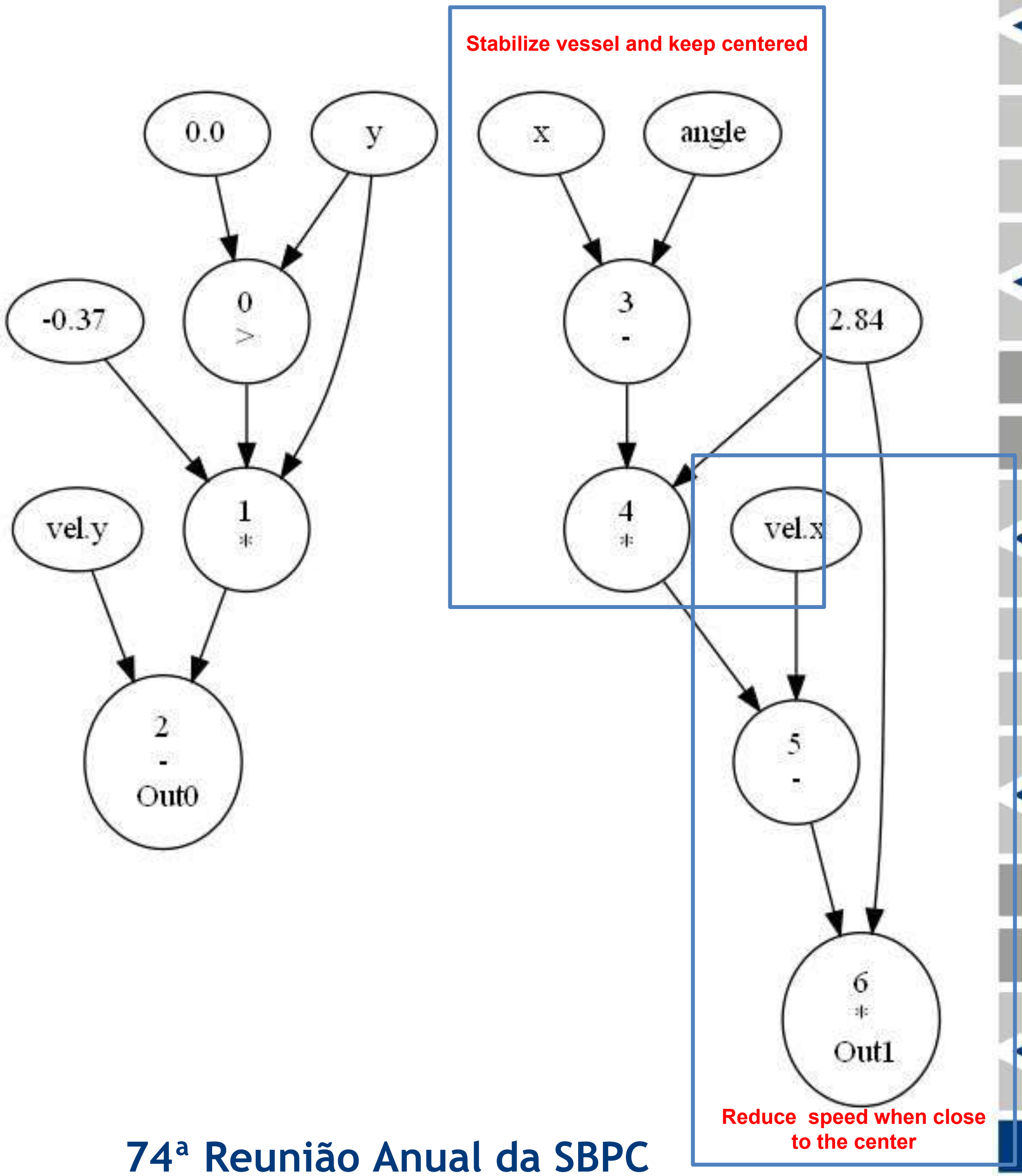




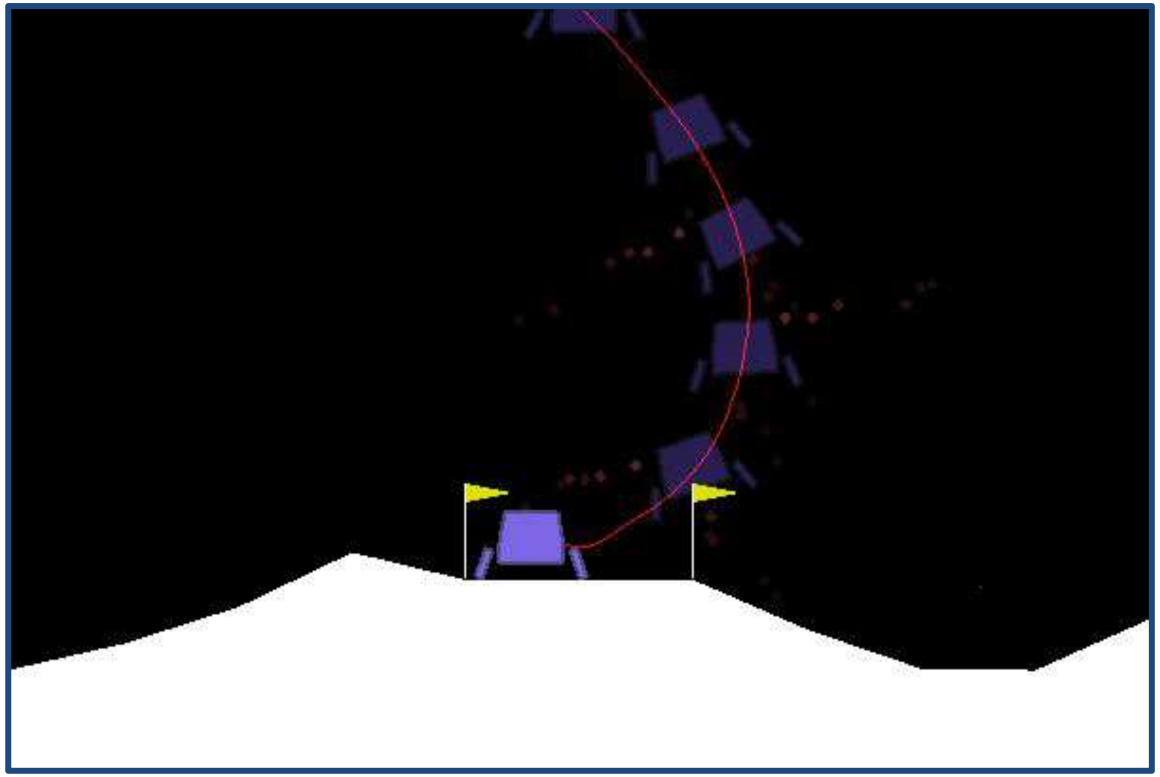
Out0: Central engine
Out1: Side engines

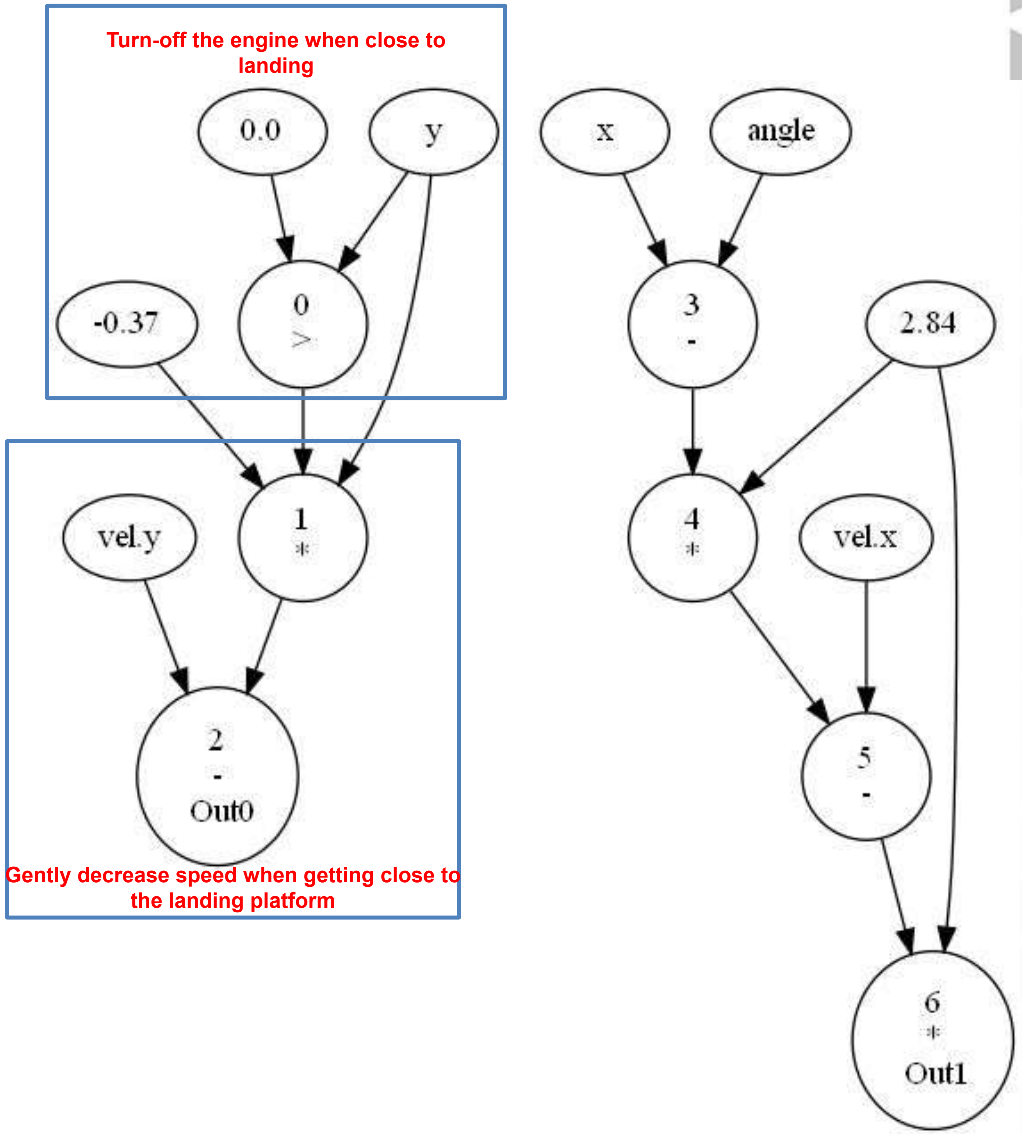
One can understand the Lunar Lander policy





One can understand the Lunar Lander policy



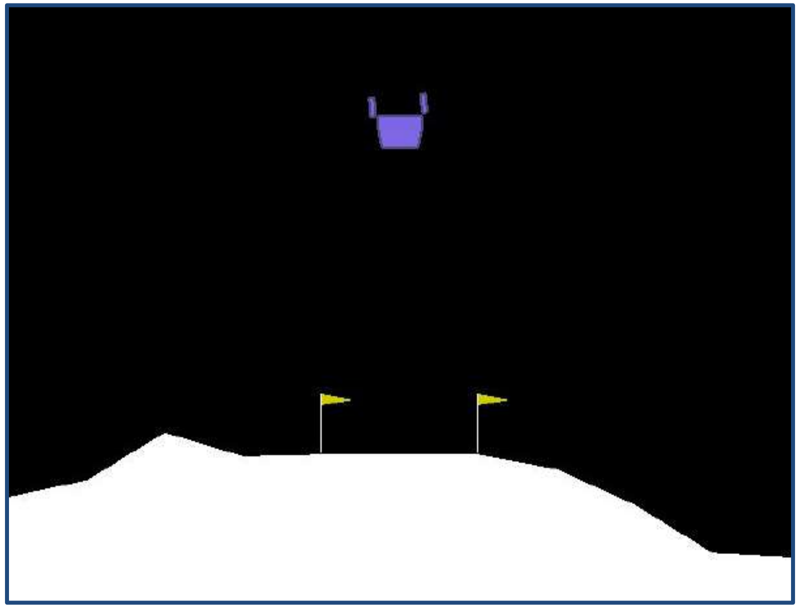


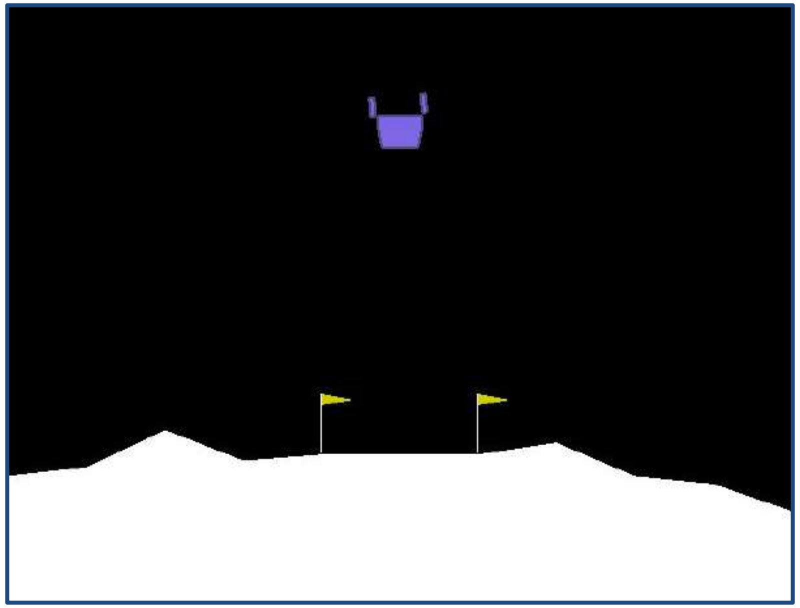
We can also observe cases in which the policy fails

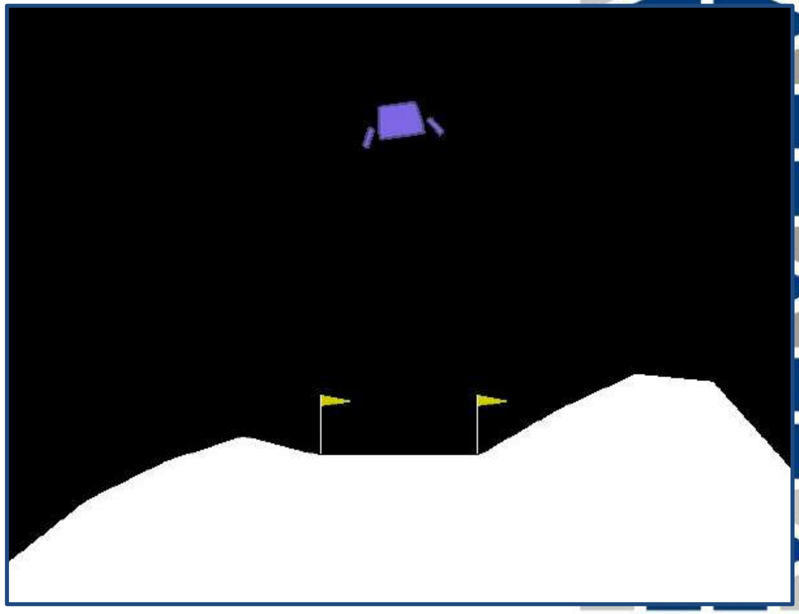
Upside-down

… with side engines off

Too large angular speed







Accelerate when getting closer to the landing zone

Does not use angular velocity

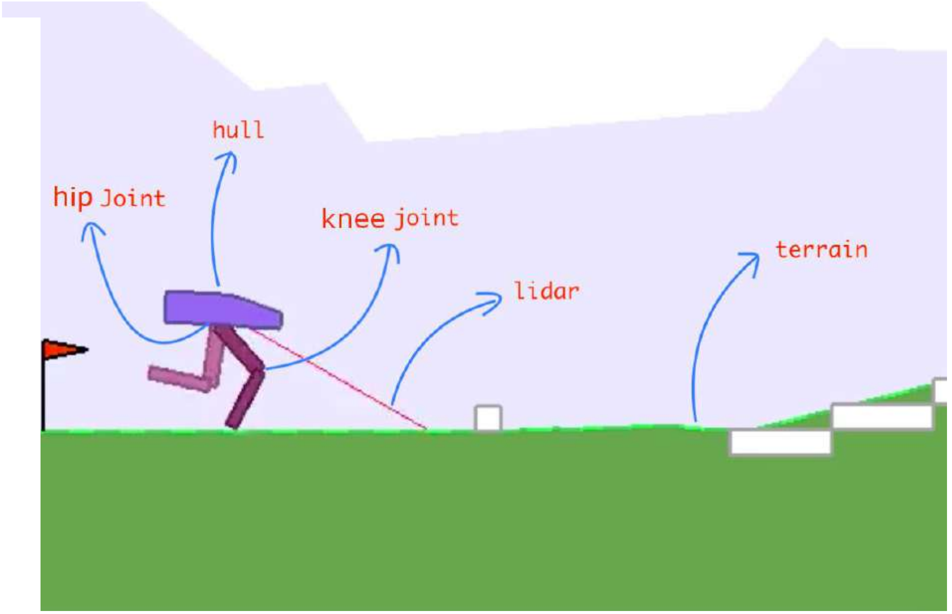
One can have an interpretable Bipedal walker policy

$$\text{hip1} : a_0 = \text{knee1}.\theta$$

$$\text{knee1} : a_1 = \frac{\text{knee1}.\dot{\theta}}{\text{lidar}_5}$$

$$\text{hip2} : a_2 = \frac{\text{lidar}_7}{\text{knee1}.\dot{\theta}} - \text{hip2}.\theta + \text{hull}.\theta$$

$$\text{knee2} : a_3 = \text{hull}.\theta - \text{knee2}.\theta$$



For complex tasks, it can be blocked on a local optimum

- On the **interpretable** side:
  - Less **intuitive/understandable** policies

- On the **performance** side:
  - **Local optimum** issue
  - **Bad exploration**

Now, we can focus on escaping the local optimum

Escaping the local optimum through quality diversity

💡 Maintaining population's diversity

- ❑ Each individual is associated with behavioral descriptor (e.g., mean amplitude of the joints)
- ❑ Starting with Map-Elites algorithm[1] then fine tuning with base algorithm to reduce policy complexity.

❑ Results on two locomotion environments:

- ✚ Better exploration
- ✚ Improved performance

Environment	QD-Tree GP	QD-Linear GP	Tree GP	Linear GP	NN
BipedalWalker	311.34	299.64	268.85	257.22	299.44*
Hopper	2152.19	1450.11	999.19	949.27	2604.91‡

* *A2C*
‡ *SAC*

[1]Mouret, J., & Clune, J. (2015). Illuminating search spaces by mapping elites. *arxiv:1504.04909*.
[1]Flageat, M., & Cully, A. (2020). Fast and stable MAP-Elites in noisy domains using deep grids. In *Artificial Life Conference*, 273-282)

Escaping the local optimum through quality diversity

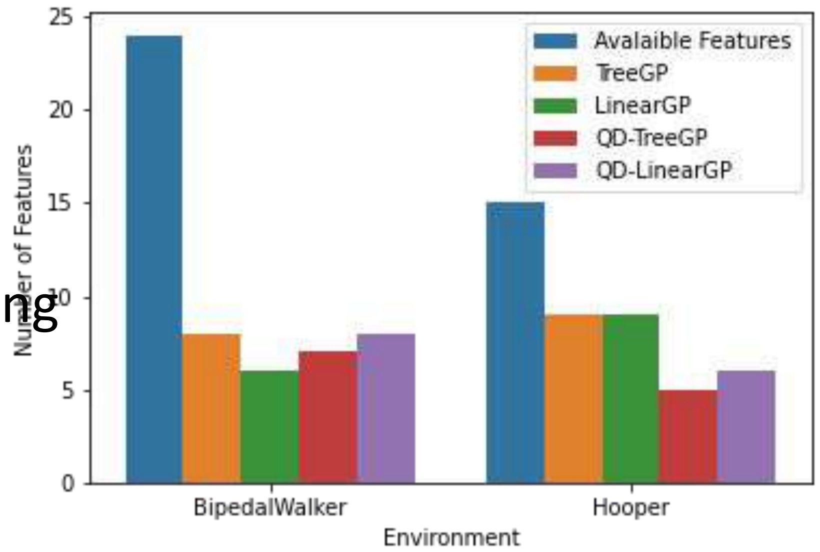
Maintain diversity of the population:

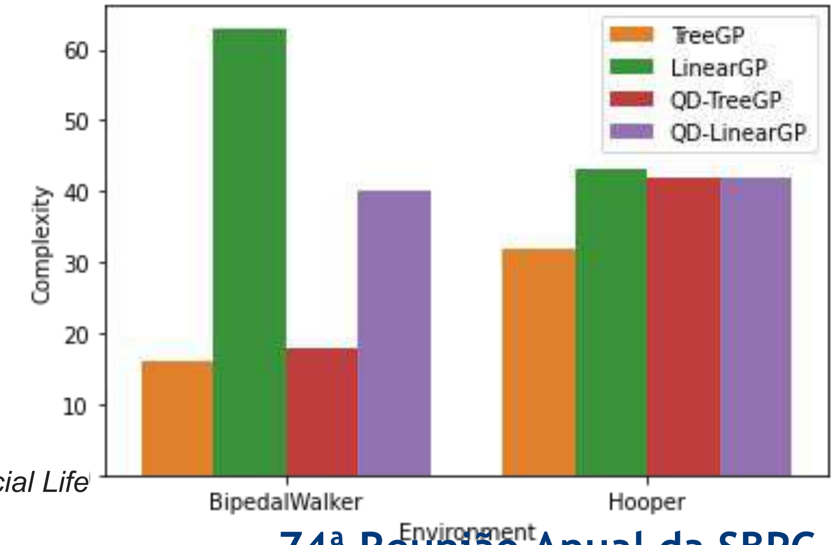
- ❑ Each individual is associated with behavioral descriptor (here mean amplitude of the joints)
- ❑ Starting with Map-Elites algorithm[1] then fine tuning with base algorithm to reduce policy complexity.

❑ Results on two locomotion environments:

- \+ Better exploration
- \+ Improved performance

- \+ No significant increase in complexity
- − Very dependent on the grid (descriptor and size)





[1] Mouret, J., & Clune, J. (2015). Illuminating search spaces by mapping elites. *arxiv:1504.04909*.
[1] Flageat, M., & Cully, A. (2020). Fast and stable MAP-Elites in noisy domains using deep grids. In *Artificial Life Conference*, 273-282)

Quality diversity leads to a better exploration strategy

Linear GP classic
(baseline)

QD - 1 descriptor
(foot joint)

QD - 2 descriptors
(leg + foot joint)

Evolutionary approaches can provide interpretable RL policies with performance comparable to deep reinforcement learning

GP-based reinforcement learning (RL) policies:

- \+ are **competitive** against neural networks on various tasks
- \+ are **portable**
- \+ can be framed as a **multi-objective optimization problem**
- \+ have **low resources** footprint
- \+ use only **episode cumulative reward**
- \- can stuck in **local optimum**
- \+ are **interpretable** (i.e., concise)
- \- can be non-**intuitive in** some **locomotion tasks**


“Thats
olks!”

74ª REUNIÃO ANUAL DA SBPC

UnB

Ciência, independência e soberania nacional

24 a 30 de julho de 2022

Realização

Sociedade Brasileira para o Progresso da Ciência

UnB

Atuante como sempre, necessária como nunca

Sócios institucionais da SBPC

Instituições parceiras

Apoio