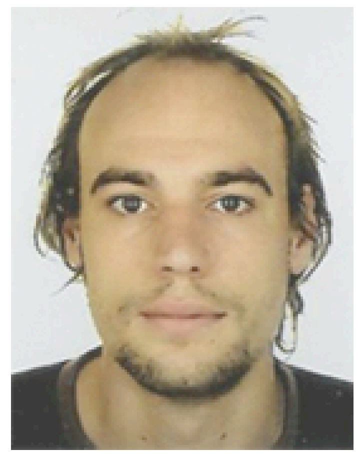
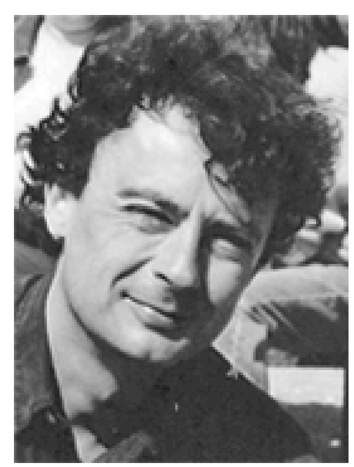
POLICY SEARCH IN A SPACE OF SIMPLE CLOSED-FORM FORMULAS: TOWARDS INTERPRETABILITY OF REINFORCEMENT LEARNING

FRANCIS MAES

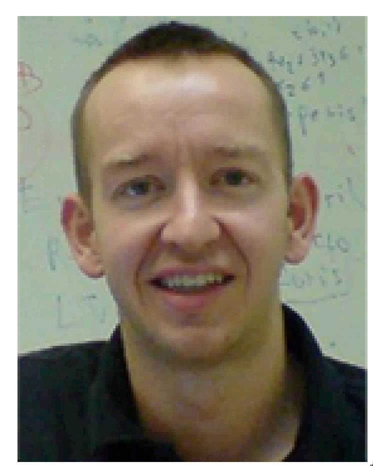


RAPHAEL FONTENEAU

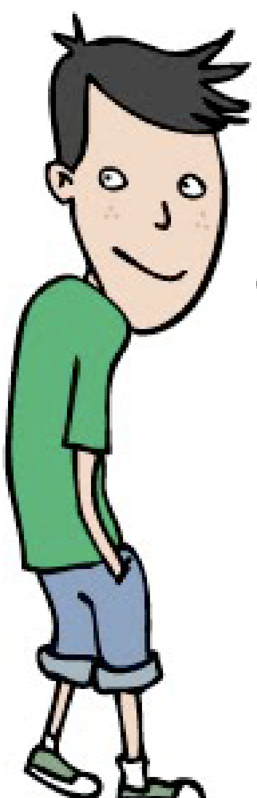
LOUIS WEHENKEL

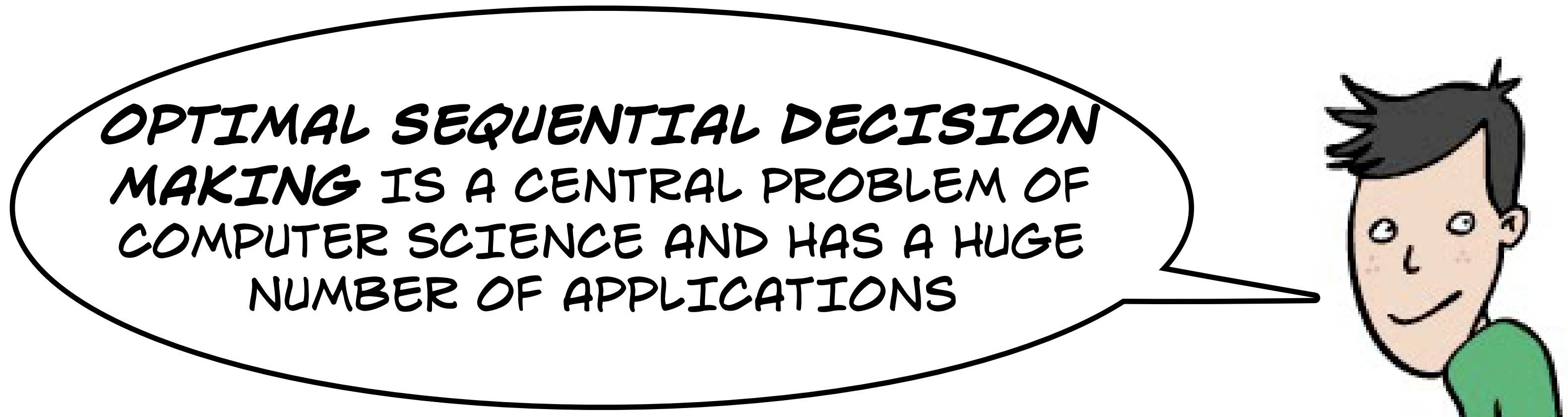


DAMIEN ERNST



Université de Liège

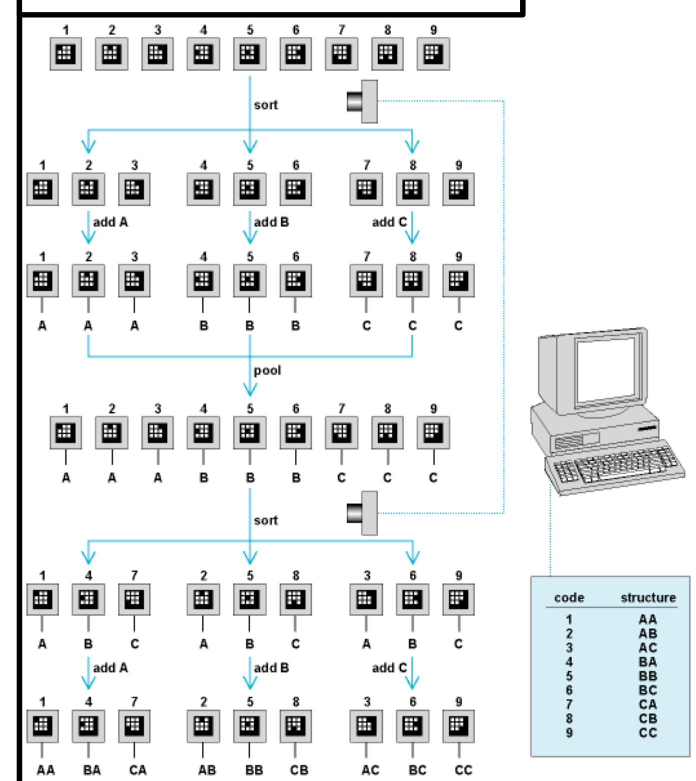




MEDICAL THERAPY OPTIMIZATION

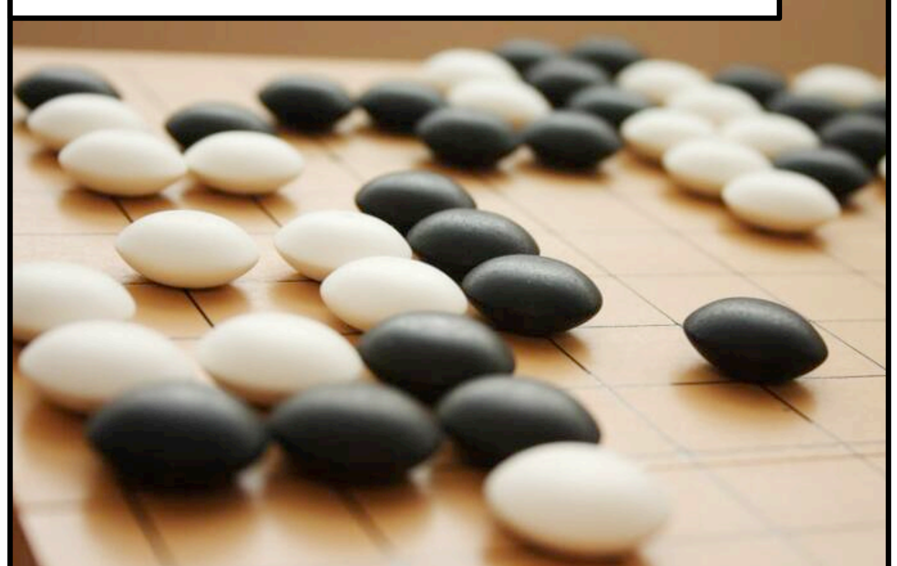
FINANCIAL TRADING

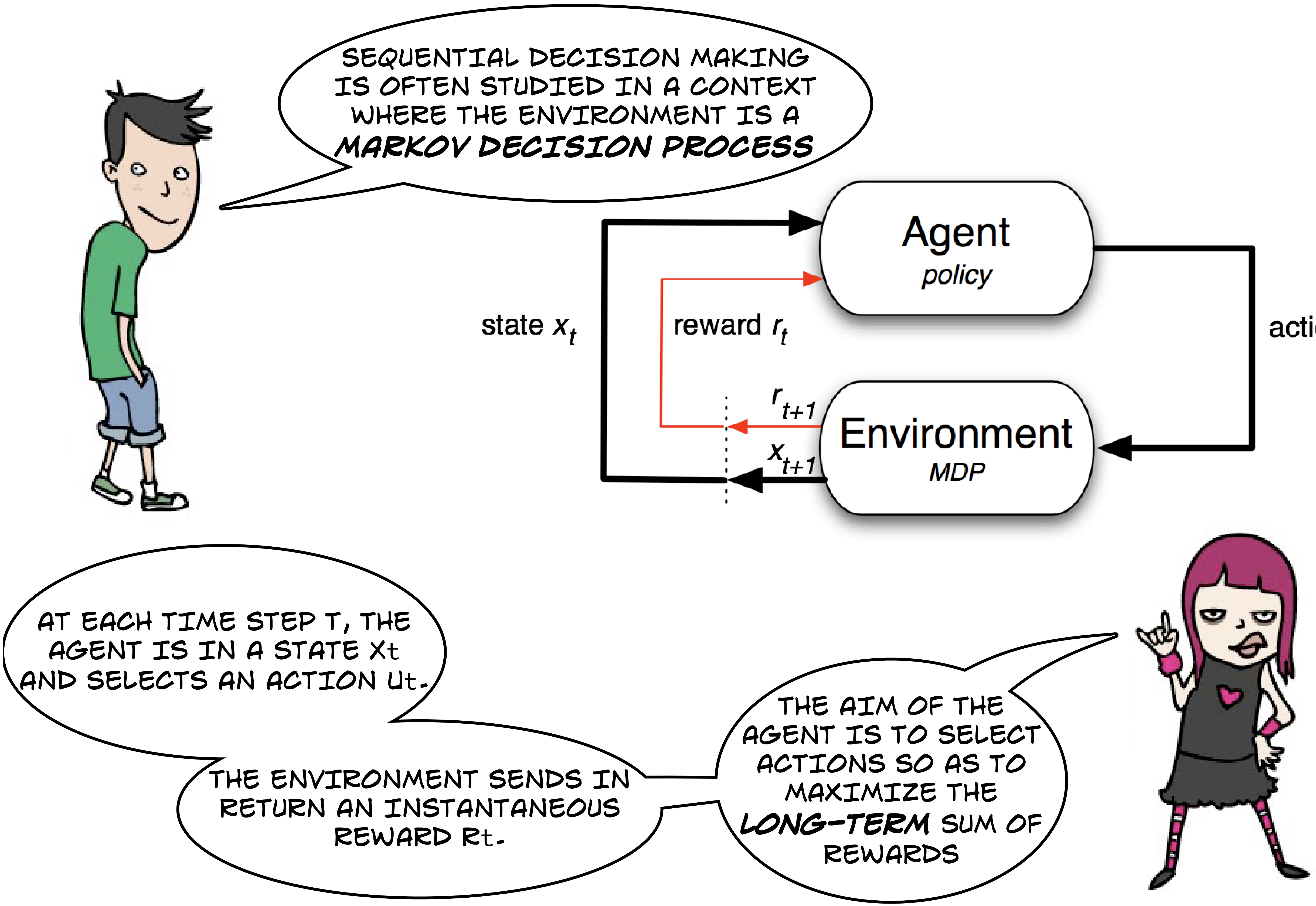
COMBINATORIAL SEARCH

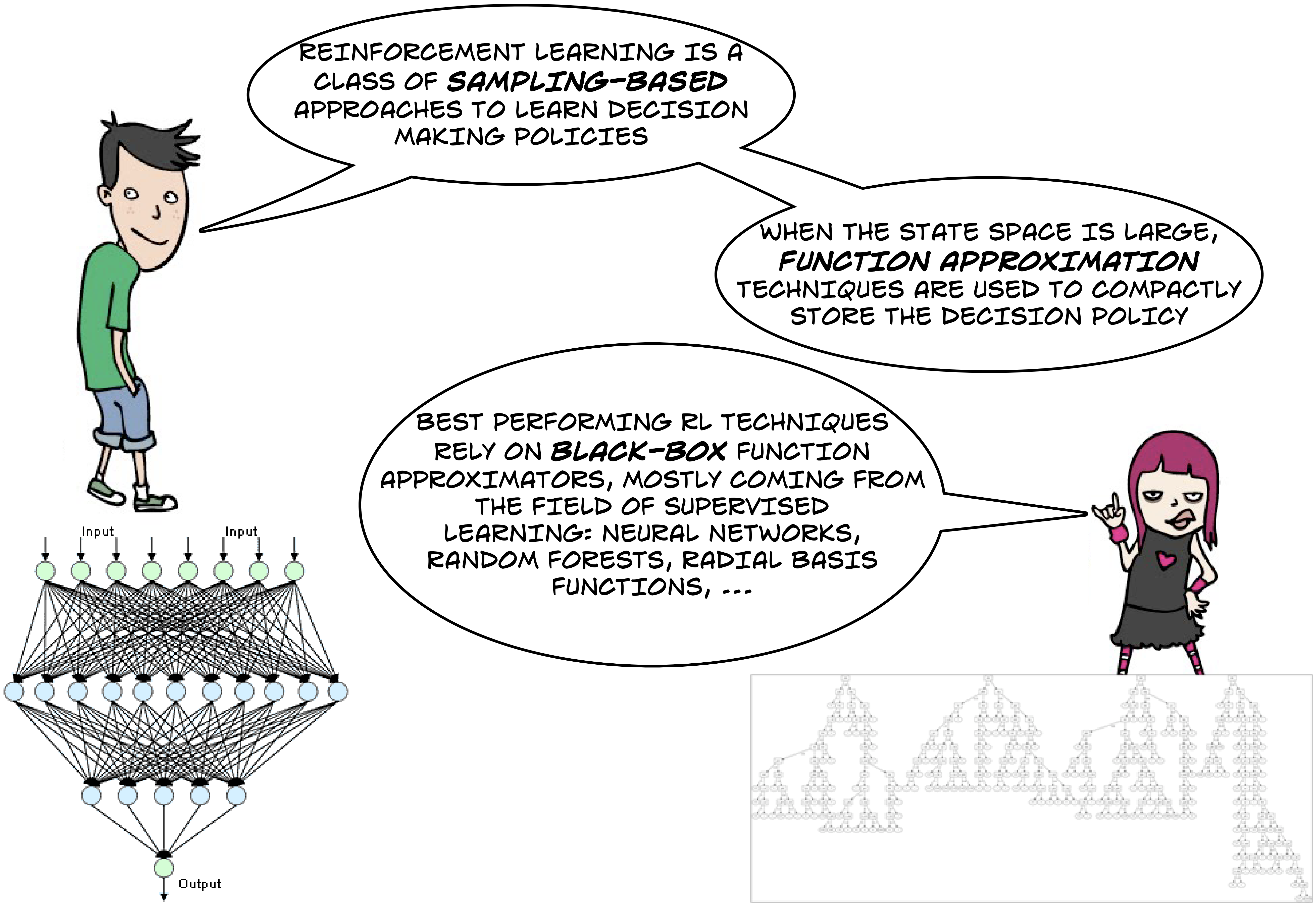


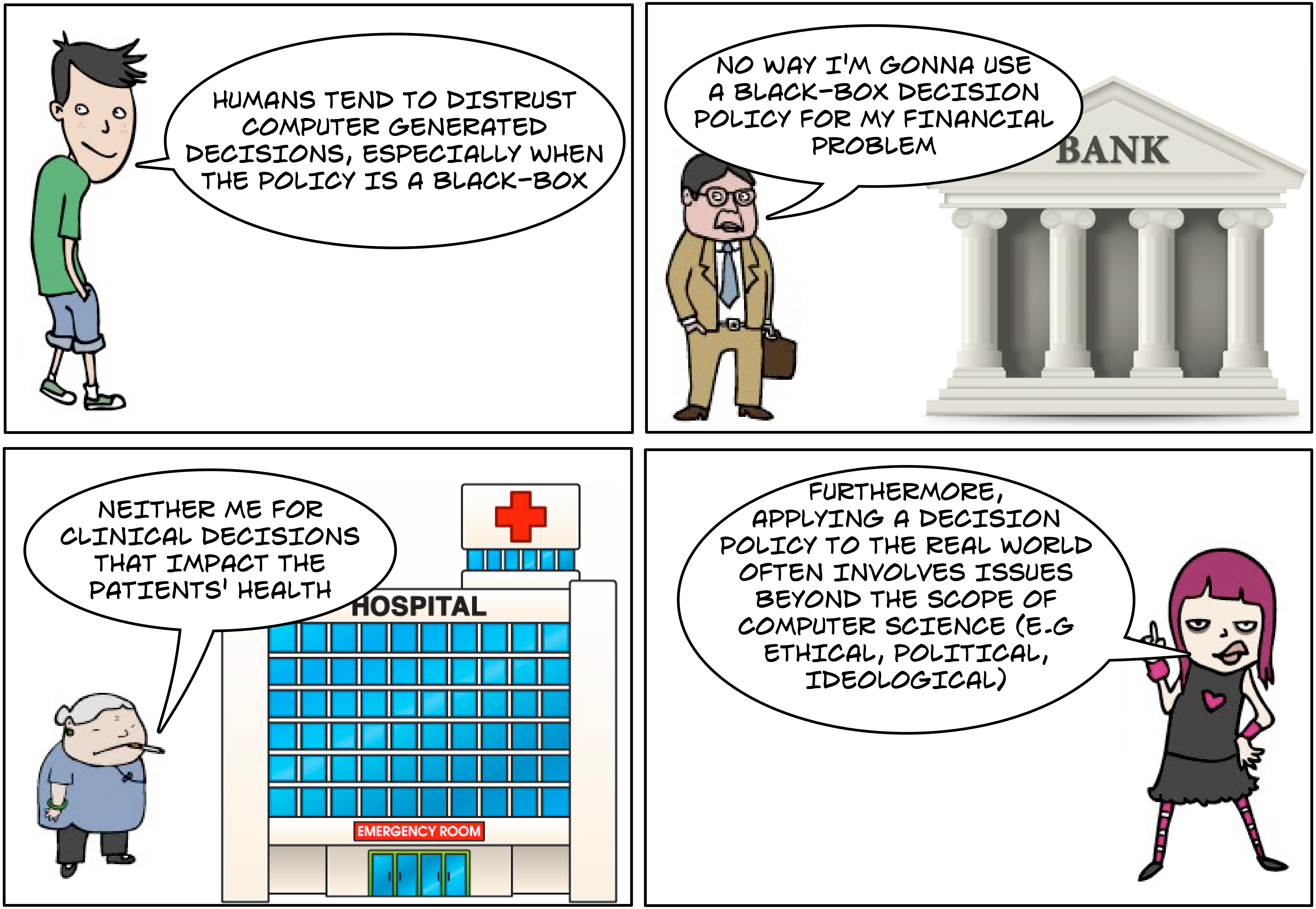
ROBOT OPTIMAL CONTROL

AUTOMATIC GAME PLAYING



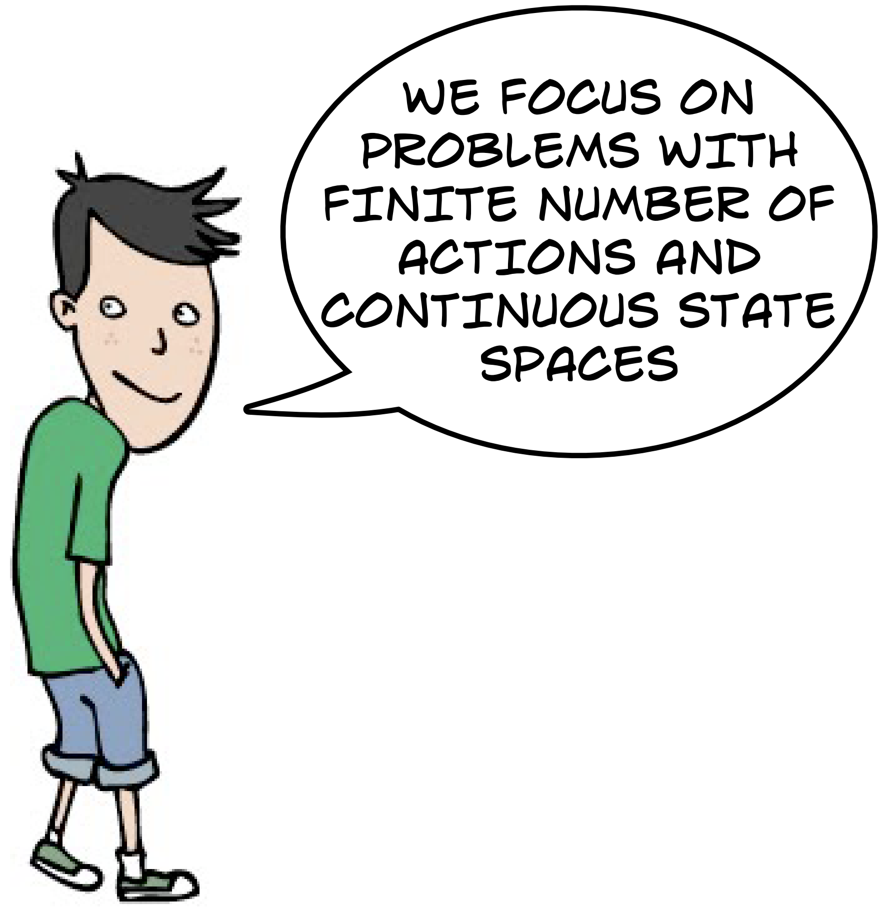

SEQUENTIAL DECISION MAKING IS OFTEN STUDIED IN A CONTEXT WHERE THE ENVIRONMENT IS A ***MARKOV DECISION PROCESS***
Agent
policy
state x_t
reward r_t
action u_t
r_{t+1}
x_{t+1}
Environment
MDP
AT EACH TIME STEP T, THE AGENT IS IN A STATE Xt AND SELECTS AN ACTION Ut.
THE ENVIRONMENT SENDS IN RETURN AN INSTANTANEOUS REWARD Rt.
THE AIM OF THE AGENT IS TO SELECT ACTIONS SO AS TO MAXIMIZE THE ***LONG-TERM*** SUM OF REWARDS


REINFORCEMENT LEARNING IS A CLASS OF **SAMPLING-BASED** APPROACHES TO LEARN DECISION MAKING POLICIES
WHEN THE STATE SPACE IS LARGE, **FUNCTION APPROXIMATION** TECHNIQUES ARE USED TO COMPACTLY STORE THE DECISION POLICY
BEST PERFORMING RL TECHNIQUES RELY ON **BLACK-BOX** FUNCTION APPROXIMATORS, MOSTLY COMING FROM THE FIELD OF SUPERVISED LEARNING: NEURAL NETWORKS, RANDOM FORESTS, RADIAL BASIS FUNCTIONS, ...
Input
Input
Output


HUMANS TEND TO DISTRUST COMPUTER GENERATED DECISIONS, ESPECIALLY WHEN THE POLICY IS A BLACK-BOX
NO WAY I'M GONNA USE A BLACK-BOX DECISION POLICY FOR MY FINANCIAL PROBLEM
BANK
NEITHER ME FOR CLINICAL DECISIONS THAT IMPACT THE PATIENTS' HEALTH
HOSPITAL
EMERGENCY ROOM
FURTHERMORE, APPLYING A DECISION POLICY TO THE REAL WORLD OFTEN INVOLVES ISSUES BEYOND THE SCOPE OF COMPUTER SCIENCE (E.G ETHICAL, POLITICAL, IDEOLOGICAL)


THERE ARE MANY EFFICIENT RL ALGORITHMS, SOME OF THEM WITH STRONG THEORETICAL GUARANTEES, BUT THESE ALGORITHMS DO NOT LEAVE THE LABORATORIES
IF WE WANT TO CHANGE THIS, WE MUST PROVIDE DECISION POLICIES THAT HUMANS CAN UNDERSTAND AND EVENTUALLY TRUST
WE THUS NEED RL ALGORITHMS PRODUCING ***INTERPRETABLE POLICIES***
INTERPRETABILITY IS AN OLD TOPIC IN SUPERVISED LEARNING, BUT SURPRISINGLY NEARLY ABSENT FROM THE FIELD OF RL

PROPOSED APPROACH



MDP DYNAMICS

$$x_{t+1} \sim p_f(.|x_t, u_t) \quad t = 0, 1, \ldots$$

$$r_t \sim p_\rho(.|x_t, u_t)$$

POLICY

$$\pi(x_t) \sim p_\pi(.|x_t)$$

OBJECTIVE: MAXIMIZE RETURN

$$J^\pi = \mathop{\mathbb{E}}_{p_0(.), p_f(.), p_\rho(.)} [\mathcal{R}^\pi(x_0)]$$

$$\mathcal{R}^\pi(x_0) = \sum_{t=0}^{\infty} \gamma^t r_t$$

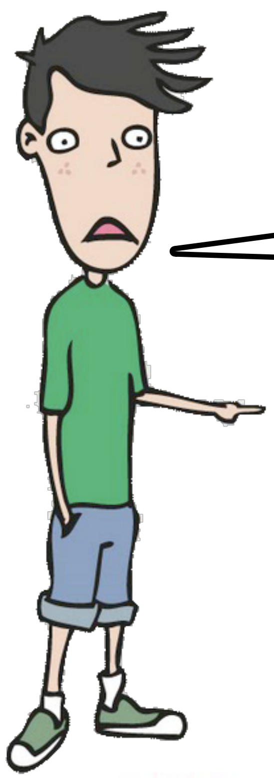
AN INDEX FUNCTION I(.,.) IS A MAPPING:

$$I : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$$

GIVEN AN INDEX FUNCTION, WE CAN DEFINE THE INDEX-BASED POLICY:

$$\forall x \in \mathcal{X}, \pi_I(x) \in \arg\max_{u \in \mathcal{U}} \quad I(x,u)$$

WE FOCUS ON THE CLASS OF INDEX-BASED POLICIES WHOSE INDEX FUNCTIONS ARE DEFINED BY ***SIMPLE CLOSED-FORM FORMULAS***



A FORMULA $F \in \mathbb{F}$ IS:

EITHER A BINARY EXPRESSION $F = B(F', F'')$,
OR AN UNARY EXPRESSION $F = U(F')$,
OR A VARIABLE $F = V$,
OR A CONSTANT $F = C$.

WE USE THE FOLLOWING OPERATORS AND CONSTANTS:

$$\mathbb{B} = \{+, -, \times, \div, \min, \max\}$$
$$\mathbb{U} = \{\sqrt{.}, \ln(.), |.|, -., \tfrac{1}{.}\}$$
$$\mathbb{C} = \{1, 2, 3, 5, 7\}$$

WE CONSIDER TWO DIFFERENT SETTINGS FOR THE VARIABLES:

LOOKAHEAD FREE

$$\mathbb{V} = \mathbb{V}_{LF} = \left\{x_t^{(1)}, \ldots, x_t^{(d_{\mathcal{X}})}, u_t^{(1)}, \ldots, u_t^{(d_{\mathcal{U}})}\right\}$$

ONE-STEP LOOKAHEAD (MODEL ACCESSIBLE)

$$\mathbb{V} = \mathbb{V}_{OL} = \left\{x_t^{(1)}, \ldots, x_t^{(d_{\mathcal{X}})}, u_t^{(1)}, \ldots, u_t^{(d_{\mathcal{U}})}, r_t, x_{t+1}^{(1)}, \ldots, x_{t+1}^{(d_{\mathcal{X}})}\right\}$$

GIVEN A POLICY π, WE DEFINE:

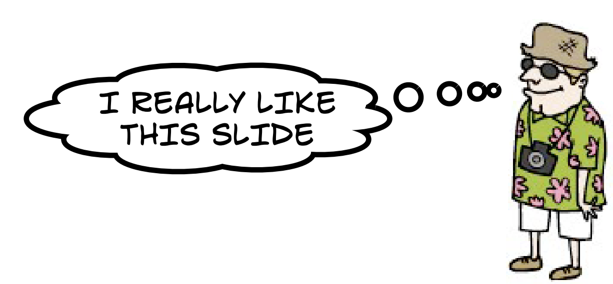
$$D_F(\pi) = \{F \in \mathbb{F} | \pi_F = \pi\}$$

THE KOLMOGOROV COMPLEXITY OF π IS:

$$\kappa(\pi) = \min_{F \in D_F(\pi)} |F|$$

GIVEN K, OUR SPACE OF INTERPRETABLE POLICIES IS DEFINED BY:

$$\Pi_{int}^K = \{\pi | D_F(\pi) \neq \emptyset \text{ and } \kappa(\pi) \leq K\}$$



CONSTRUCTING Π^K_{int} IS NON TRIVIAL (EXCEPT FOR FINITE STATE SPACES)

WE INSTEAD APPROXIMATE THIS SPACE BY COMPARING POLICIES ON A FINITE SET OF SAMPLES

1. WE ENUMERATE ALL FORMULAS $|F| \leq K$

2. GIVEN A FINITE SET OF STATE POINTS $\mathcal{S} = \{s_i\}_{i=1}^{S}$,

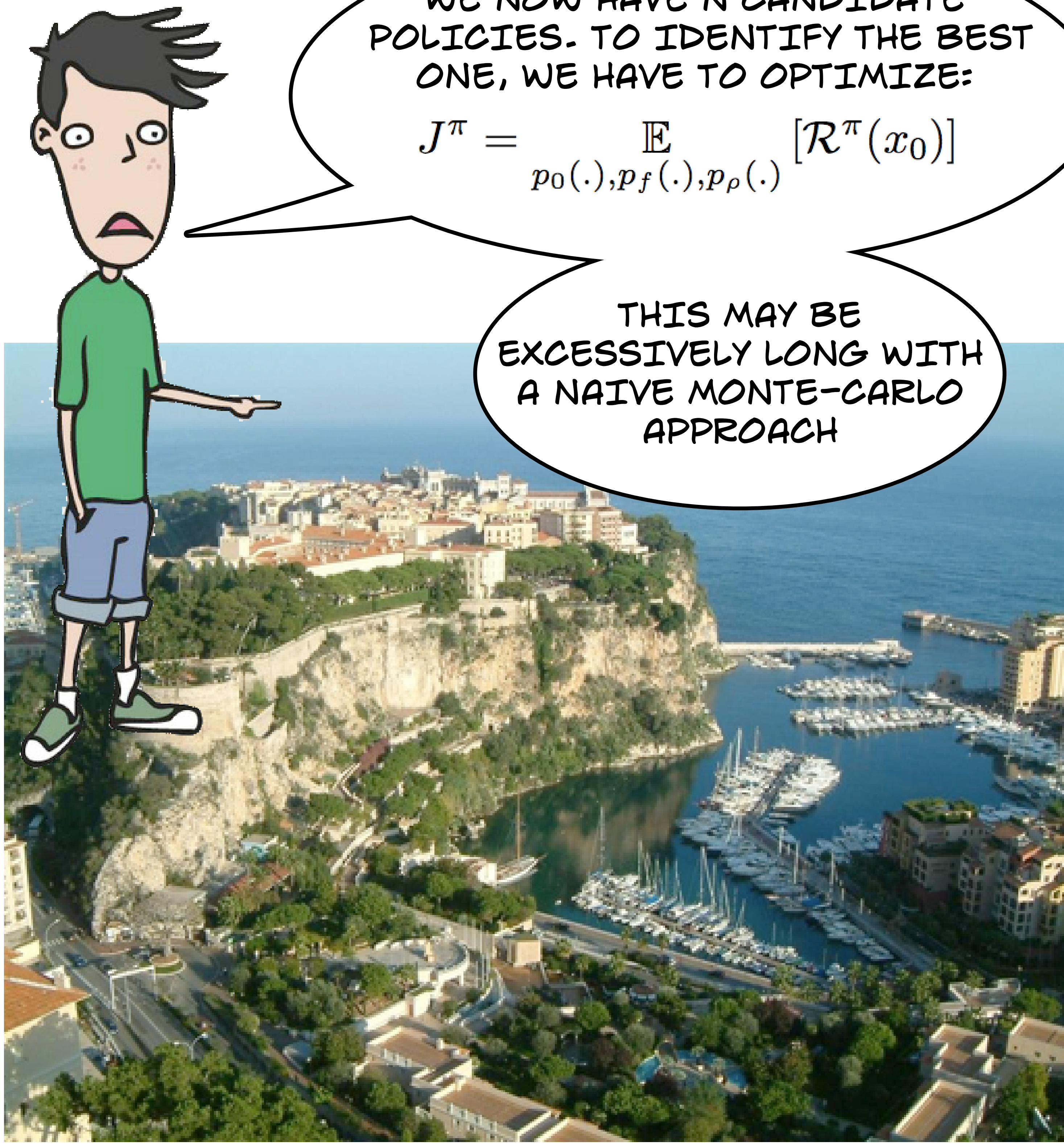
WE CLUSTERIZE FORMULAS. TWO FORMULAS F AND F' ARE EQUIVALENT IFF:

$$\forall s \in \{s_1, \ldots, s_S\},$$
$$\arg\max_{u \in \mathcal{U}} F(s,u,r,y) = \arg\max_{u \in \mathcal{U}} F'(s,u,r,y)$$

3. AMONG EACH CLUSTER, WE SELECT A FORMULA OF MINIMAL LENGTH

4. WE GATHER ALL SELECTED FORMULAS OF MINIMAL LENGTH AND WE DENOTE:

$$\tilde{\Pi}^K_{int} = \{\pi_{F_1}, \ldots, \pi_{F_N}\}$$


WE NOW HAVE N CANDIDATE POLICIES. TO IDENTIFY THE BEST ONE, WE HAVE TO OPTIMIZE:
$J^{\pi} = \underset{p_0(.),p_f(.),p_\rho(.)}{\mathbb{E}} [\mathcal{R}^{\pi}(x_0)]$
THIS MAY BE EXCESSIVELY LONG WITH A NAIVE MONTE-CARLO APPROACH


I MAY HAVE A SOLUTION TO THIS PROBLEM !


WELCOME
TO Fabulous
LAS VEGAS
NEVADA

A MULTI-ARMED BANDIT PROBLEM IS A SEQUENTIAL GAME

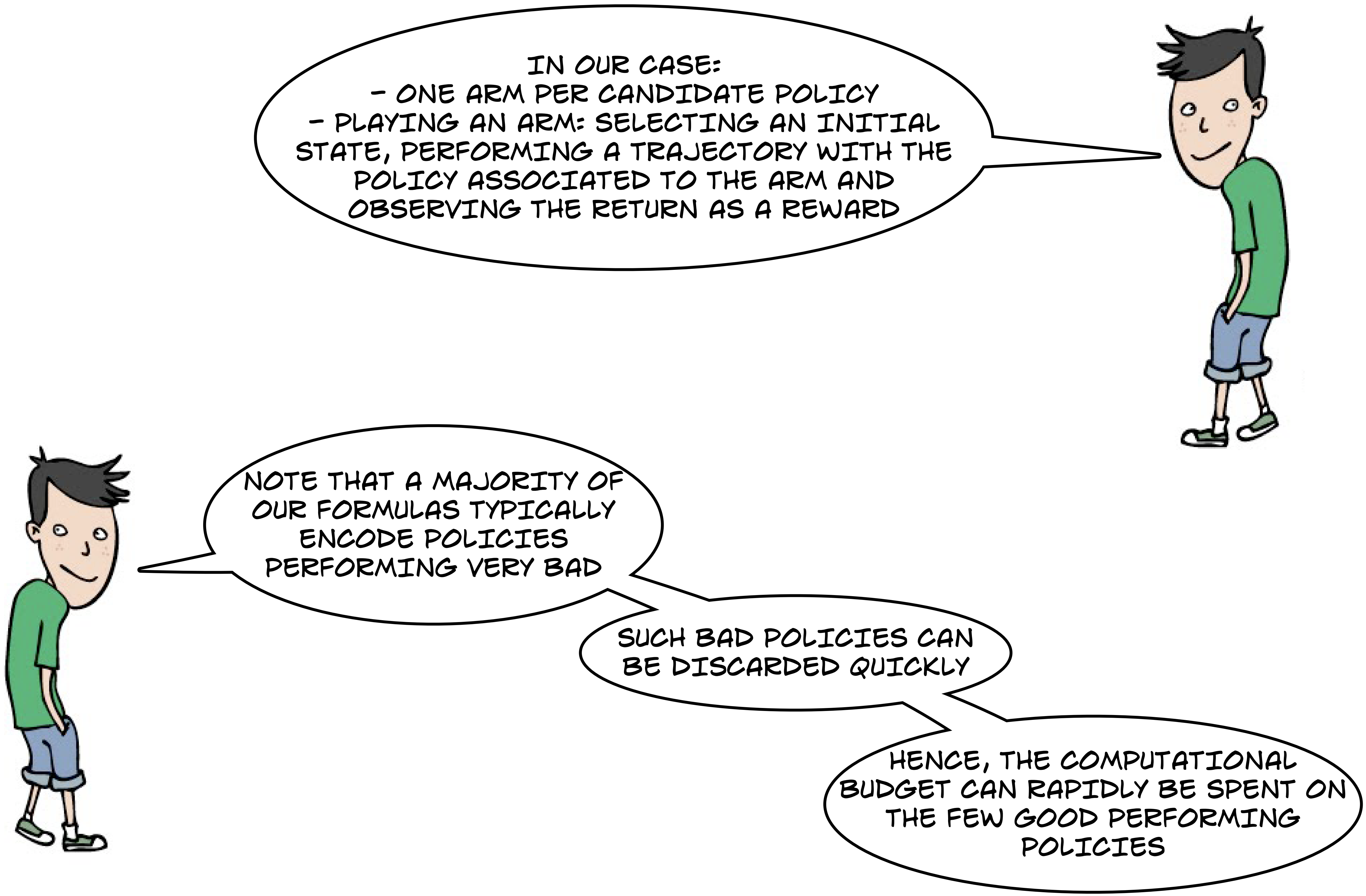
- N ARMS WITH UNKNOWN REWARD DISTRIBUTIONS
- AT EACH STEP, THE PLAYER SELECTS ONE OF THE ARMS AND RECEIVES A REWARD DRAWN FROM THE ASSOCIATED DISTRIBUTION
- BEST ARM IDENTIFICATION: WITHIN A FINITE NUMBER OF STEPS T, SELECT WHICH ARMS TO PLAY SO AS TO IDENTIFY THE ARM WHICH PERFORMS BEST ON AVERAGE

=> EXPLORATION / EXPLOITATION DILEMMA

EXPLORATION:
TRYING ARMS THAT MAY POTENTIALLY BE GOOD

EXPLOITATION:
FOCUSING ON ARMS THAT WE ALREADY KNOW TO BE GOOD

A GOOD MULTI-ARMED BANDIT STRATEGY BALANCES EXPLORATION AND EXPLOITATION


IN OUR CASE:
- ONE ARM PER CANDIDATE POLICY
- PLAYING AN ARM: SELECTING AN INITIAL STATE, PERFORMING A TRAJECTORY WITH THE POLICY ASSOCIATED TO THE ARM AND OBSERVING THE RETURN AS A REWARD
NOTE THAT A MAJORITY OF OUR FORMULAS TYPICALLY ENCODE POLICIES PERFORMING VERY BAD
SUCH BAD POLICIES CAN BE DISCARDED QUICKLY
HENCE, THE COMPUTATIONAL BUDGET CAN RAPIDLY BE SPENT ON THE FEW GOOD PERFORMING POLICIES

1. CONSTRUCT THE APPROXIMATE SET OF CANDIDATE POLICIES

$$\tilde{\Pi}_{int}^{K} = \{\pi_{F_1}, \ldots, \pi_{F_N}\}$$

2. PLAY EACH ARM ONCE (= DRAW ONE TRAJECTORY PER CANDIDATE POLICY)

3. WHILE THERE IS TRAINING TIME:

   A) SELECT THE ARM WHICH MAXIMIZES $A_{n,t} = \bar{r}_{n,t} + \frac{\alpha}{\theta_{n,t}}$

   WHERE $\bar{r}_{n,t}$ IS THE EMPIRICAL MEAN OF RETURNS ASSOCIATED TO π_{F_n}

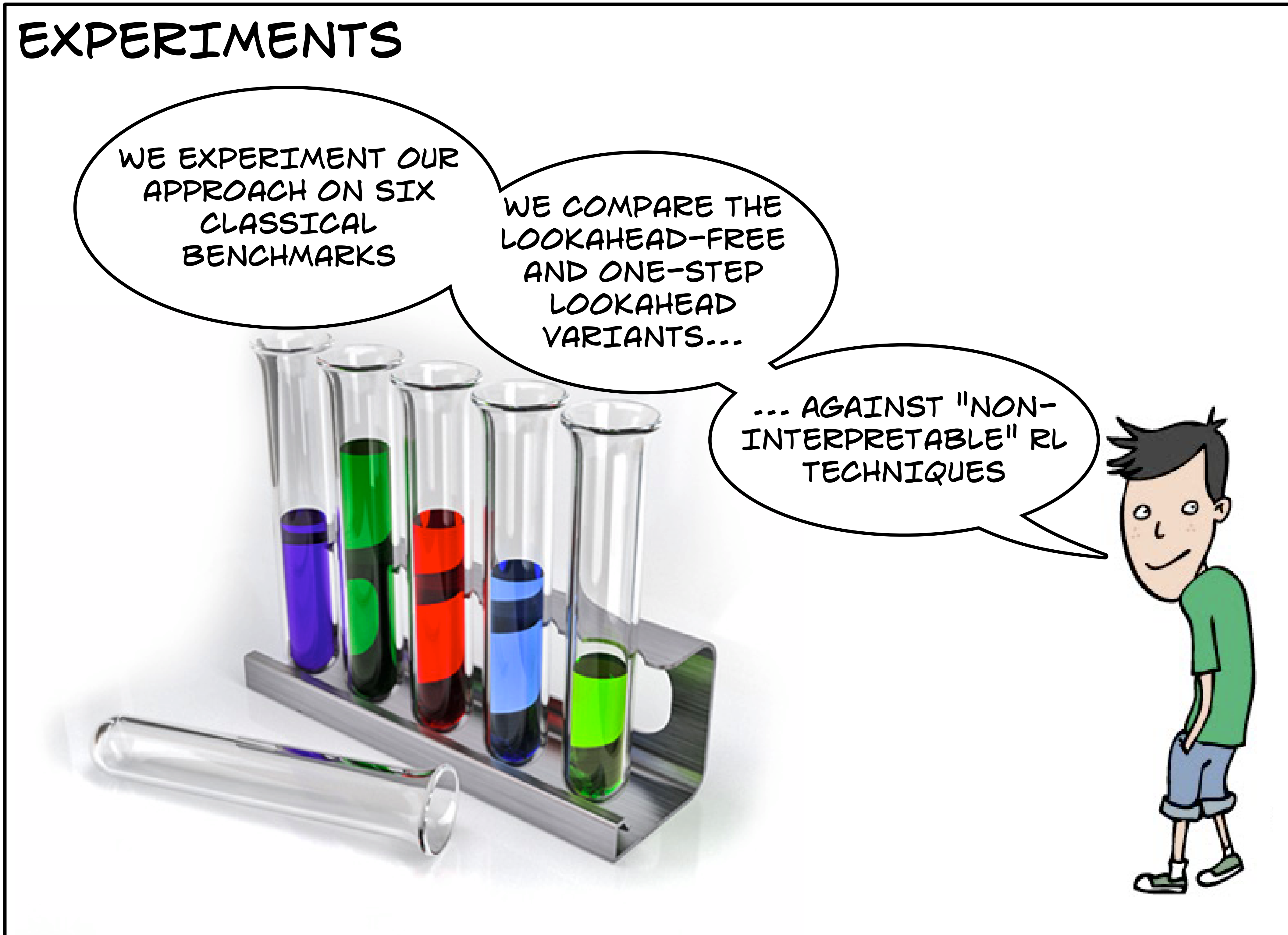
   $\theta_{n,t}$ IS THE NUMBER OF TIMES π_{F_n} HAS BEEN PLAYED

   $\alpha > 0$ IS AN EXPLORATION/EXPLOITATION TRADEOFF CONSTANT

   B) DRAW AN INITIAL STATE, PERFORM ONE TRAJECTORY WITH π_{F_n} AND OBSERVE THE RETURN

   C) UPDATE $\bar{r}_{n,t}$ AND $\theta_{n,t}$

4. RETURN THE POLICY (OR POLICIES) THAT MAXIMIZE(S) $\bar{r}_{n,t}$


EXPERIMENTS
WE EXPERIMENT OUR APPROACH ON SIX CLASSICAL BENCHMARKS
WE COMPARE THE LOOKAHEAD-FREE AND ONE-STEP LOOKAHEAD VARIANTS...
... AGAINST "NON-INTERPRETABLE" RL TECHNIQUES

BENCHMARK	LP	LOR	CAR	ACR	B	HIV
$d_{\mathcal{X}}$	2	1	2	4	5	6
$d_{\mathcal{U}}$	1	1	1	1	2	2
m	2	2	2	2	9	4
Stoch.	no	yes	no	no	yes	no
$\#\mathbb{V}_{LF}$	3	2	3	5	7	8
$\#\mathbb{V}_{OL}$	6	4	6	10	13	15
γ	.9	.75	.95	.95	.98	.98
T	50	20	1000	100	5e4	300

WE TESTED OUR APPROACH WITH 4 SETTINGS

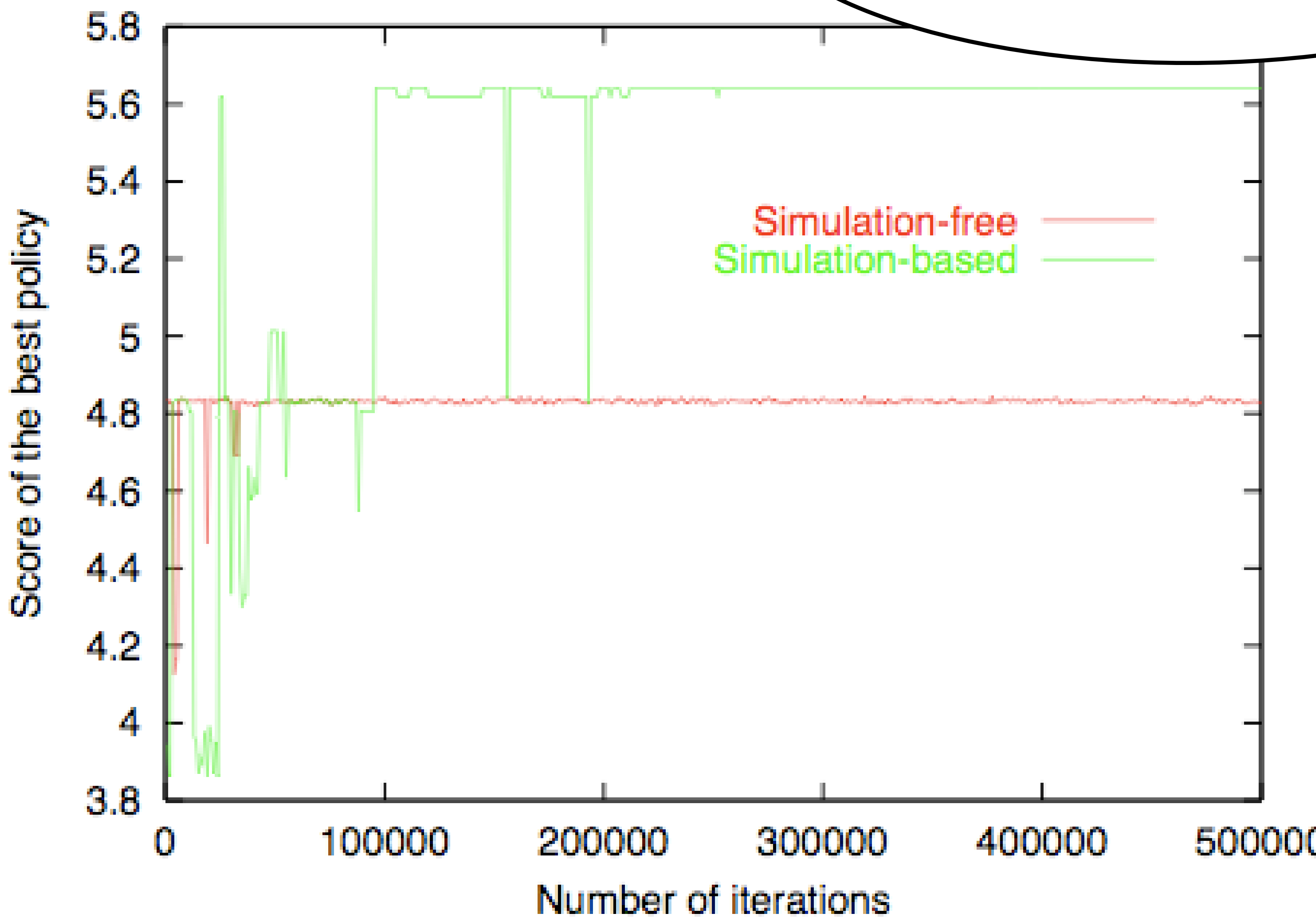
TO DISCRIMINATE AMONGS THE FORMULAS, WE USE SAMPLES OF 100 STATE POINTS

WE PERFORM LEARNING WITH:
K=5 => 10^6 BANDIT STEPS
K=6 => 10^7 BANDIT STEPS
WE USE A FIXED VALUE $\alpha = 2$

WE COMPARE AGAINST THE FOLLOWING TECHNIQUES:
- RANDOM POLICY
- LOOKAHEAD POLICIES
- FITTED Q-ITERATION

	CANDIDATE FORMULAS	CANDIDATE POLICIES
K=5, LOOKAHEAD-FREE	80,000 - 340,000	500 - 11,500
K=5, ONE-STEP LOOKAHEAD	140,000 - 990,000	3,800 - 95,000
K=6, LOOKAHEAD-FREE	1,000,000 - 5,500,000	3,600 - 132,000
K=6, ONE-STEP LOOKAHEAD	2,100,000 - 18,500,000	31,000 - 1,200,000


THIS FIGURE SHOWS A TYPICAL RUN OF THE ALGORITHM (K=5, LINEAR POINT BENCHMARK)
IN THIS SETTING, WE HAVE 907 (RESP. 12,214) CANDIDATE POLICIES IN THE LOOKAHEAD FREE (RESP. ONE-STEP LOOKAHEAD) SETTING

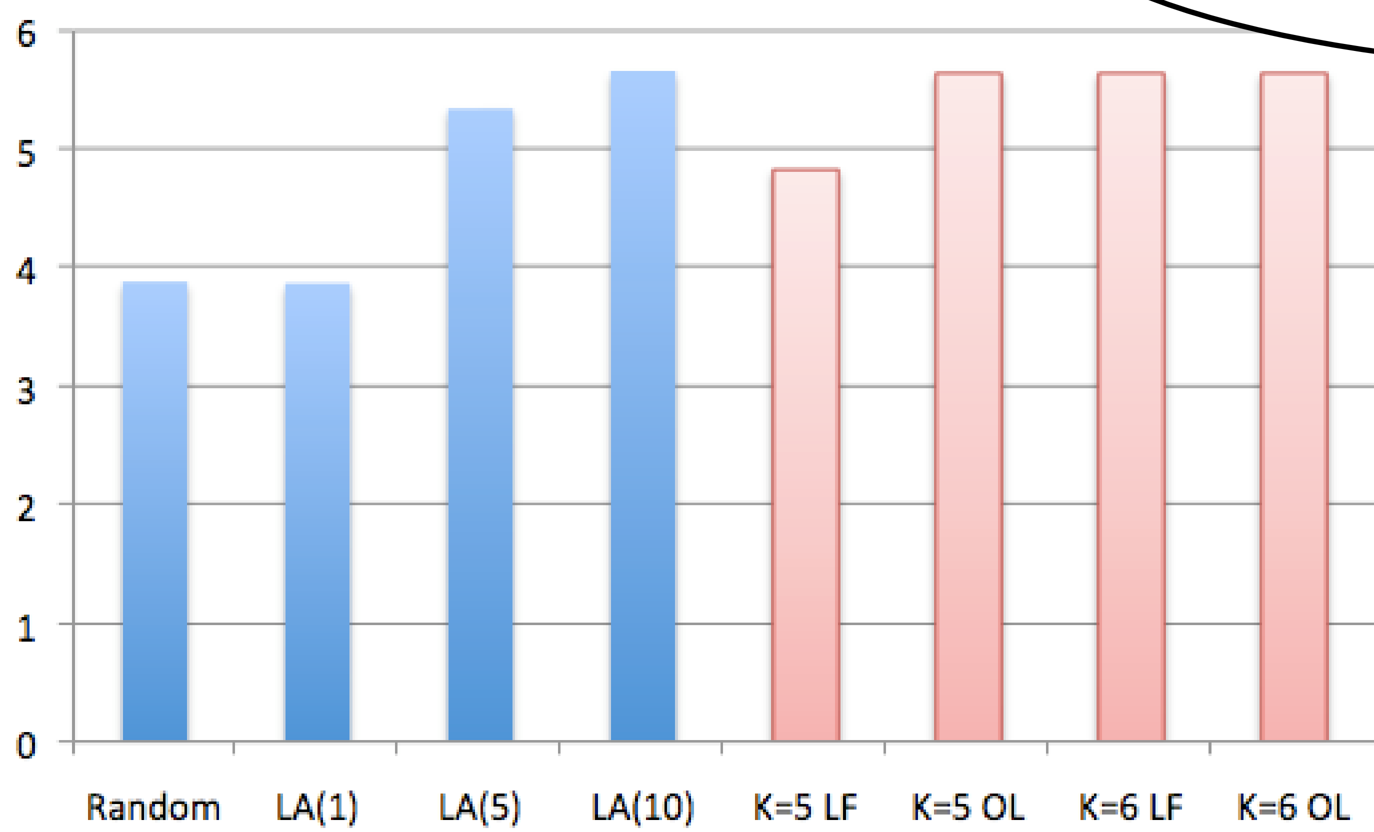

Score of the best policy
5.8
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
0
100000
200000
300000
400000
500000
Number of iterations
Simulation-free
Simulation-based

IN THREE CASES, WE FIND AS GOOD POLICIES AS THE LOOKAHEAD POLICY OF DEPTH 10

THE "K=6 LF" DISCOVERED FORMULA IS: $F^* = (-y - v)a$

WHERE y AND v ARE THE STATE VARIABLES AND a IS THE ACTION VARIABLE

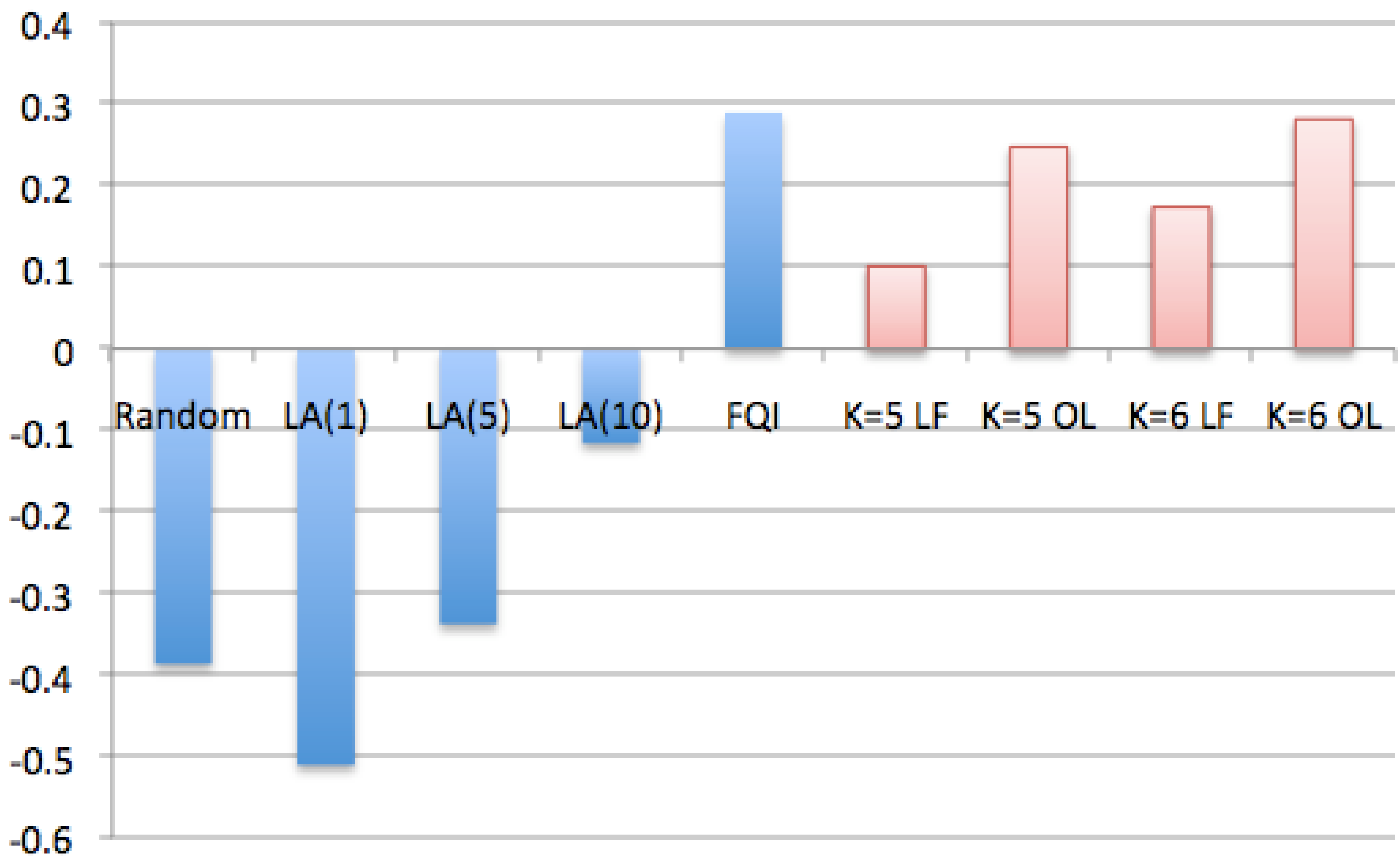
SINCE THE ACTION a IS EITHER −1 OR +1, THE POLICY CAN BE TERMED AS: "CHOOSE −1 WHEN y > -v AND +1 OTHERWISE"

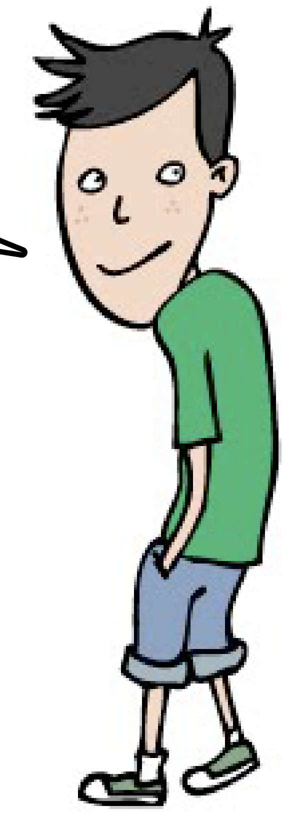


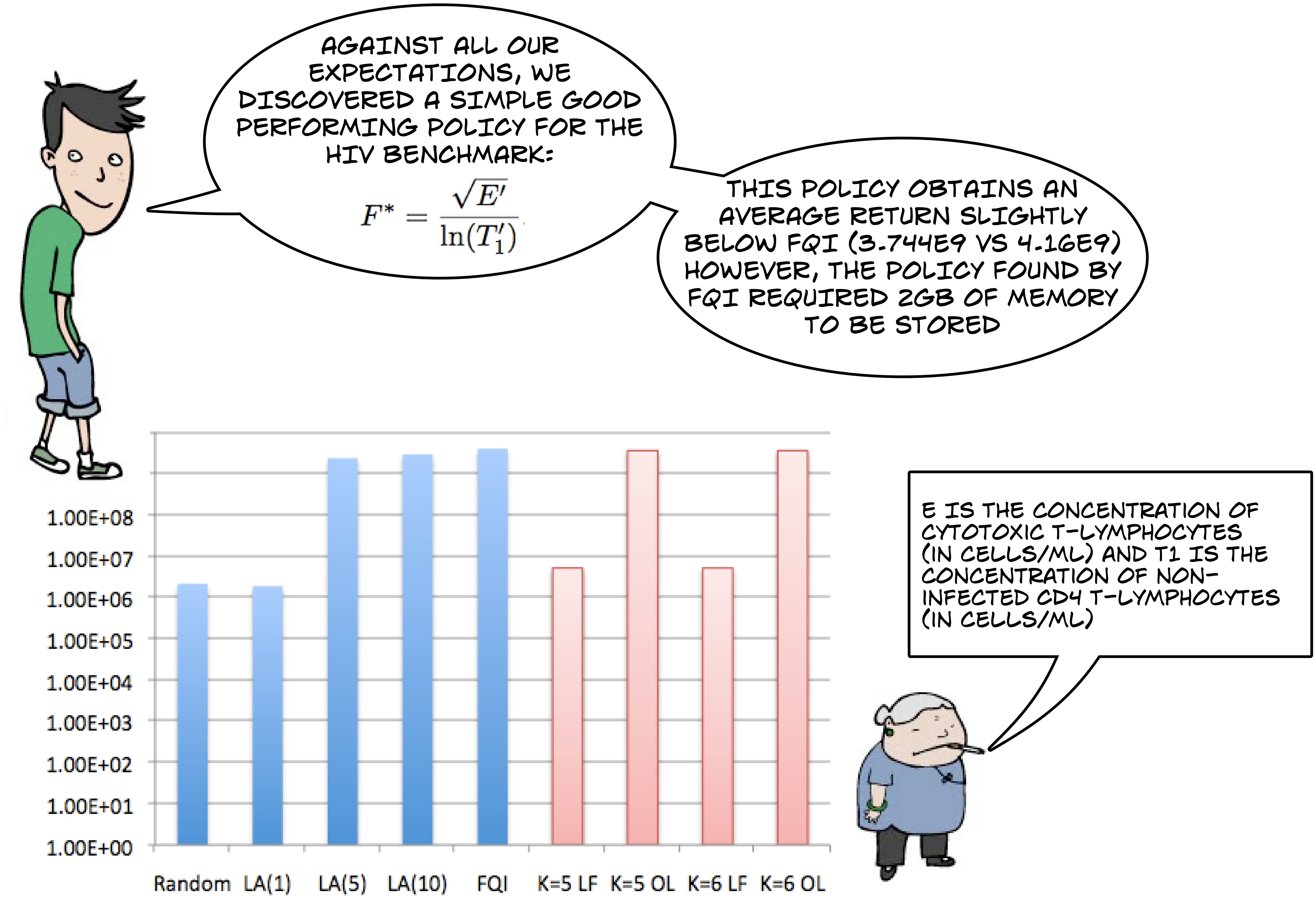
OUR BEST INTERPRETABLE POLICY PERFORMS NEARLY AS WELL AS FQI (0.282 VS 0.29)

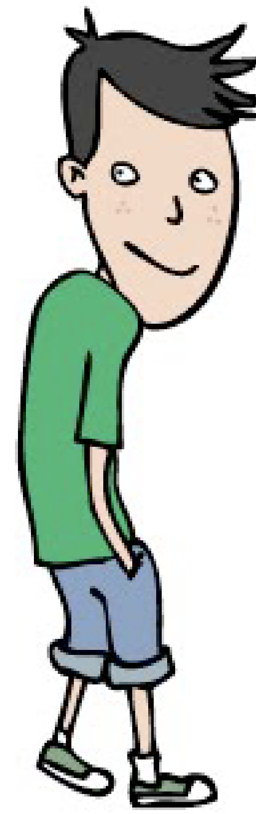
THE K=6 OL DISCOVERED FORMULA IS:

$$r - \frac{1}{\max(p', s')}$$






AGAINST ALL OUR EXPECTATIONS, WE DISCOVERED A SIMPLE GOOD PERFORMING POLICY FOR THE HIV BENCHMARK:
$F^* = \frac{\sqrt{E'}}{\ln(T_1')}$
THIS POLICY OBTAINS AN AVERAGE RETURN SLIGHTLY BELOW FQI (3.744E9 VS 4.16E9) HOWEVER, THE POLICY FOUND BY FQI REQUIRED 2GB OF MEMORY TO BE STORED
1.00E+08
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
1.00E+02
1.00E+01
1.00E+00
Random
LA(1)
LA(5)
LA(10)
FQI
K=5 LF
K=5 OL
K=6 LF
K=6 OL
E IS THE CONCENTRATION OF CYTOTOXIC T-LYMPHOCYTES (IN CELLS/ML) AND T1 IS THE CONCENTRATION OF NON-INFECTED CD4 T-LYMPHOCYTES (IN CELLS/ML)

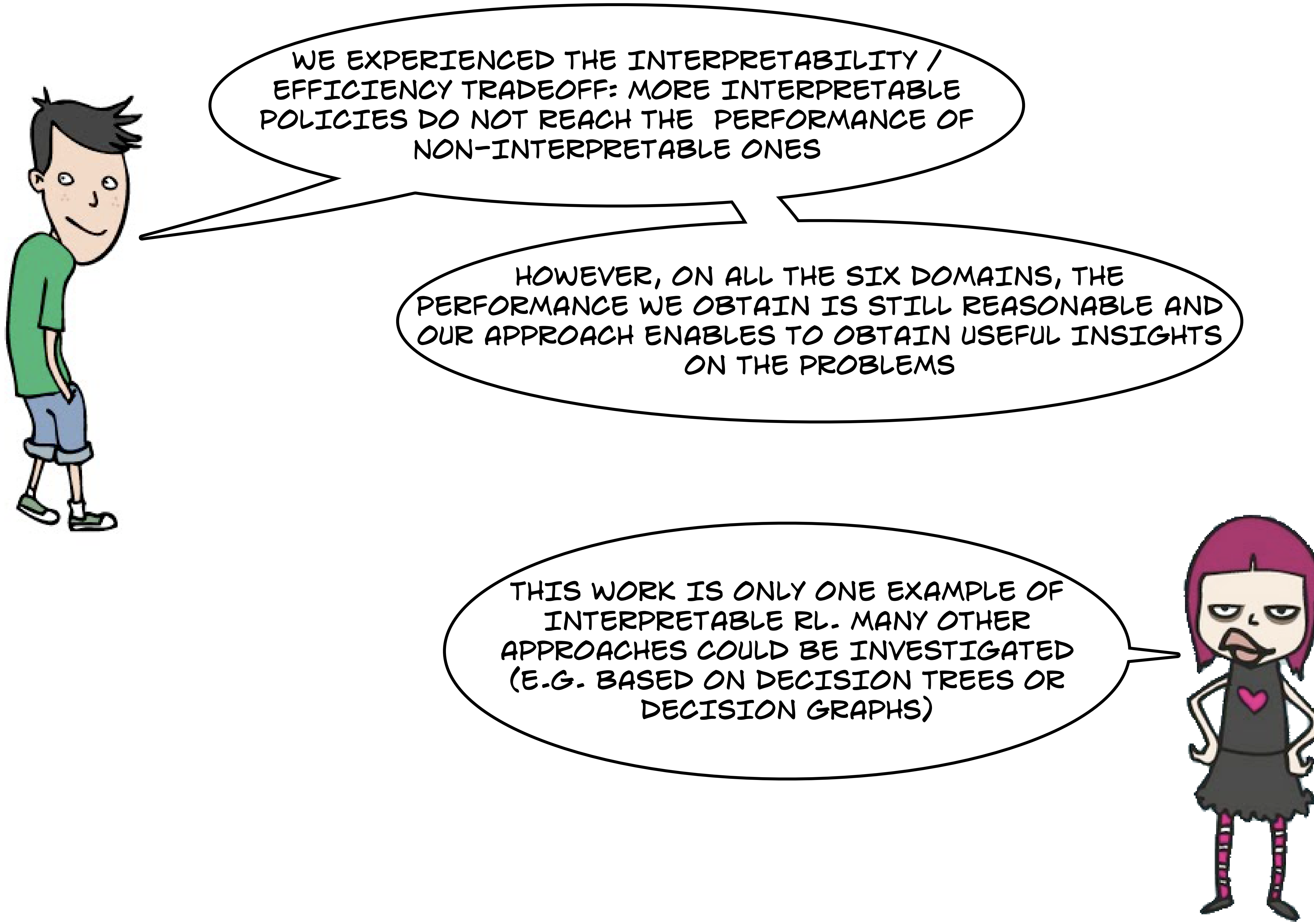


LOOKAHEAD-FREE FORMULAS
$(-y-v)a$
$u/(x-\sqrt{5})$
$u(\sqrt{7}-s)$
$\max(\dot{\theta}_2/u, \sqrt{2})$
$\psi\dot{\theta}-\lvert d\rvert$
$1/(\epsilon_1-\frac{T_2}{T_1^*})$

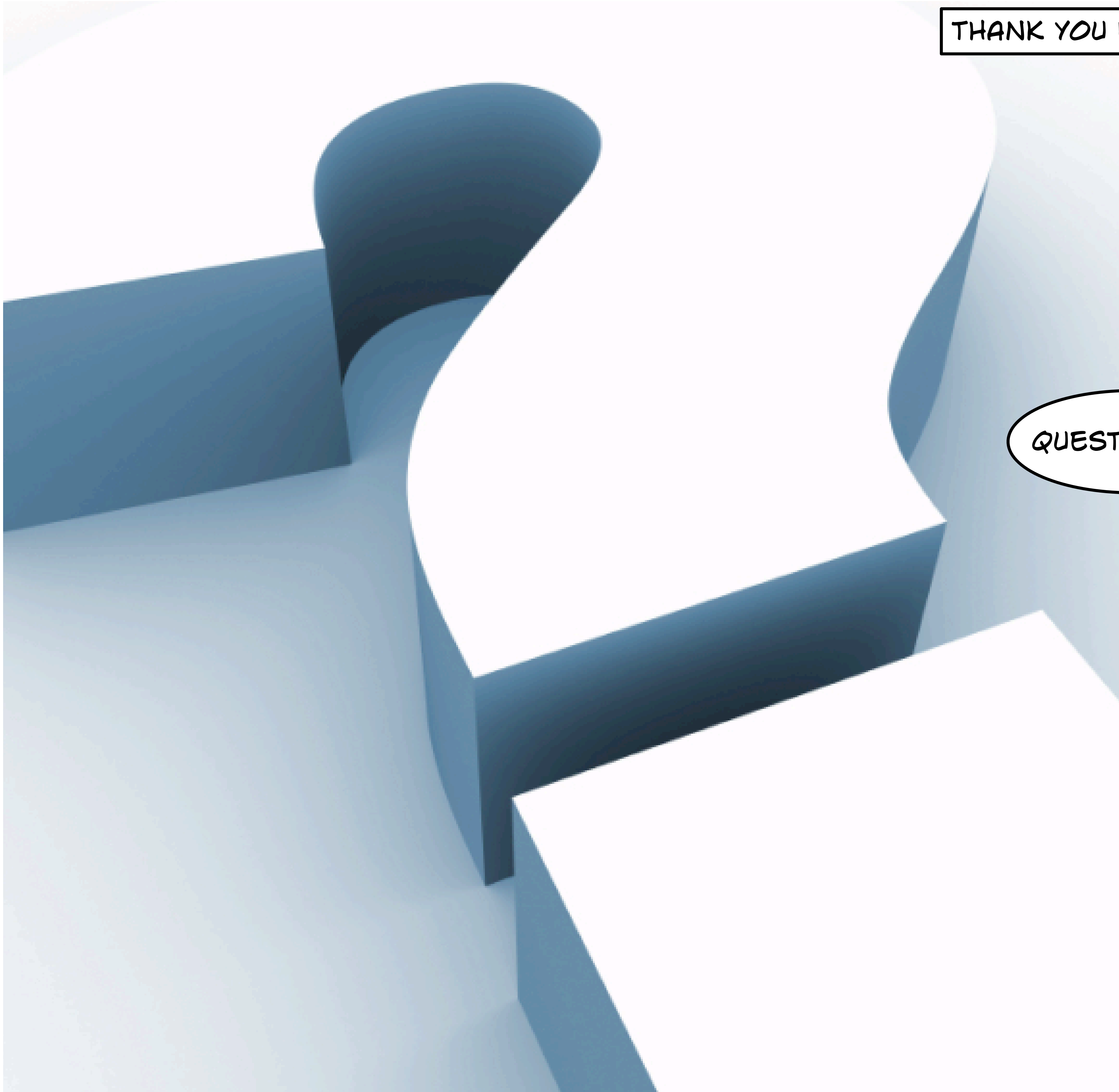
ONE-STEP LOOKAHEAD FORMULAS
$y'-\lvert y+v'\rvert$
$u/(x-\sqrt{7})$
$r-\frac{1}{\max(p',s')}$
$\dot{\theta}_2\lvert\dot{\theta}_2'\rvert-u$
$1/(7-\dot{\theta}'/\dot{\omega}')$
$\sqrt{E'}/\ln(T_1')$

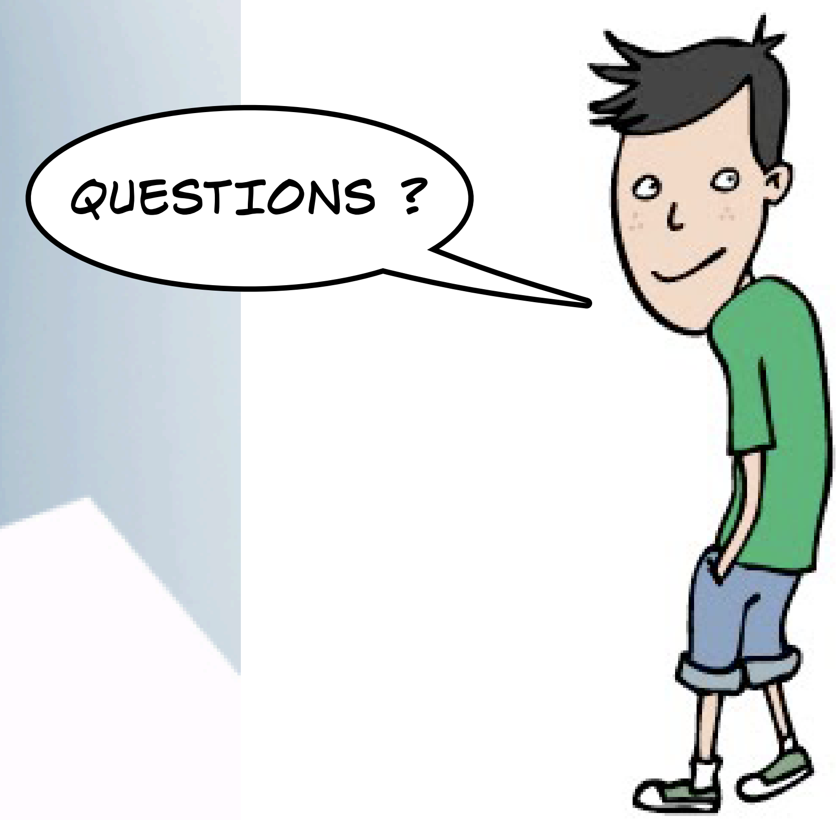
CONCLUSIONS


WE INTRODUCED THE ISSUE OF INTERPRETABILITY IN REINFORCEMENT LEARNING
ALTHOUGH IT IS AN OLD TOPIC IN SUPERVISED LEARNING, THIS ISSUE HAS RECEIVED VERY FEW ATTENTION IN THE FIELD OF RL
WE FOCUSED ON THE CLASS OF INDEX-BASED POLICIES DESCRIBED BY SMALL INDEX FORMULAS
WE INTRODUCED A DIRECT POLIC SEARCH SCHEME BASED ON MULTI-ARMED BANDITS
WE TESTED OUR APPROACH ON SIX BENCHMARK PROBLEMS AND DISCOVERED SIMPLE – REASONABLY EFFICIENT – POLICIES IN EACH CASE


WE EXPERIENCED THE INTERPRETABILITY / EFFICIENCY TRADEOFF: MORE INTERPRETABLE POLICIES DO NOT REACH THE PERFORMANCE OF NON-INTERPRETABLE ONES
HOWEVER, ON ALL THE SIX DOMAINS, THE PERFORMANCE WE OBTAIN IS STILL REASONABLE AND OUR APPROACH ENABLES TO OBTAIN USEFUL INSIGHTS ON THE PROBLEMS
THIS WORK IS ONLY ONE EXAMPLE OF INTERPRETABLE RL. MANY OTHER APPROACHES COULD BE INVESTIGATED (E.G. BASED ON DECISION TREES OR DECISION GRAPHS)

THANK YOU FOR YOUR ATTENTION !





SEE ALSO ...

Monte Carlo Search Algorithm Discovery for One Player Games. Francis Maes, David Lupien St-Pierre and Damien Ernst.

Automatic discovery of ranking formulas for playing with multi-armed bandits. Francis Maes, Louis Wehenkel and Damien Ernst. In 9th European workshop on reinforcement learning (EWRL'11), Athens, Greece, September 2011.

Meta-Learning of Exploration/Exploitation Strategies: The Multi-Armed Bandit Case. Francis Maes, Louis Wehenkel and Damien Ernst.

Learning exploration/exploitation strategies for single trajectory reinforcement learning. Michael Castronovo, Francis Maes, Raphael Fonteneau and Damien Ernst. In 10th European Workshop on Reinforcement Learning (EWRL'12), Edinburgh, Scotland, June 2012.