ROB-GY 6323
reinforcement learning and optimal control for robotics

Lecture 14
Beyond the class

Course material

All necessary material will be posted on Brightspace
Code will be posted on the Github site of the class

https://github.com/righetti/optlearningcontrol

Discussions/Forum with Slack

Contact

ludovic.righetti@nyu.edu
Office hours in person
Wednesday 3pm to 4pm
370 Jay street - room 801

Course Assistant

Armand Jordana
aj2988@nyu.edu
Office hours Monday 1pm to 2pm
Rogers Hall 515

any other time by appointment only

Schedule

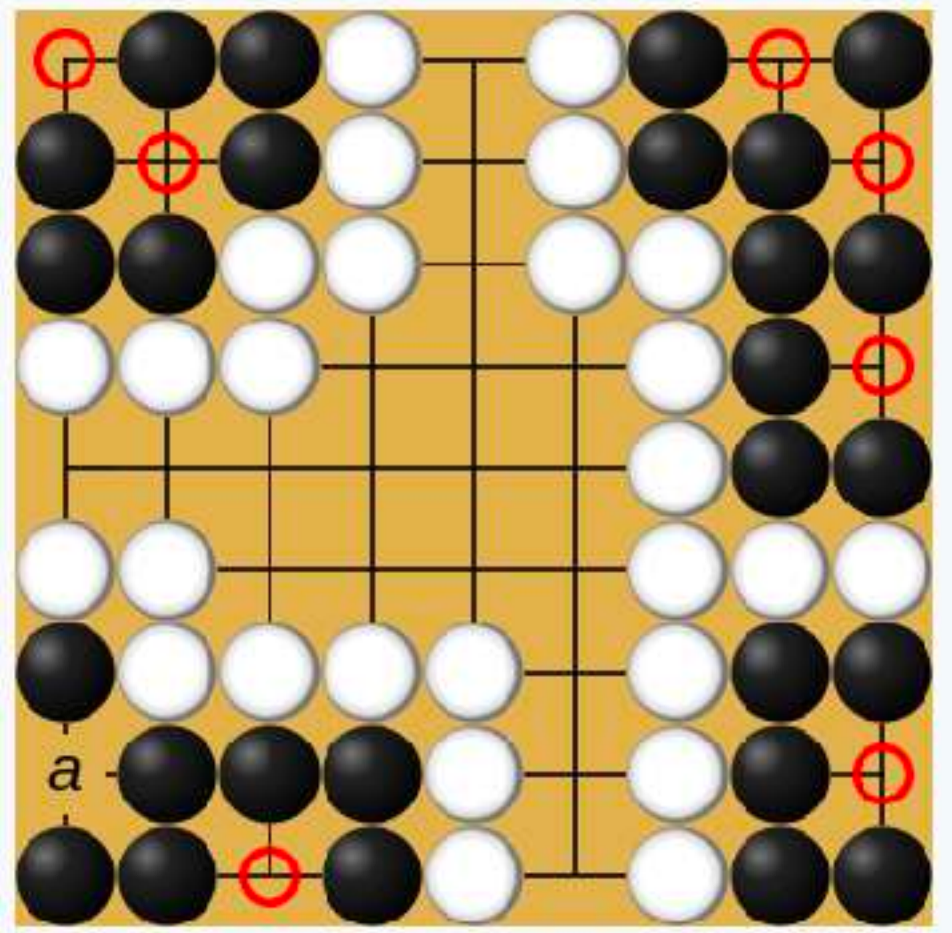
Week	Lecture		Homework	Project
1	Intro	Lecture 1: introduction		
2	Trajectory optimization	Lecture 2: Basics of optimization	HW 1	
3		Lecture 3: QPs		
4		Lecture 4: Nonlinear optimal control		
5		Lecture 5: Model-predictive control		
6		Lecture 6: Sampling-based optimal control	HW 2	
7	Policy optimization	Lecture 7: Bellman's principle		
8		Lecture 8:Value iteration / policy iteration		Project 1
9		Lecture 9: Q-learning	HW 3	
10		Lecture 10: Deep Q learning		
11		Lecture 11:Actor-critic algorithms		
12		Lecture 12: Learning by demonstration	HW 4	Project 2
13		Lecture 13: Monte-Carlo Tree Search		
14		Lecture 14: Beyond the class		
15	Finals week			

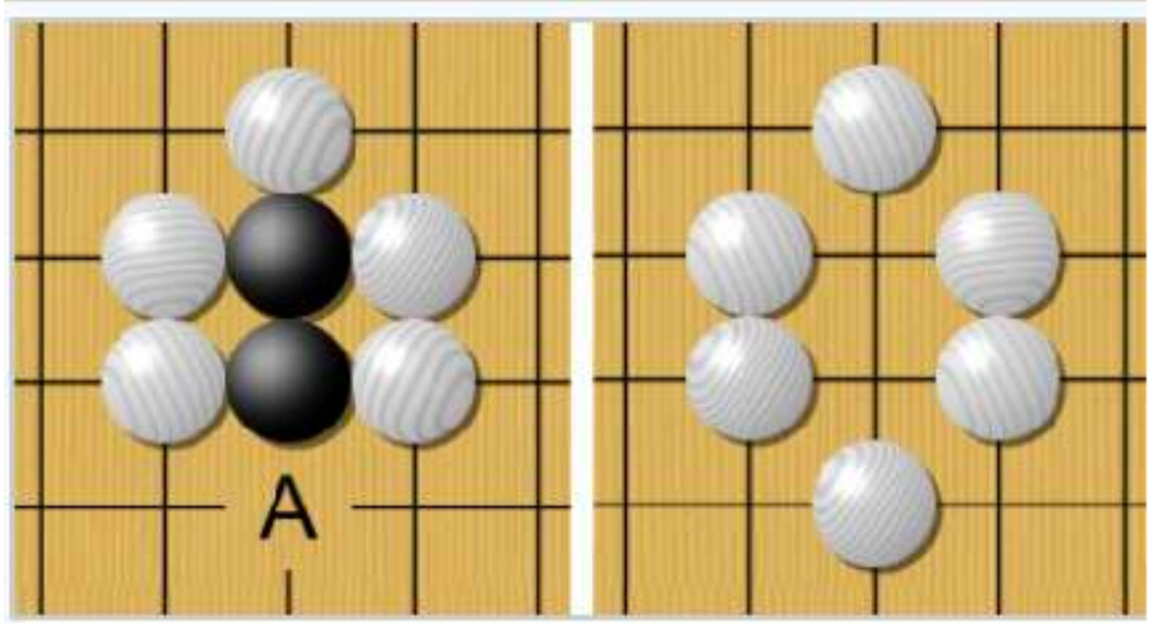
Paper report is due December 19th

Project 2 is due December 19th

Playing the game of Go

A mixture of imitation learning, OC and RL


a


A

Deciding how to play with tree search

Go branching factor

For a game typically b^d number of moves to test
b is “breadth”, i.e. number of legal moves at each turn
d is the “depth”, i.e. the game length

For Go b~250 and d~150

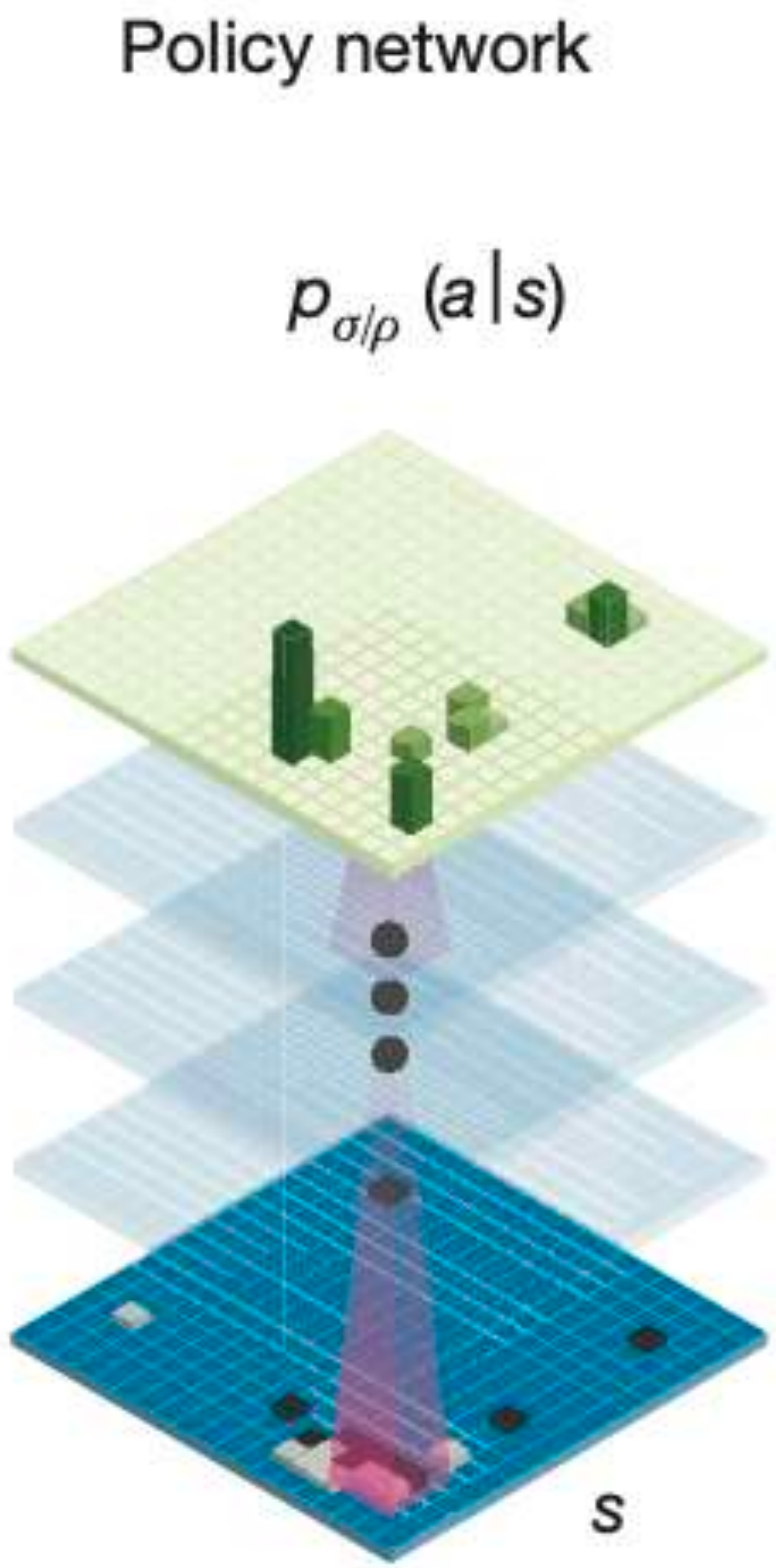
Speeding up search: Monte-Carlo Tree Search

Invented by R. Coulom in 2006 (not new!)

4 steps to be repeated N times

1. Selection
2. Expansion
3. Simulation
4. Backup

Stage 1: learn a policy from Human players



$p_\sigma(a|s)$

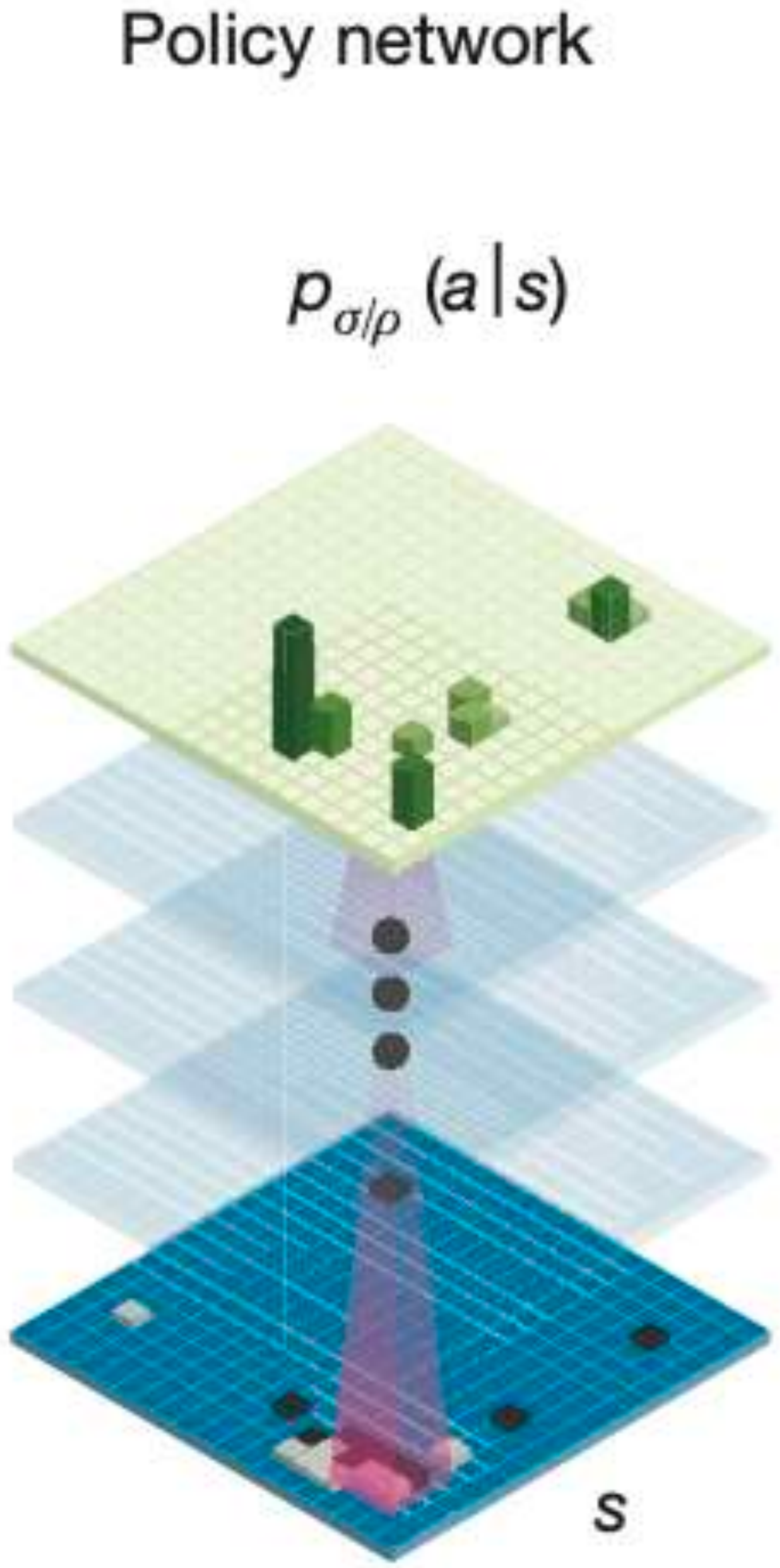
Policy learned with supervised learning SL-policy

13 layers neural network - accurate (57% / 55%) but slow to evaluate (3ms)

$p_\pi(a|s)$

Policy with smaller network - less accurate (24%) but fast to evaluate (2us)

Stage 2: improve policy using RL policy gradient



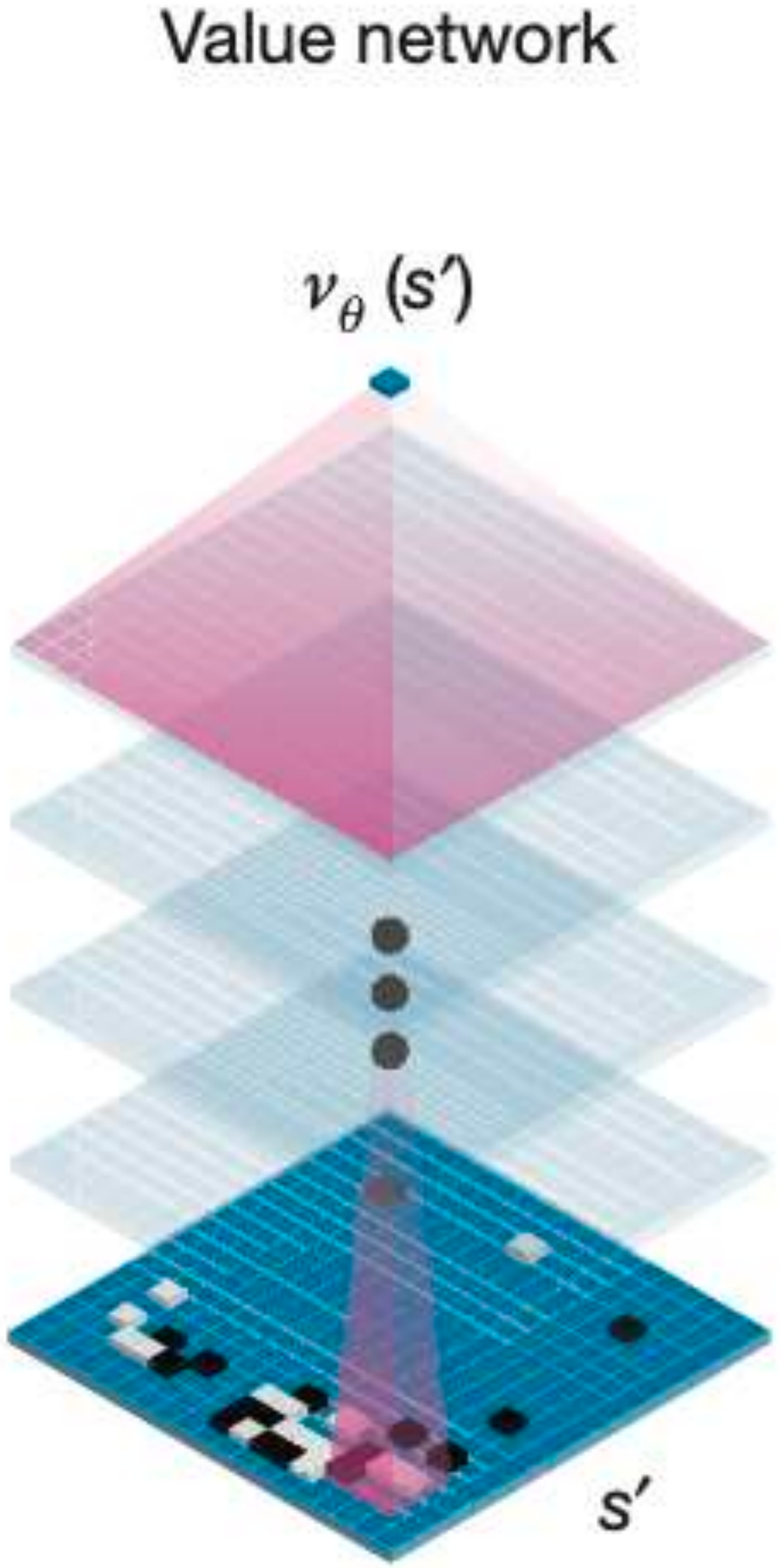
p_ρ Policy learned using RL

$$\Delta\rho \propto \frac{\partial \log p_\rho(a_t|s_t)}{\partial\rho} z_t$$

$$z_t = \pm r(s_T)$$

RL policy won 80% games agains SL policy

Stage 3: learning a value function

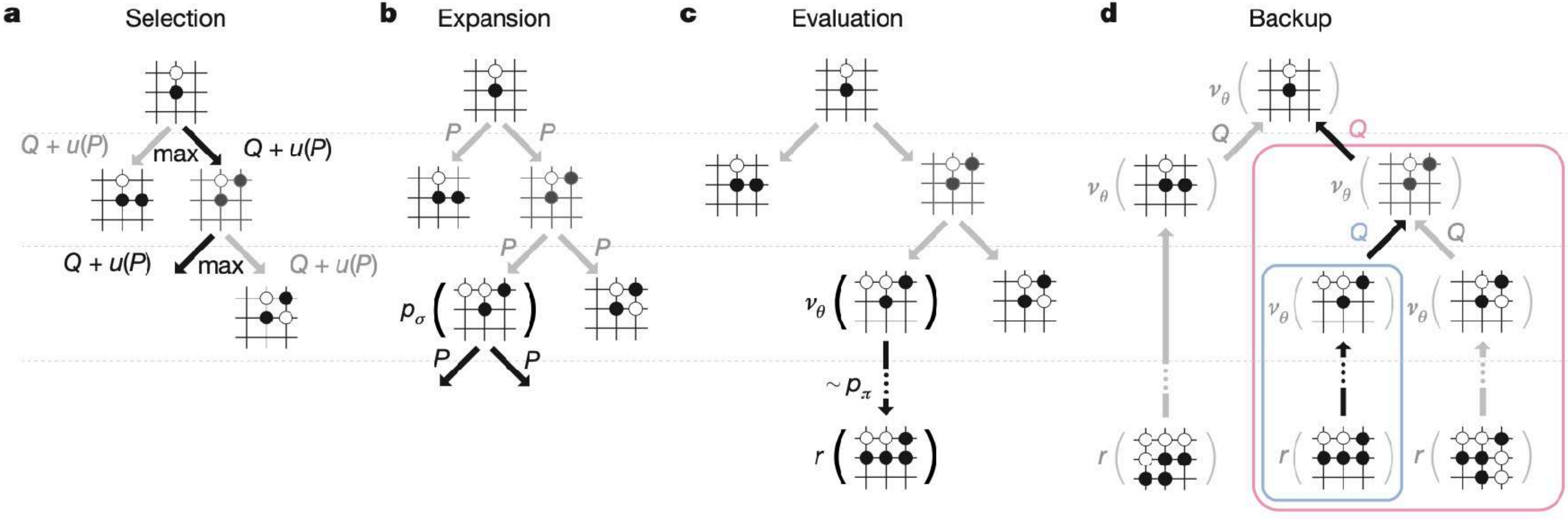


Self-play and randomization

$$v^p(s) = \mathbb{E}[z_t | s_t = s,\ a_{t \dots T} \sim p]$$

$$\Delta\theta \propto \frac{\partial v_\theta(s)}{\partial \theta}(z - v_\theta(s))$$

Stage 4: Monte-Carlo Tree Search



AlphaGoZero (2017)

Use self-play to learn a policy and value function (no SL)
Input features are only black and white stones (no other features)
Single NN for both policy and value
Simpler tree search - no evaluation through simulated play

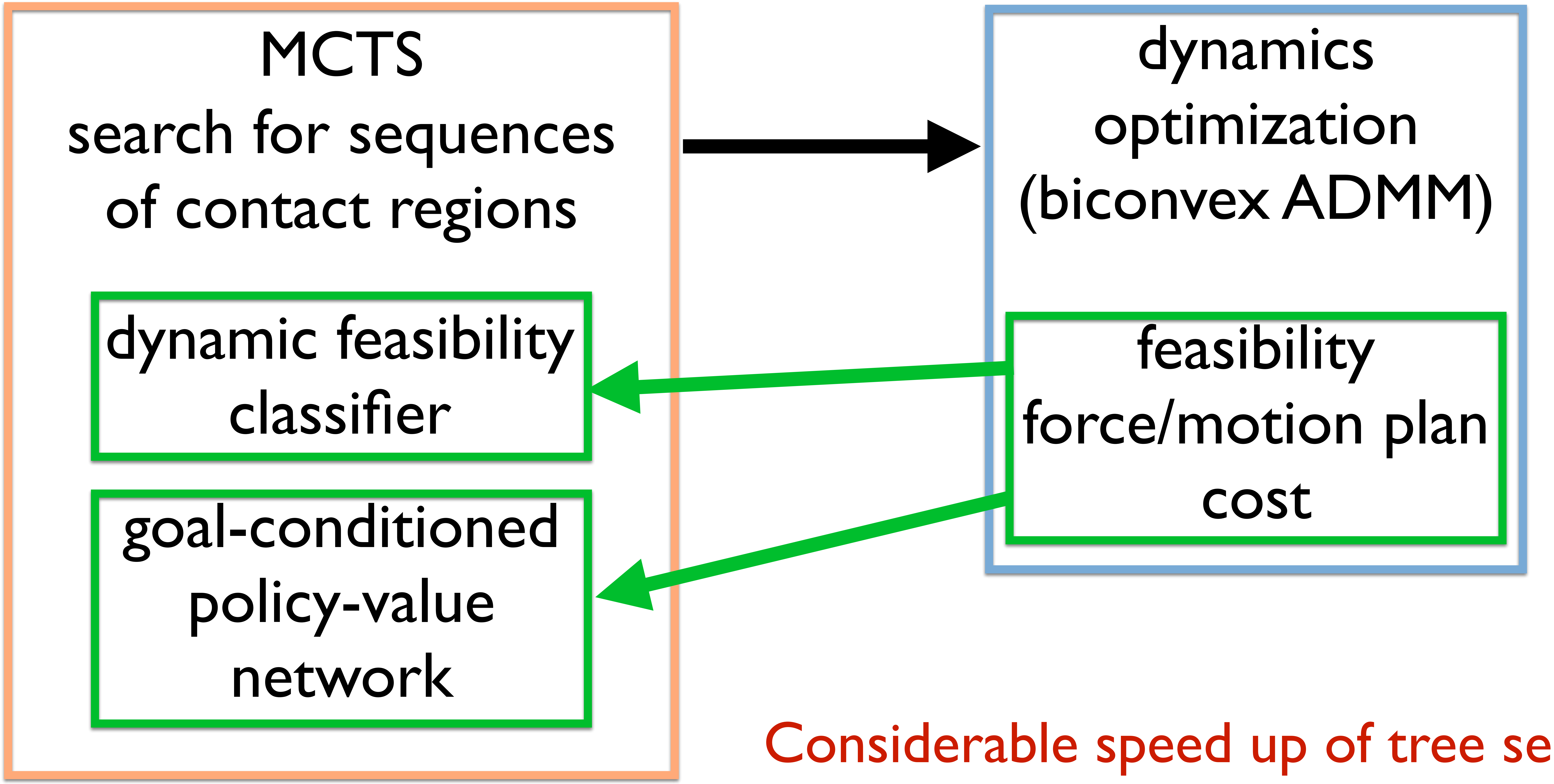
AlphaZero (2017)

Similar to AlphaGoZero but to play also Chess and Shogi

MuZero (2019)

Similar to AlphaGoZero but also learns the game model

Efficient manipulation planning with Monte-Carlo Tree Search

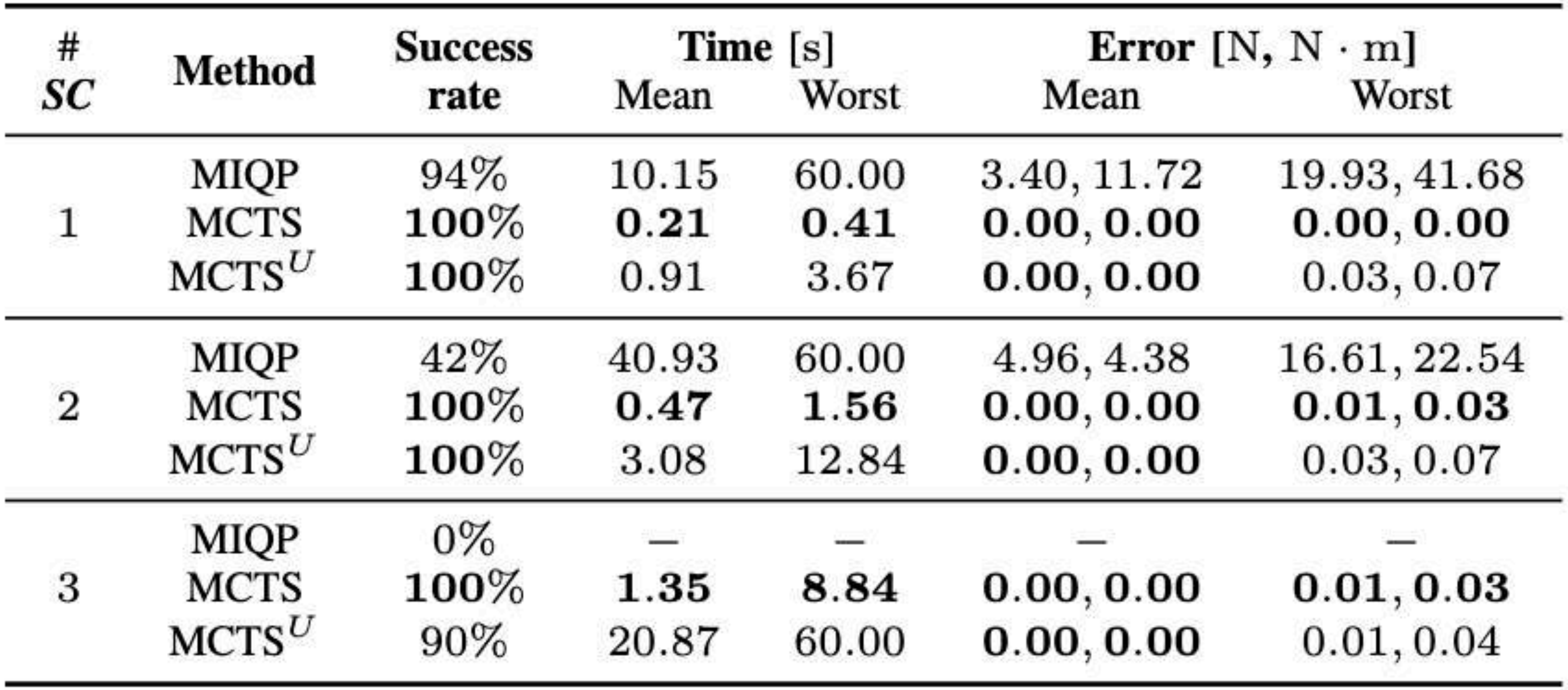

MCTS
search for sequences of contact regions
dynamic feasibility classifier
goal-conditioned policy-value network
dynamics optimization (biconvex ADMM)
feasibility
force/motion plan
cost

Considerable speed up of tree search

[Zhu et al., IROS 2023]

Efficient manipulation planning with Monte-Carlo Tree Search

sequence length

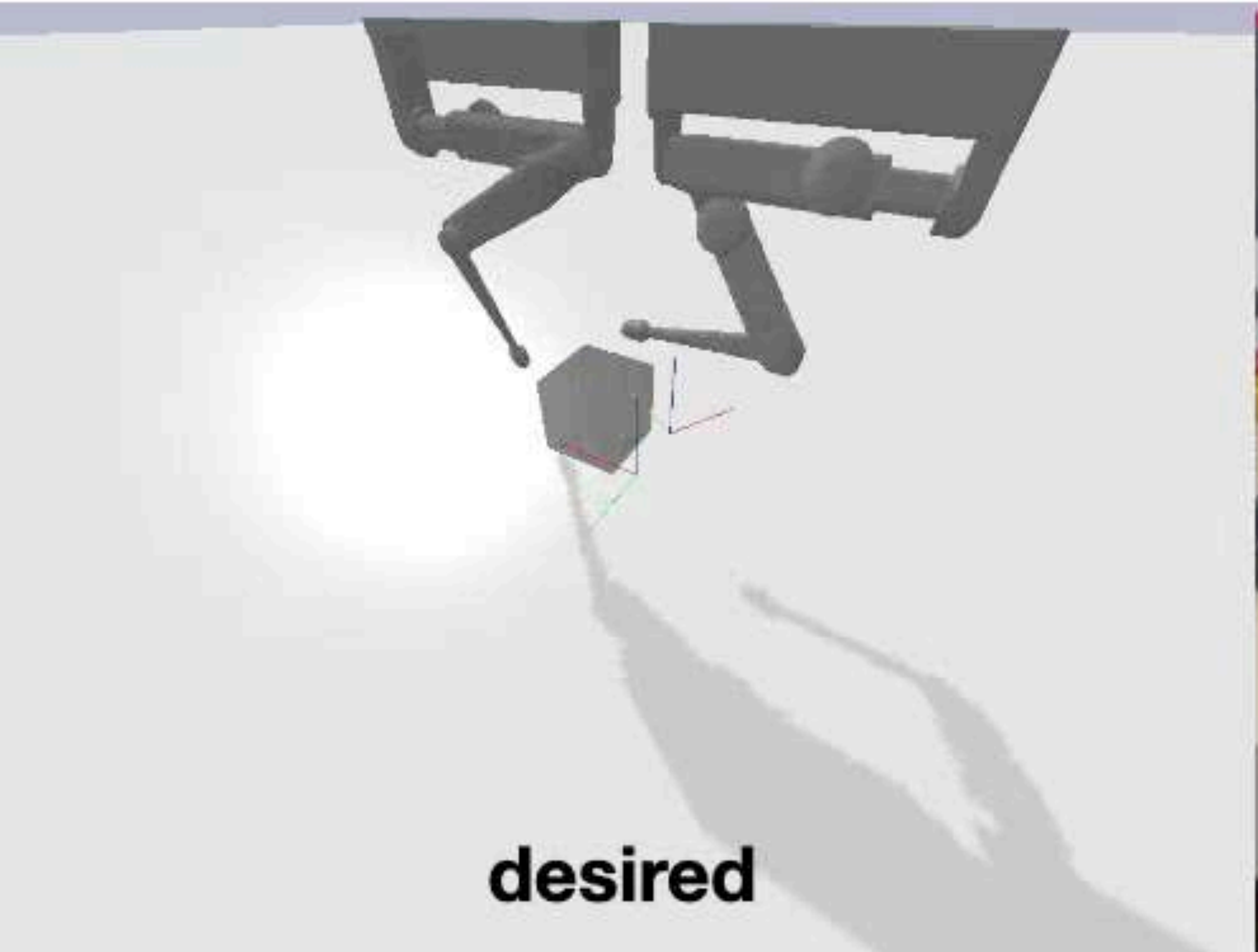


# ***SC***	**Method**	**Success rate**	**Time** [s] Mean	Worst	**Error** [N, N · m] Mean	Worst
1	MIQP	94%	10.15	60.00	$3.40, 11.72$	$19.93, 41.68$
	MCTS	**100%**	**0.21**	**0.41**	$\mathbf{0.00, 0.00}$	$\mathbf{0.00, 0.00}$
	MCTS^U	**100%**	0.91	3.67	$\mathbf{0.00, 0.00}$	$0.03, 0.07$
2	MIQP	42%	40.93	60.00	$4.96, 4.38$	$16.61, 22.54$
	MCTS	**100%**	**0.47**	**1.56**	$\mathbf{0.00, 0.00}$	$\mathbf{0.01, 0.03}$
	MCTS^U	**100%**	3.08	12.84	$\mathbf{0.00, 0.00}$	$0.03, 0.07$
3	MIQP	0%	–	–	–	–
	MCTS	**100%**	**1.35**	**8.84**	$\mathbf{0.00, 0.00}$	$\mathbf{0.01, 0.03}$
	MCTS^U	90%	20.87	60.00	$\mathbf{0.00, 0.00}$	$0.01, 0.04$

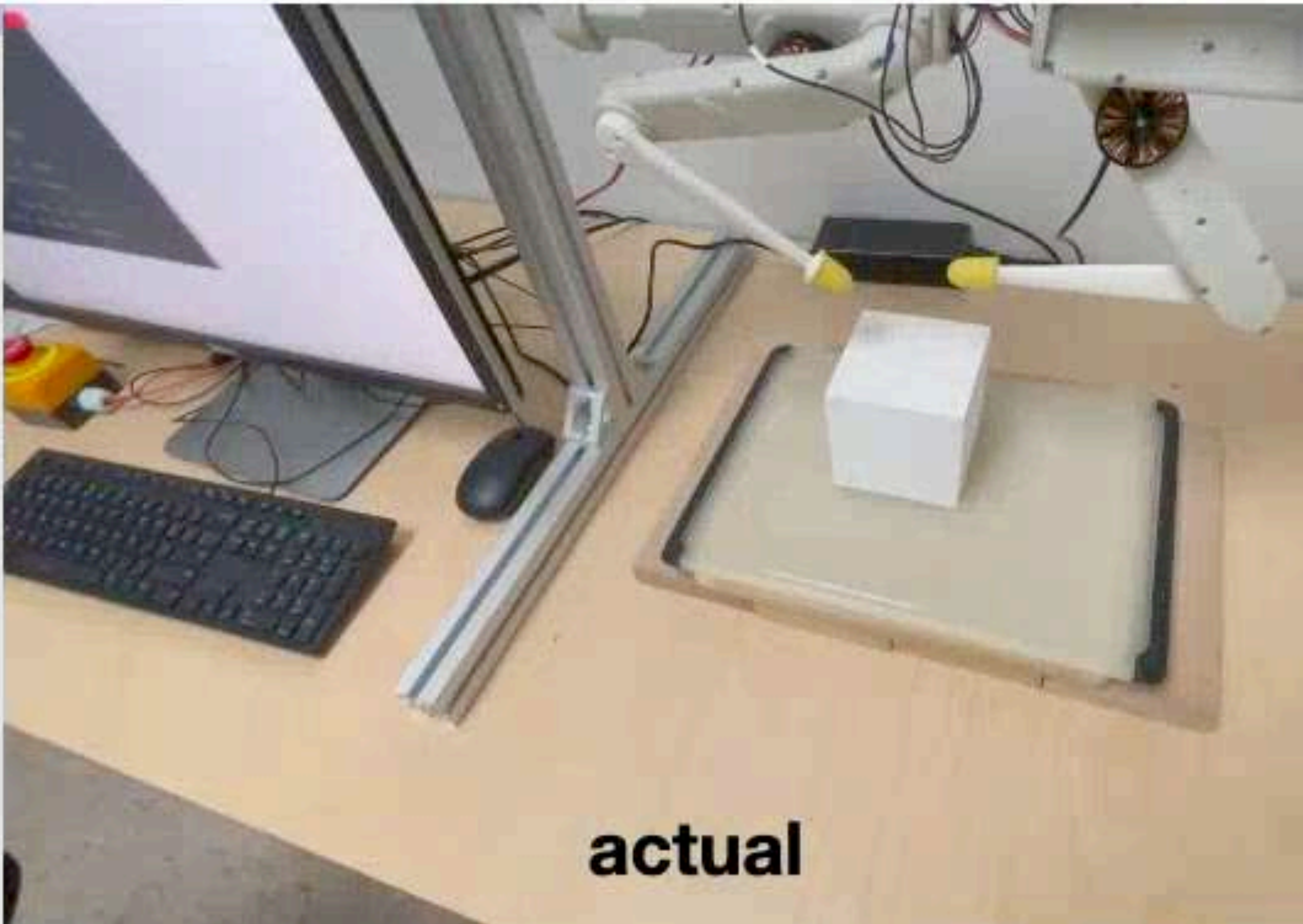
Composition of long manipulation sequences scales

[Zhu et al., IROS 2023]

Slide with Curvature Twice



desired



actual

MCTS is increasingly being used in robotics for problems with discrete state/actions

Works great with learned policy/value function

optimal control and reinforcement learning… what now?

Self-driving cars

Challenges

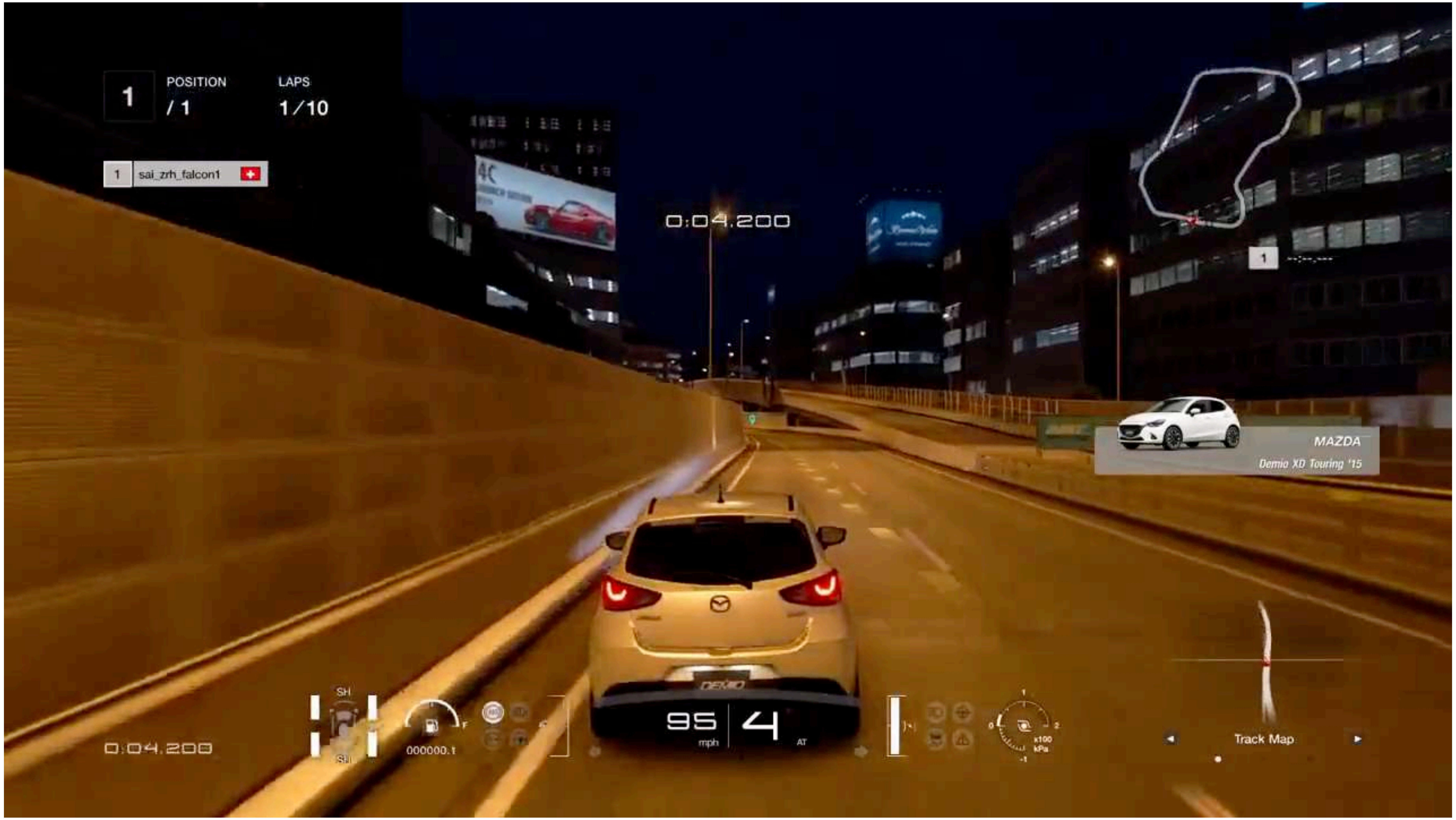
- Need to guarantee safety <u>at all times</u>
- Need to perceive the environment (pedestrian, traffic, etc)
- Need to predict what cars / pedestrian will do

=> usually complex systems with many components / not just RL or OC

Model predictive control

[Gandhi et al. 2020]

Reinforcement learning based



[Fuchs et al. 2021]

The problem… with optimal control

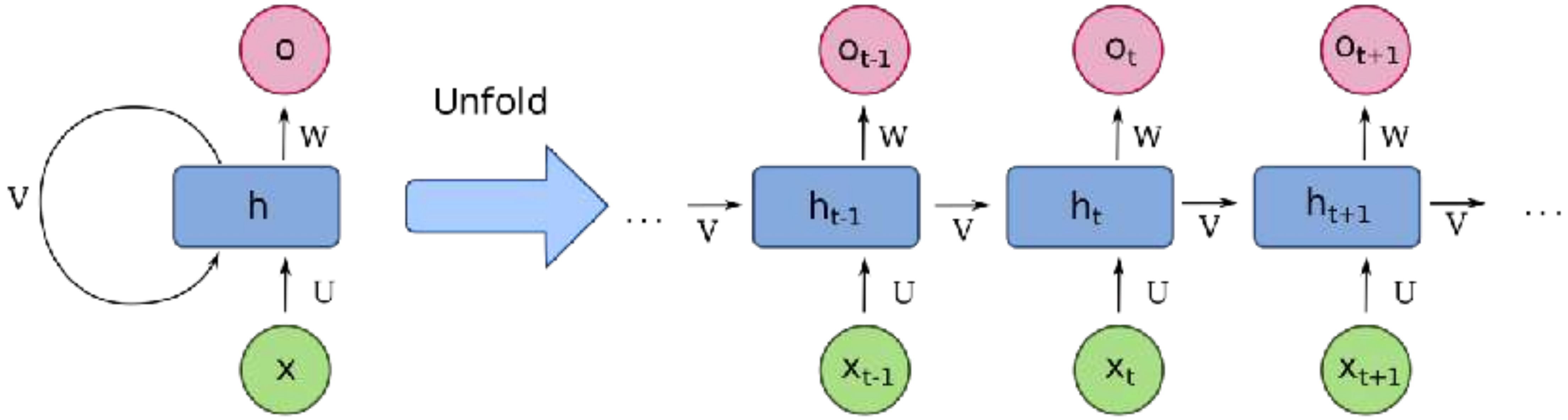
The problem… with states

What if we cannot measure all the states?

=> We go from a Markov Decision Process (MDP) to a Partially Observable Markov Decision Process (POMDP)

=> Reconstruct all the states from the measurement (e.g. Kalman filter)

=> In reinforcement learning, use recurrent neural networks or sequences of past information as input to reconstruct the “real” states



What if we don’t know what the states should be?

What if we want to do something else?

If we learned/computed a policy or value function

=> for a different cost function / task the policy and value functions will be different

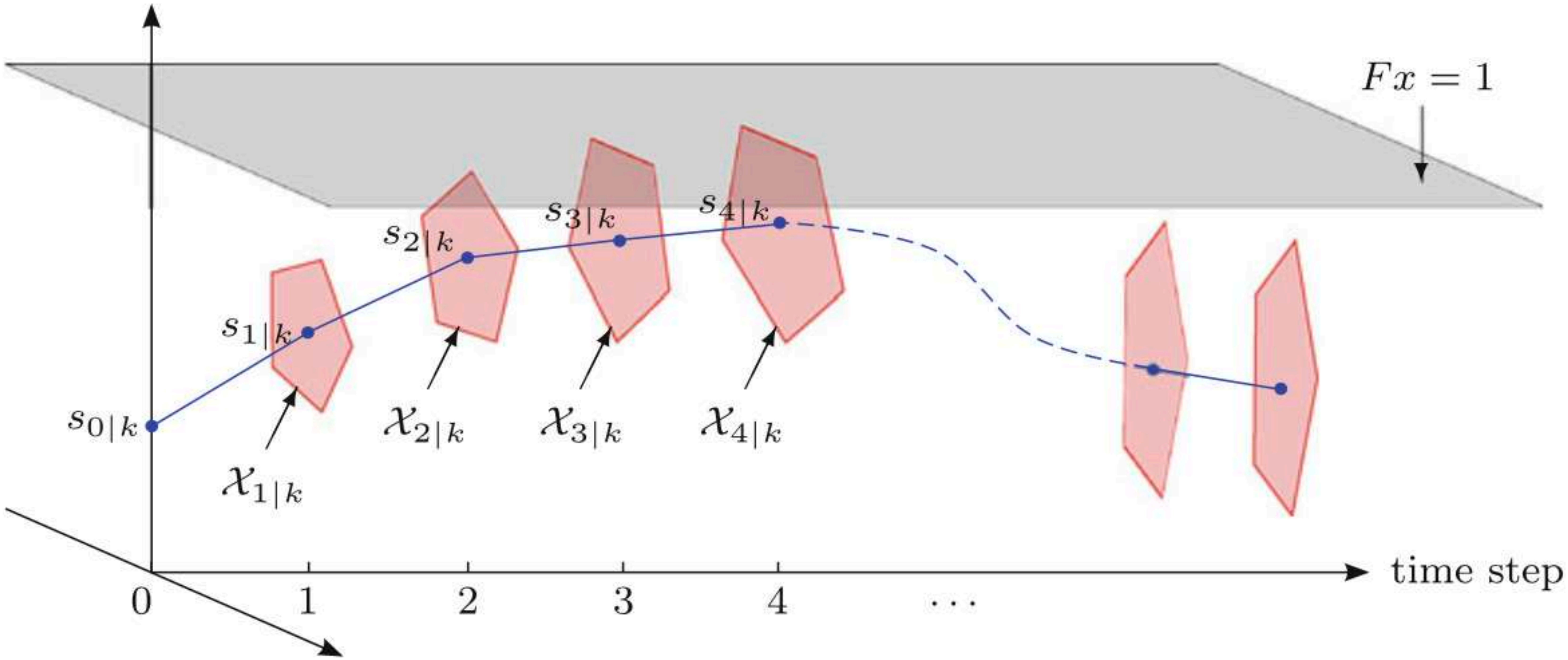
If we have a model: change the cost and we are done

Dealing with uncertainty…

From model predictive control… to stochastic MPC or robust MPC

Robust MPC:

- Dynamic model includes a model of finite disturbances
- Compute a worst case behavior to ensure constraint satisfaction e.g. using tubes



From model predictive control… to stochastic MPC or robust MPC

Stochastic MPC:

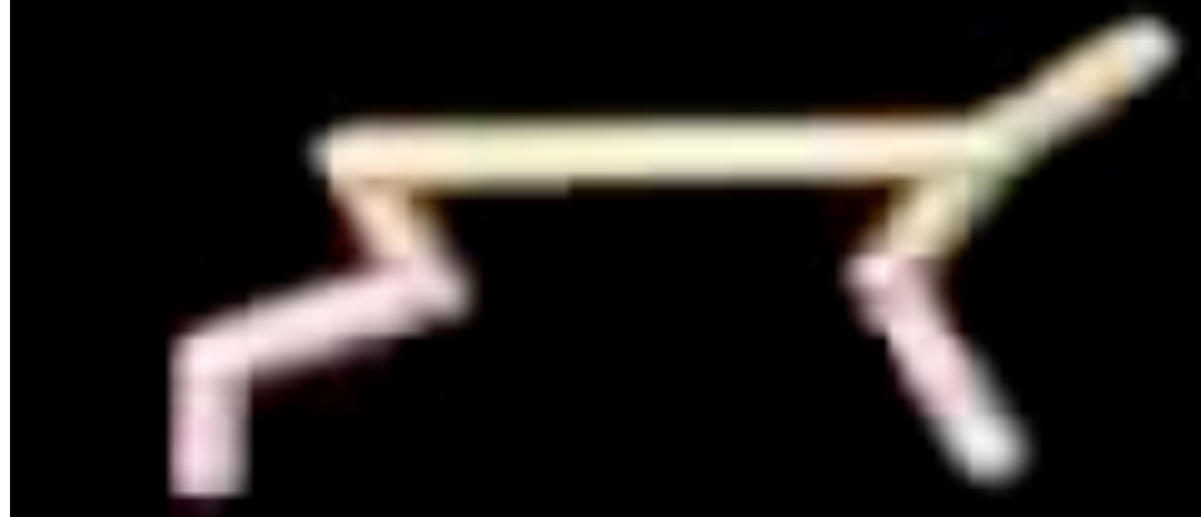
- Dynamic model includes a model of stochastic disturbances
- Compute a behavior to ensure probabilistic constraint satisfaction

$$\mathbf{Gx} \leq 0 \qquad \Rightarrow \qquad \Pr\{\mathbf{Gx} \leq 0\} \leq \epsilon$$

Domain randomization in RL

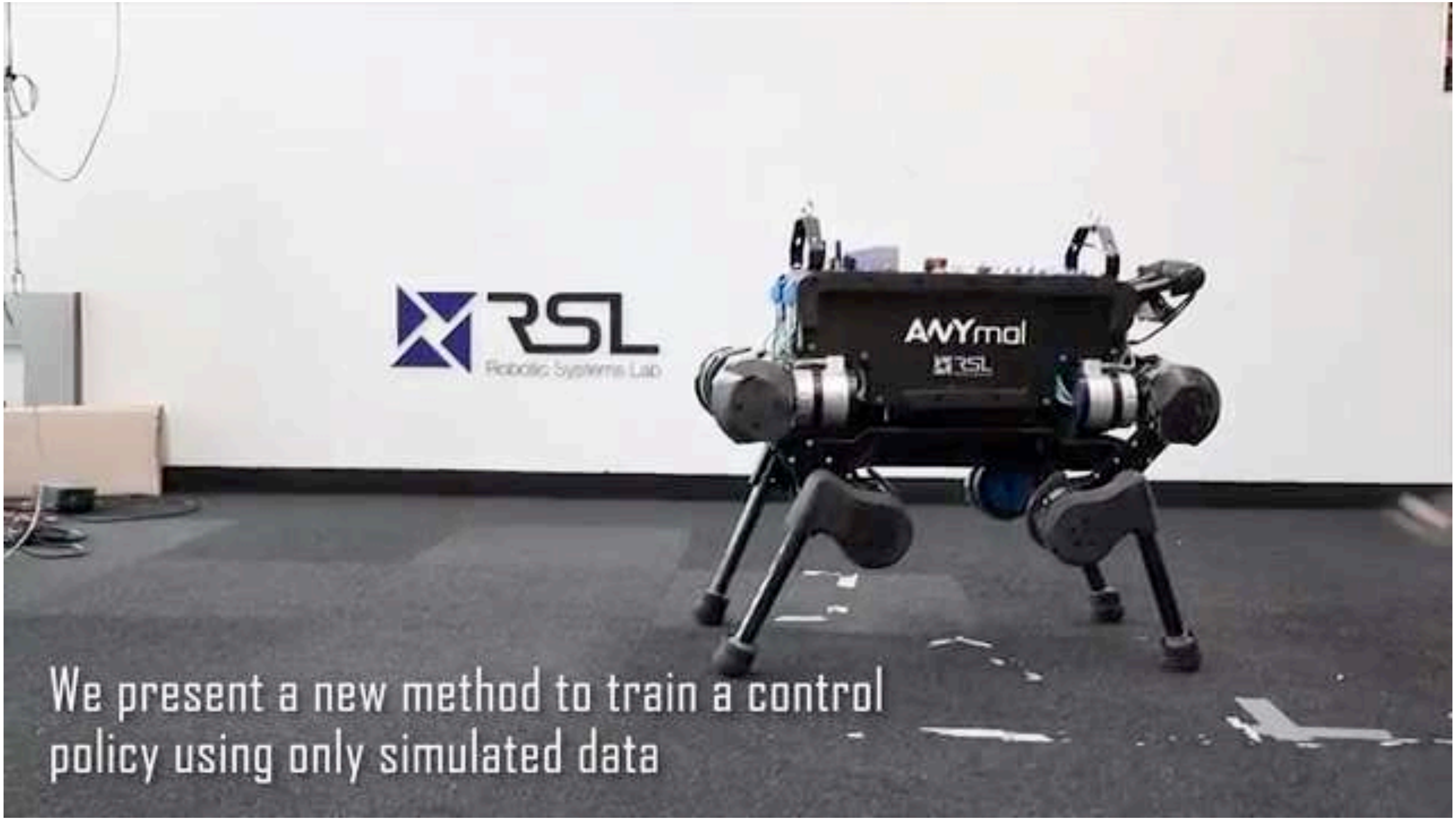
Randomize the simulation to get more robust policies

The problem with learning/optimizing in simulation



Learning skills that transfer to real robots

[Hwangbo et al. 2019]



We present a new method to train a control policy using only simulated data

Sim2Real is the problem of transferring policies learned in simulation onto a real robot - mostly a RL issue / active field of research

MPC-based controller tend to be “easy” to put on robots

[Meduri et al. 2022]

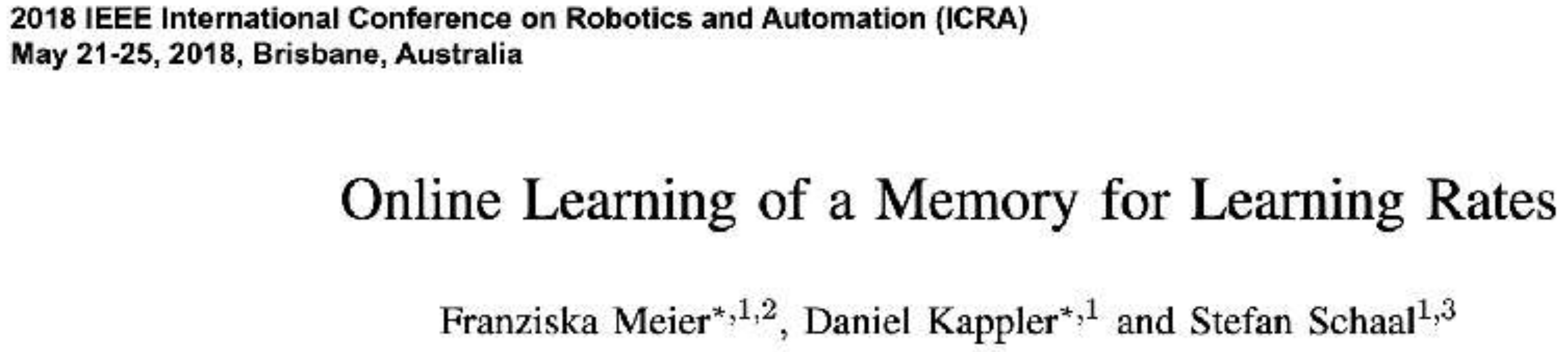
Meta-learning for RL (“Learning to learn”)

Meta-learning for RL (“Learning to learn”)

=> reuse previously learned tasks to learn faster new “similar” tasks

Different methods:

- Learn the parameters of the optimizer

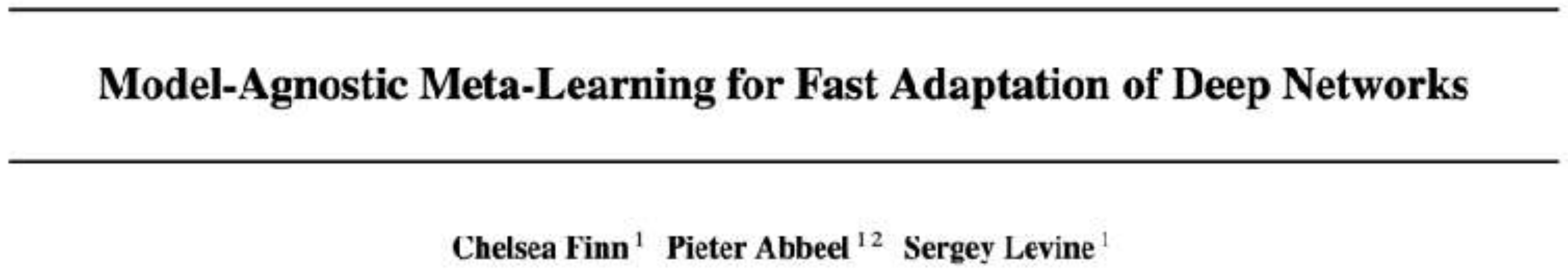


2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Online Learning of a Memory for Learning Rates

Franziska Meier[*,1,2], Daniel Kappler[*,1] and Stefan Schaal[1,3]

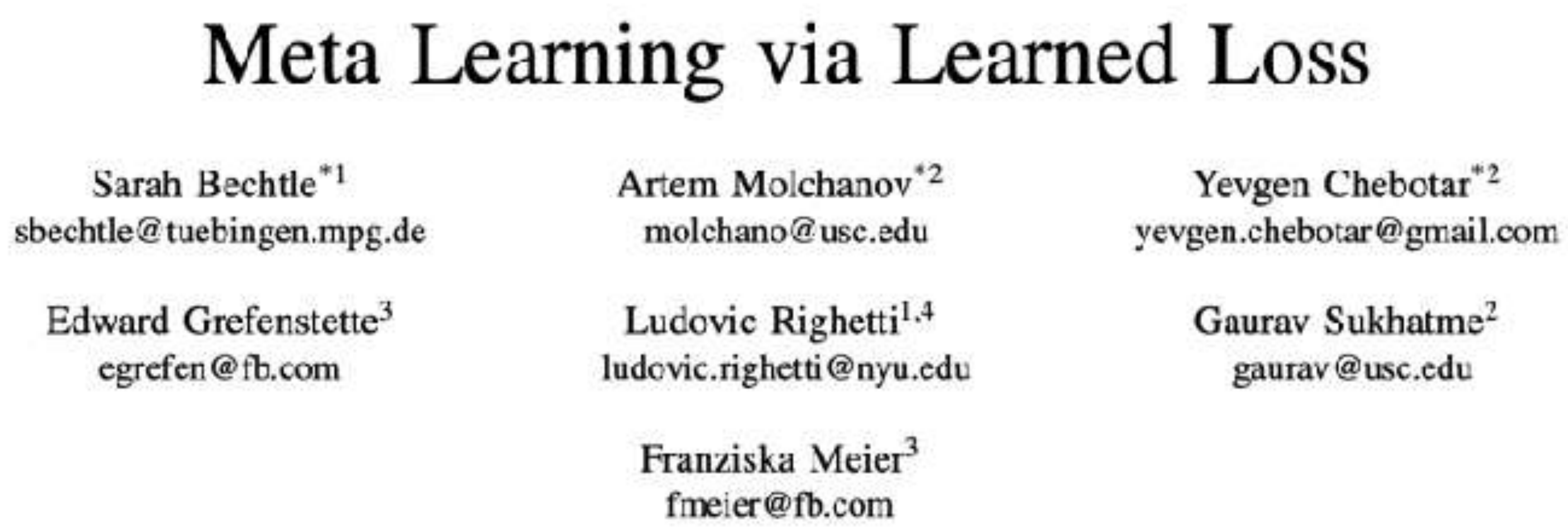
- Learn a model optimized over several tasks



Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn[1] Pieter Abbeel[1 2] Sergey Levine[1]

- Learn a cost function



Meta Learning via Learned Loss

Sarah Bechtle[*1]
sbechtle@tuebingen.mpg.de

Artem Molchanov[*2]
molchano@usc.edu

Yevgen Chebotar[*2]
yevgen.chebotar@gmail.com

Edward Grefenstette[3]
egrefen@fb.com

Ludovic Righetti[1,4]
ludovic.righetti@nyu.edu

Gaurav Sukhatme[2]
gaurav@usc.edu

Franziska Meier[3]
fmeier@fb.com

What about humans/animals?

Do humans have models?

Internal models for motor control and trajectory planning

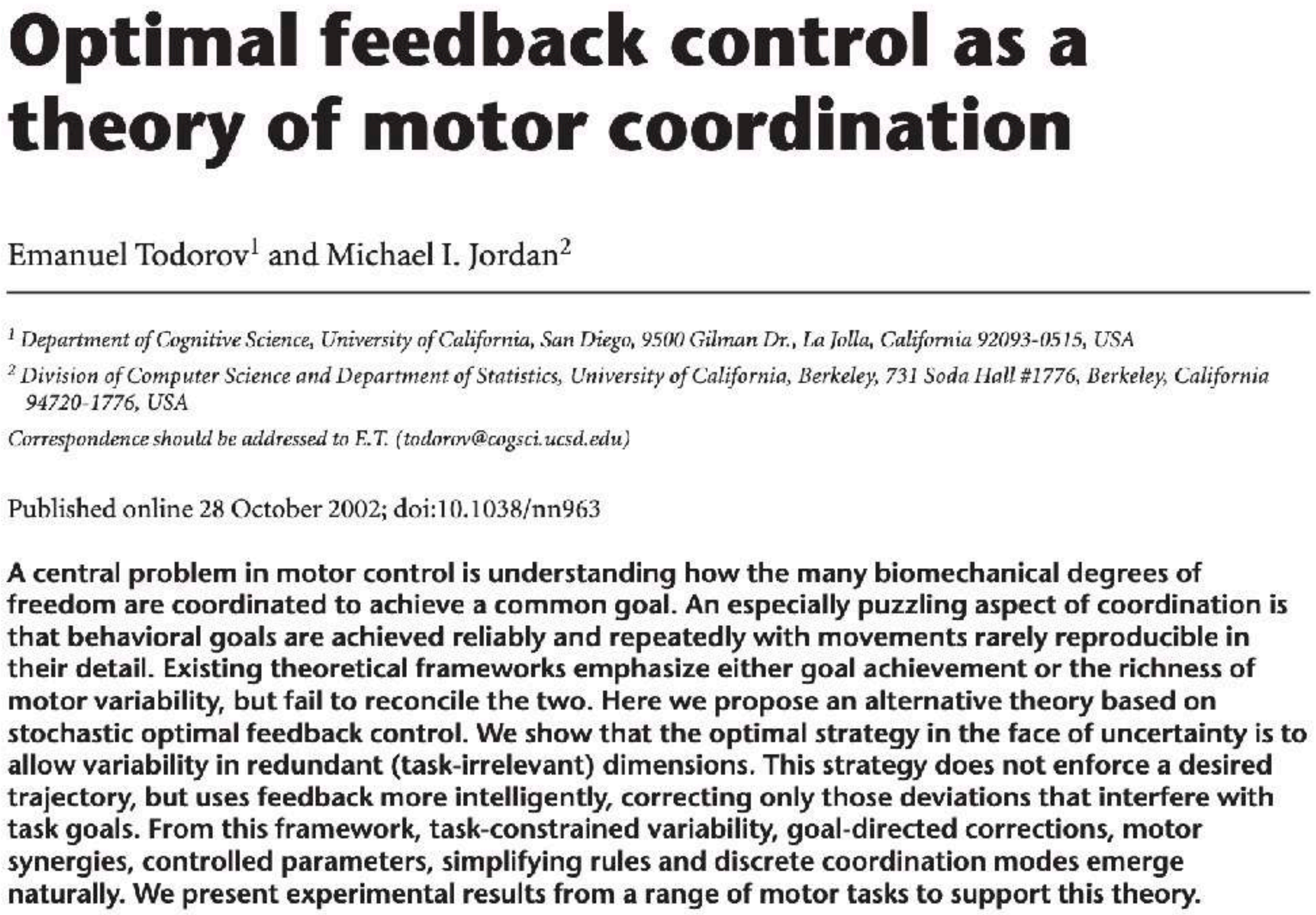
Mitsuo Kawato

A number of internal model concepts are now widespread in neuroscience and cognitive science. These concepts are supported by behavioral, neurophysiological, and imaging data; furthermore, these models have had their structures and functions revealed by such data. In particular, a specific theory on inverse dynamics model learning is directly supported by unit recordings from cerebellar Purkinje cells. Multiple paired forward inverse models describing how diverse objects and environments can be controlled and learned separately have recently been proposed. The 'minimum variance model' is another major recent advance in the computational theory of motor control. This model integrates two furiously disputed approaches on trajectory planning, strongly suggesting that both kinematic and dynamic internal models are utilized in movement planning and control.

Addresses
ATR Human Information Processing Research Laboratories and
Kawato Dynamic Brain Project, ERATO, JST, 2-2 Hikaridai, Seika-cho,
Soraku-gun, Kyoto 619-0288, Japan; e-mail: kawato@hip.atr.co.jp

Current Opinion in Neurobiology 1999, 9:718–727

Do humans behave optimally?



Optimal feedback control as a theory of motor coordination

Emanuel Todorov[1] and Michael I. Jordan[2]

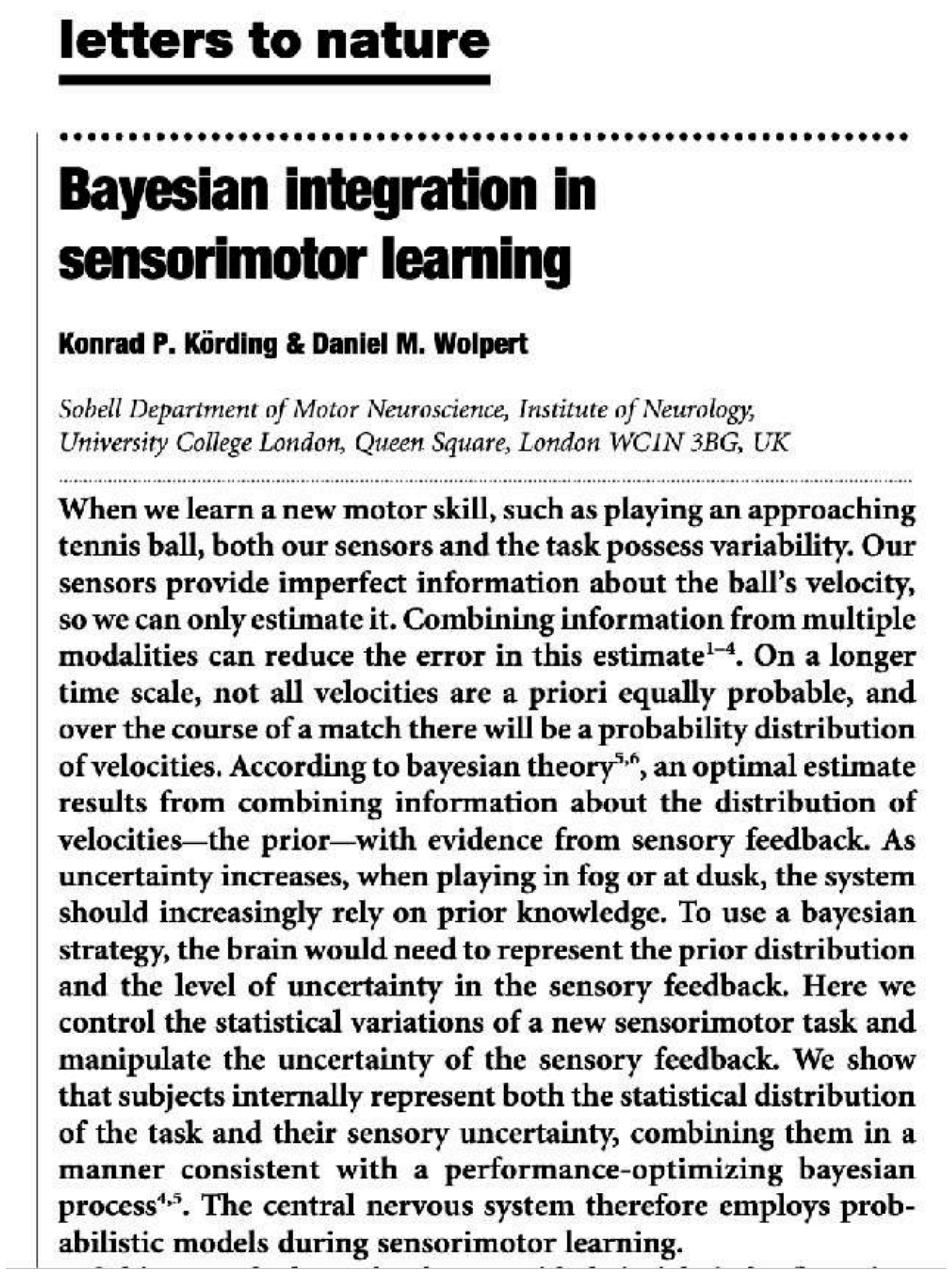
[1] *Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr., La Jolla, California 92093-0515, USA*

[2] *Division of Computer Science and Department of Statistics, University of California, Berkeley, 731 Soda Hall #1776, Berkeley, California 94720-1776, USA*

Correspondence should be addressed to E.T. (todorov@cogsci.ucsd.edu)

Published online 28 October 2002; doi:10.1038/nn963

A central problem in motor control is understanding how the many biomechanical degrees of freedom are coordinated to achieve a common goal. An especially puzzling aspect of coordination is that behavioral goals are achieved reliably and repeatedly with movements rarely reproducible in their detail. Existing theoretical frameworks emphasize either goal achievement or the richness of motor variability, but fail to reconcile the two. Here we propose an alternative theory based on stochastic optimal feedback control. We show that the optimal strategy in the face of uncertainty is to allow variability in redundant (task-irrelevant) dimensions. This strategy does not enforce a desired trajectory, but uses feedback more intelligently, correcting only those deviations that interfere with task goals. From this framework, task-constrained variability, goal-directed corrections, motor synergies, controlled parameters, simplifying rules and discrete coordination modes emerge naturally. We present experimental results from a range of motor tasks to support this theory.

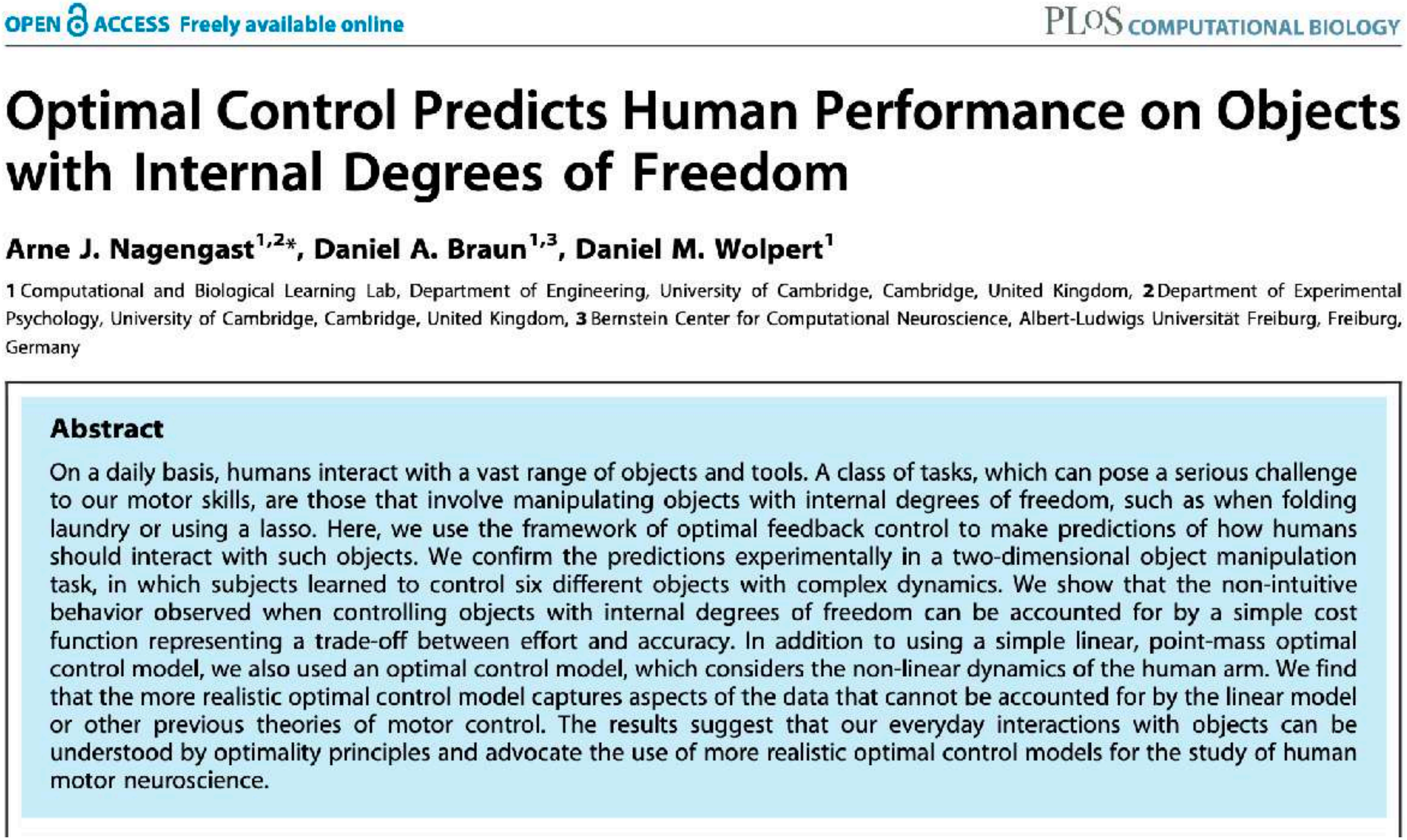


Bayesian integration in sensorimotor learning

Konrad P. Körding & Daniel M. Wolpert

*Sobell Department of Motor Neuroscience, Institute of Neurology,
University College London, Queen Square, London WC1N 3BG, UK*

When we learn a new motor skill, such as playing an approaching tennis ball, both our sensors and the task possess variability. Our sensors provide imperfect information about the ball's velocity, so we can only estimate it. Combining information from multiple modalities can reduce the error in this estimate[1–4]. On a longer time scale, not all velocities are a priori equally probable, and over the course of a match there will be a probability distribution of velocities. According to bayesian theory[5,6], an optimal estimate results from combining information about the distribution of velocities—the prior—with evidence from sensory feedback. As uncertainty increases, when playing in fog or at dusk, the system should increasingly rely on prior knowledge. To use a bayesian strategy, the brain would need to represent the prior distribution and the level of uncertainty in the sensory feedback. Here we control the statistical variations of a new sensorimotor task and manipulate the uncertainty of the sensory feedback. We show that subjects internally represent both the statistical distribution of the task and their sensory uncertainty, combining them in a manner consistent with a performance-optimizing bayesian process[4,5]. The central nervous system therefore employs probabilistic models during sensorimotor learning.



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Optimal Control Predicts Human Performance on Objects with Internal Degrees of Freedom

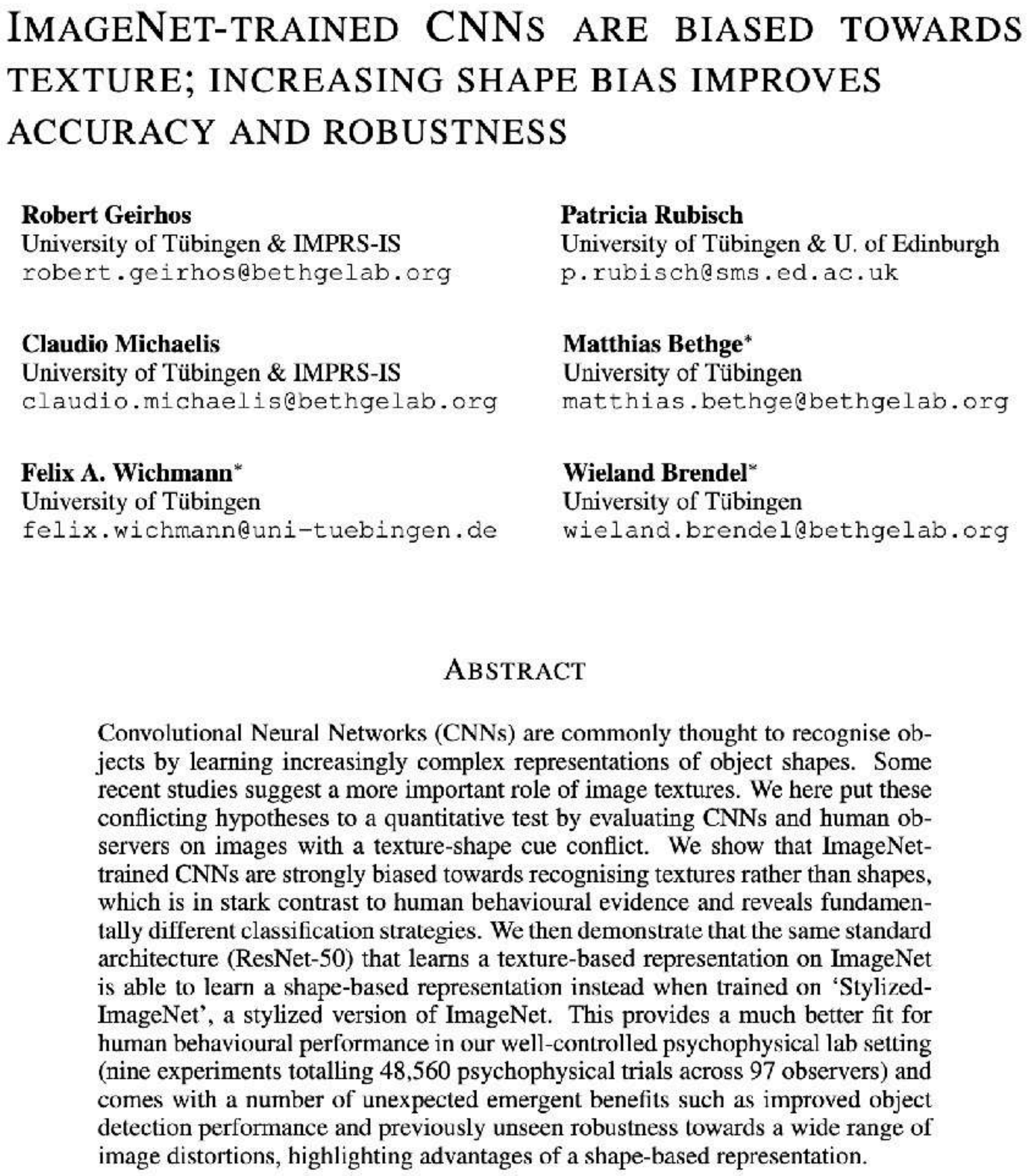
Arne J. Nagengast[1,2]*, Daniel A. Braun[1,3], Daniel M. Wolpert[1]

1 Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Cambridge, United Kingdom, 2 Department of Experimental Psychology, University of Cambridge, Cambridge, United Kingdom, 3 Bernstein Center for Computational Neuroscience, Albert-Ludwigs Universität Freiburg, Freiburg, Germany

Abstract

On a daily basis, humans interact with a vast range of objects and tools. A class of tasks, which can pose a serious challenge to our motor skills, are those that involve manipulating objects with internal degrees of freedom, such as when folding laundry or using a lasso. Here, we use the framework of optimal feedback control to make predictions of how humans should interact with such objects. We confirm the predictions experimentally in a two-dimensional object manipulation task, in which subjects learned to control six different objects with complex dynamics. We show that the non-intuitive behavior observed when controlling objects with internal degrees of freedom can be accounted for by a simple cost function representing a trade-off between effort and accuracy. In addition to using a simple linear, point-mass optimal control model, we also used an optimal control model, which considers the non-linear dynamics of the human arm. We find that the more realistic optimal control model captures aspects of the data that cannot be accounted for by the linear model or other previous theories of motor control. The results suggest that our everyday interactions with objects can be understood by optimality principles and advocate the use of more realistic optimal control models for the study of human motor neuroscience.

How do humans/animals learn like machines?



IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

ABSTRACT

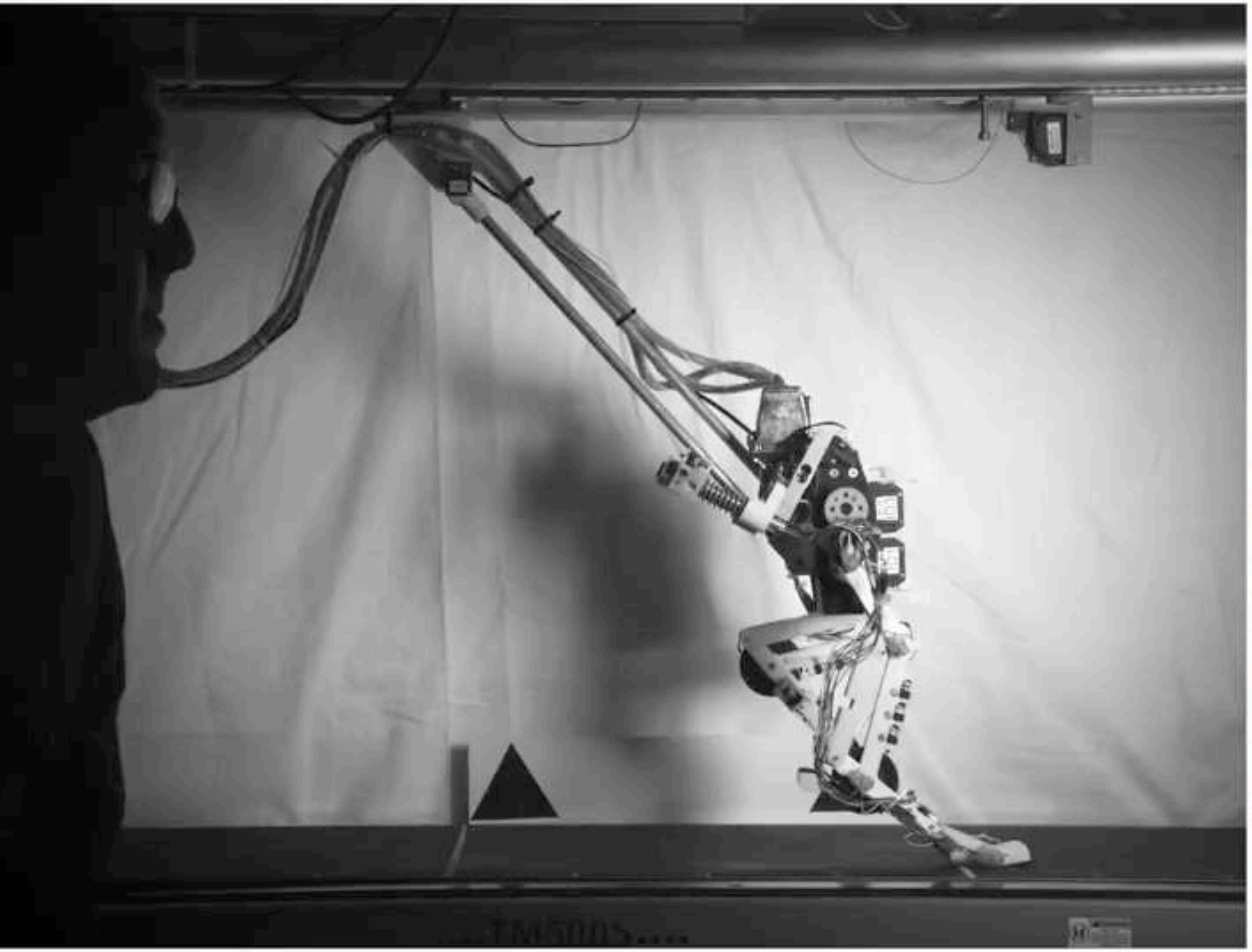
Convolutional Neural Networks (CNNs) are commonly thought to recognise objects by learning increasingly complex representations of object shapes. Some recent studies suggest a more important role of image textures. We here put these conflicting hypotheses to a quantitative test by evaluating CNNs and human observers on images with a texture-shape cue conflict. We show that ImageNet-trained CNNs are strongly biased towards recognising textures rather than shapes, which is in stark contrast to human behavioural evidence and reveals fundamentally different classification strategies. We then demonstrate that the same standard architecture (ResNet-50) that learns a texture-based representation on ImageNet is able to learn a shape-based representation instead when trained on 'Stylized-ImageNet', a stylized version of ImageNet. This provides a much better fit for human behavioural performance in our well-controlled psychophysical lab setting (nine experiments totalling 48,560 psychophysical trials across 97 observers) and comes with a number of unexpected emergent benefits such as improved object detection performance and previously unseen robustness towards a wide range of image distortions, highlighting advantages of a shape-based representation.

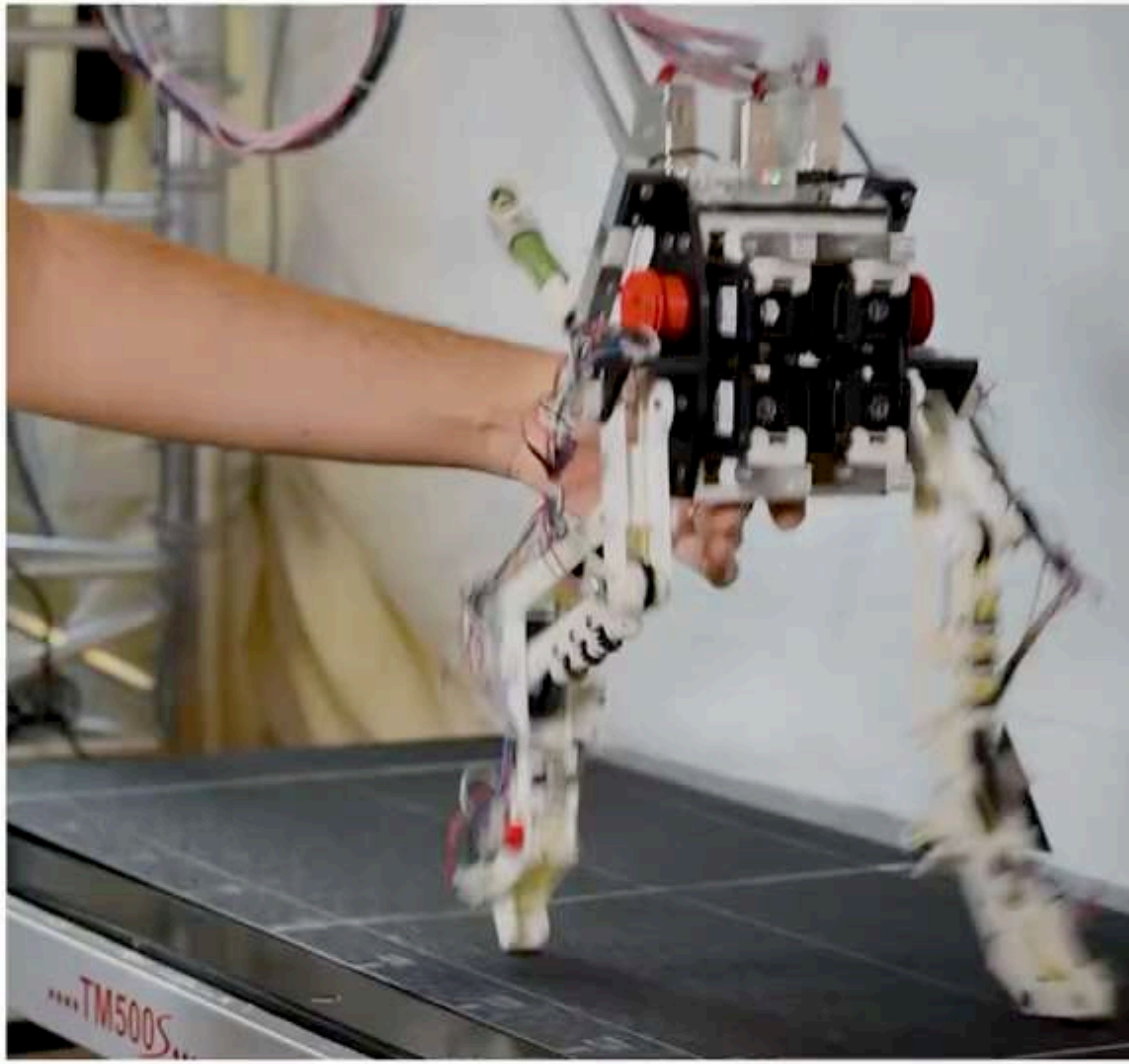
In RL/AI we tend to use an “anthropomorphized language”
we talk about “intelligence”, “learning”, “rewards”, etc

BUT

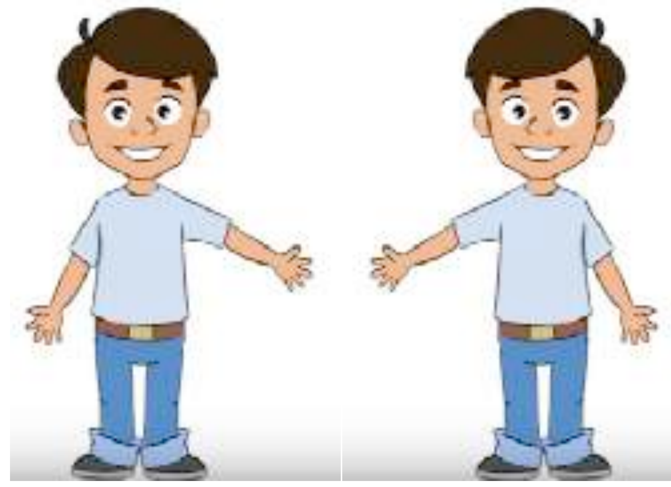
these are just algorithms designed by engineers -
unrelated to what the human brain does

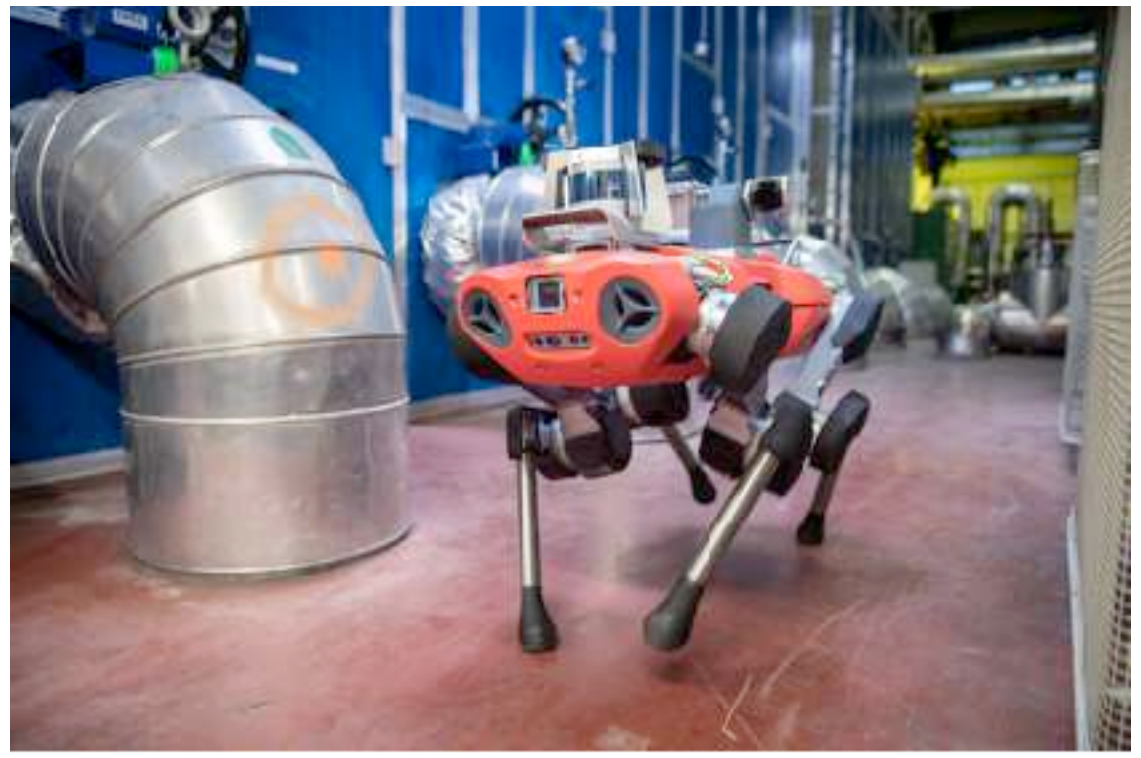
Physical intelligence…





Why?





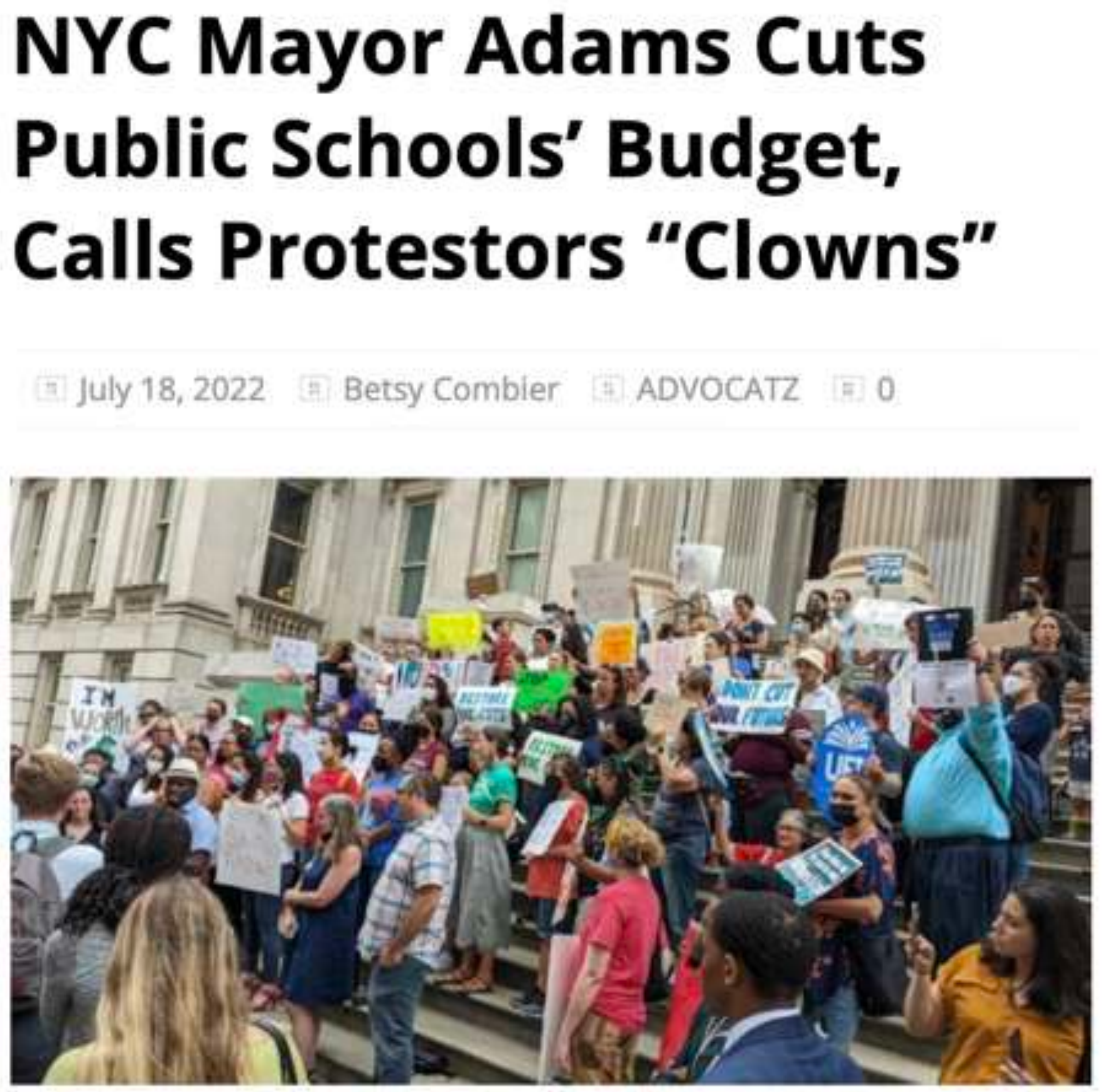
Shall an engineer care about the usage of its robots?

Isn’t it society that needs to decide what it wants?

The origin of the word engineer…

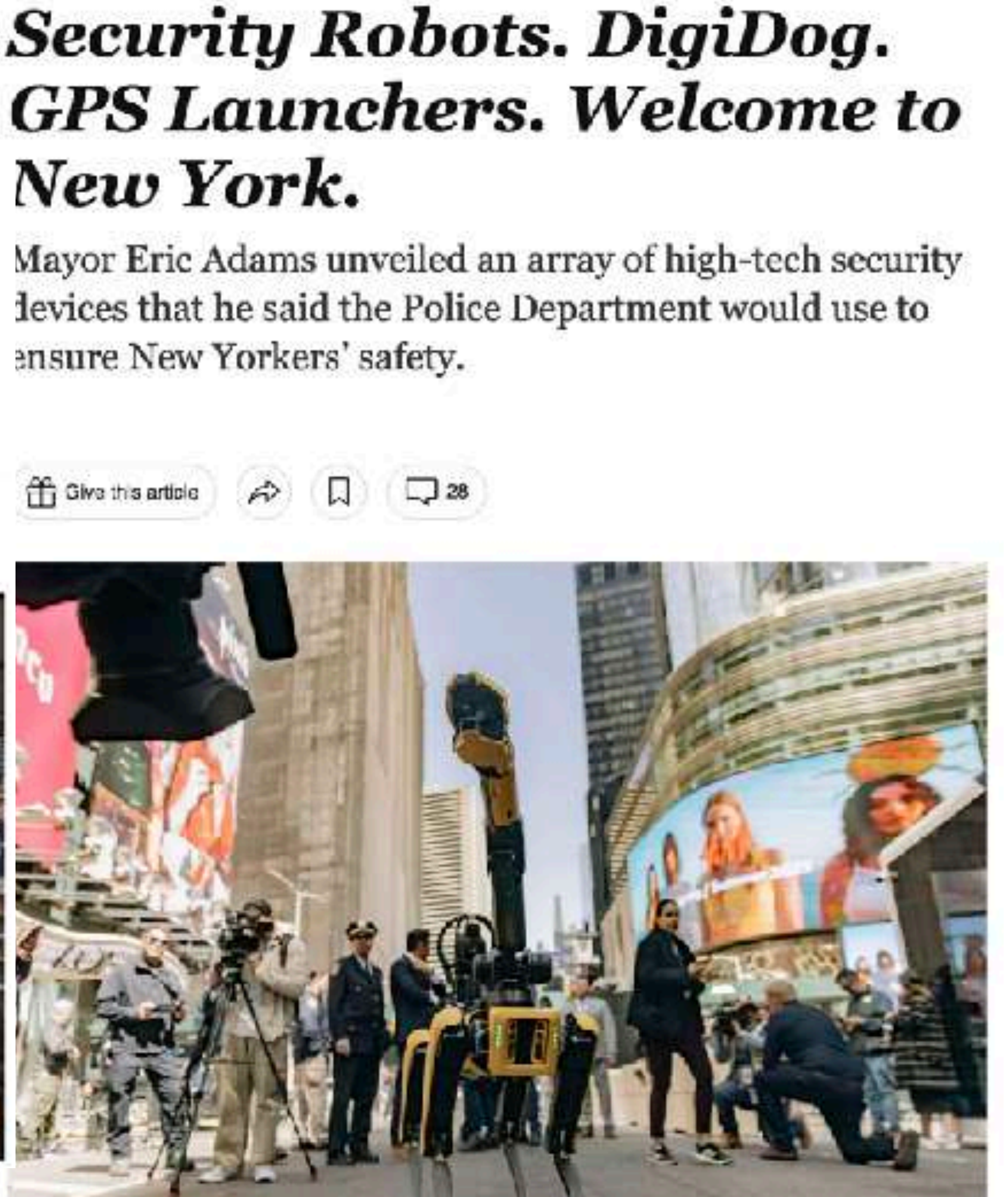
Engineer was coined around the 14th century to describe "a designer and constructor of fortifications and weapons"

Technology is not value free
it shapes possibilities



NYC Mayor Adams Cuts Public Schools' Budget, Calls Protestors "Clowns"

July 18, 2022 Betsy Combier ADVOCATZ 0



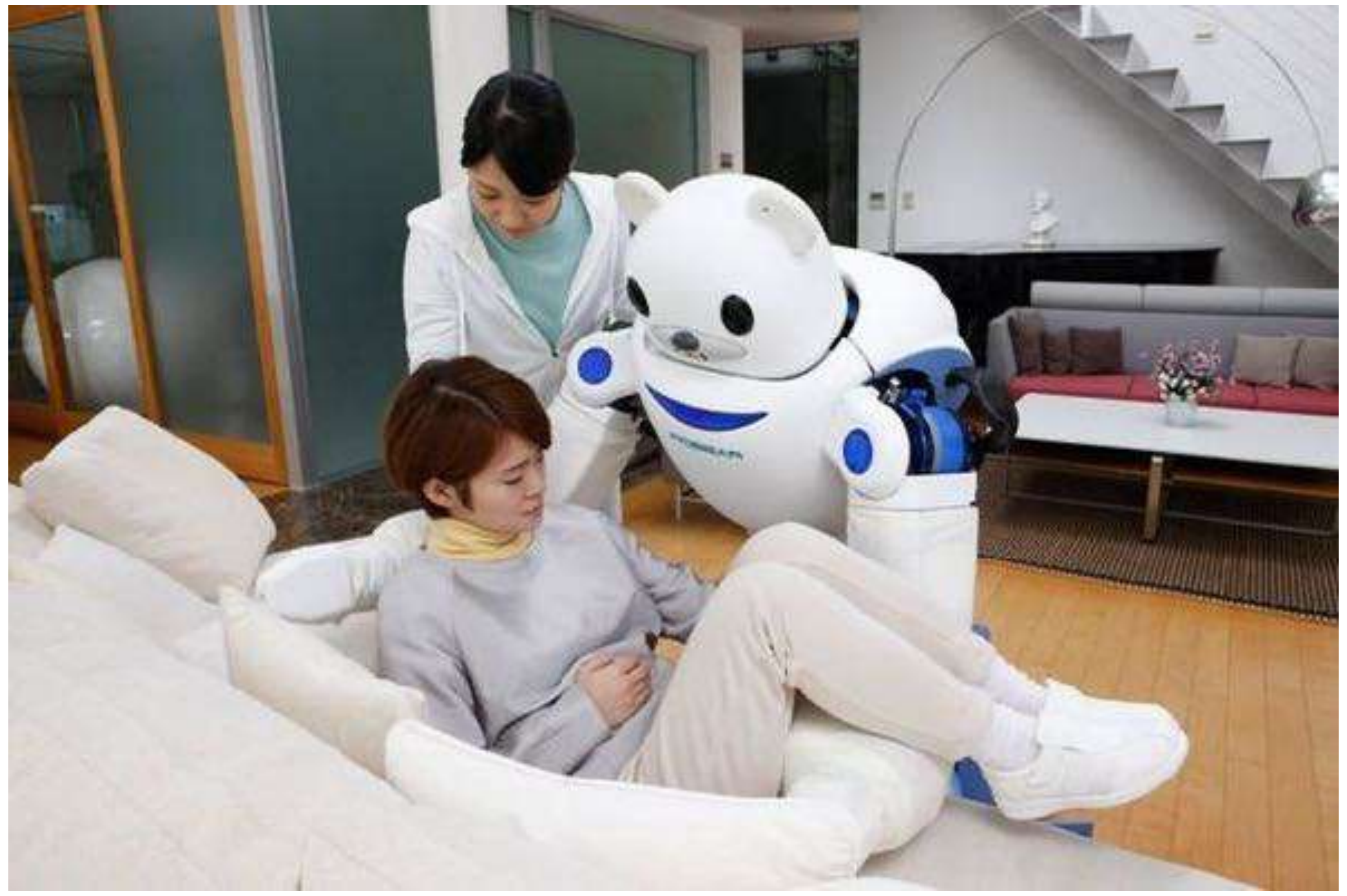
Security Robots. DigiDog. GPS Launchers. Welcome to New York.

Mayor Eric Adams unveiled an array of high-tech security devices that he said the Police Department would use to ensure New Yorkers' safety.

Once I have decided what applications I care about, can I stop thinking about it?

Self-Driving Car

What possible ethical problems with “care” robots?



Algorithms are not well conceived

Data is not neutral

Lack of diversity in the use cases / training dataset

Lack of diversity in the tech industry

G

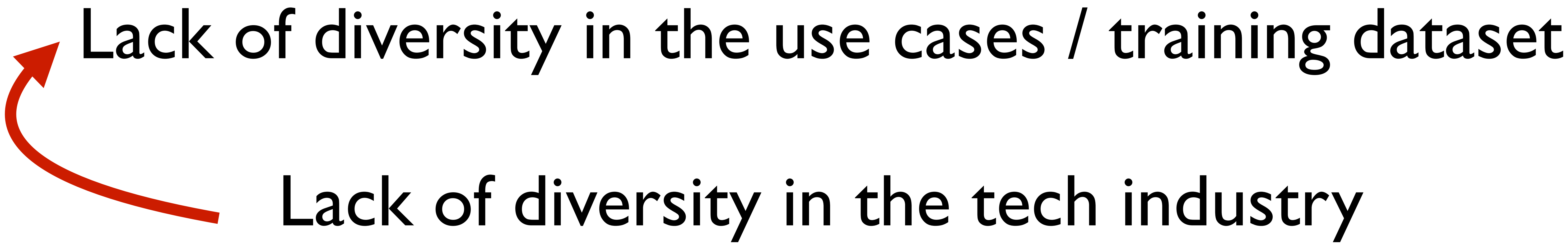
en

Black wo

orkforce

Algorithms are not well conceived

Data is not neutral



Lack of diversity in the use cases / training dataset

Lack of diversity in the tech industry

Intrinsic bias might not be removable
(historical data might be unfair)

Difficult to detect/remove "problematic"
behaviors because they are unknown

Our own unconscious/unknown biases

A crash course on responsible innovation

3 steps to ensure greater acceptability, sustainability and societal desirability of the research and innovation process and its marketable products

1) **baseline data collection and analysis** of the technology or decision that is to be risk assessed

2) **map out what could go wrong** identify and characterize the spectrum of peace and security risks associated with the development, diffusion and (mis)use of the technology

3) **identify means of intervention** identify a) functional requirements for the design, development and diffusion of new research product or services; or b) means for engagement within our outside one workplace/community

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

A crash course on responsible innovation

When to do it? It’s a process! à at key junctures in the life cycle of a technology, or research and innovation process

Who should it? People that design the technology together with relevant decision makers from the university or companies such as project/product managers or people involved in ethical screening and internal compliance programs

Who should be involved? The process can usefully be informed by engagement with external stakeholders who may bring important additional expertise or perspective (e.g. users’ perspective) e.g. through focus-groups, ethical review board etc.

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

Concrete case: autonomous aerial drone with computer vision for fighting wildfires

A company that specializes in software for drones has developed a computer vision (CV) application to help fight wildfires. An autonomous quadcopter drone with the software uses infra-red cameras to detect fire and a sensor suite to analyse the fire pattern (size, location, direction and local weather). The computation is done via an onboard processor and requires no internet connection. The drone can fly autonomously to pre-programmed areas and loiter over an area as big as 2 square kilometres for 45 min. The drone relays the data to a remote-control device, which visualizes the data on a topical map.

The company is now looking into two additional use cases:

- Search and rescue: The CV software would detect people that need rescue during wildfire.
- Poacher identification: The CV software would detect people that may be potential poachers in protected natural reserves.

With these uses, the company is seeking to broaden its customer base from fire-responders to environmental protection professionals (like park rangers).

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

1. Baseline data collection

1. What is the technology? What needs/problems does it solve?
2. What is the level of maturity of the technology?
3. Who are the intended users?
4. How will it be diffused? (how will be people get access to it?)
5. Is the technology already regulated?

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

2. What could go wrong?

1. What could be unintended negative consequences of the intended use?
2. What could be form of misuse? By who?
3. How difficult would it be to misuse the technology?
4. When and where would the issues materialize?
5. Who would be affected? (people, states, companies, etc)

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

3. Identifying means of intervention

1. Are the risks identified so significant and intolerable that the research or development of the technology should not be pursued?
2. Can the risks be managed at least partially at the technical level?
3. Does the problem require limiting the diffusing of the technology?
4. Do I have enough expertise to understand the problem / deploy risk mitigation measures? If not, who should I turn to?
5. Should the problem be addressed by the robotics community as a whole?

Material courtesy of C. Ovink (UNODA) and V. Boulanin (SIPRI)

NYU Tandon Center for Responsible AI
NYU Alliance for Public Interest Technology

IEEE Global initiative on Ethics of Autonomous and Intelligent Systems
Key ressources: 1) Ethically aligned design report, 2) IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being; 3) IEEE standard model process for addressing ethical concerns during system design; 4) IEEE ethics certification program for autonomous and intelligent systems; 5) A call for action for business using AI

The High-Level Expert Group on Artificial Intelligence (AI HLEG)
Key ressources: Ethics Guidelines for Trustworthy Artificial Intelligence; The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment

Partnership on AI => private led
The Global Partnership on AI => State led
Montreal Ethics Institute => NGO

Moving forward…

Technical responses…

=> testing/certification (formal methods to guarantee certain performance/safety, etc)

=> algorithmic fairness (formal methods to remove bias)

Fundamental limits on algorithmic guarantees

[Kleinberg et al. 2016]

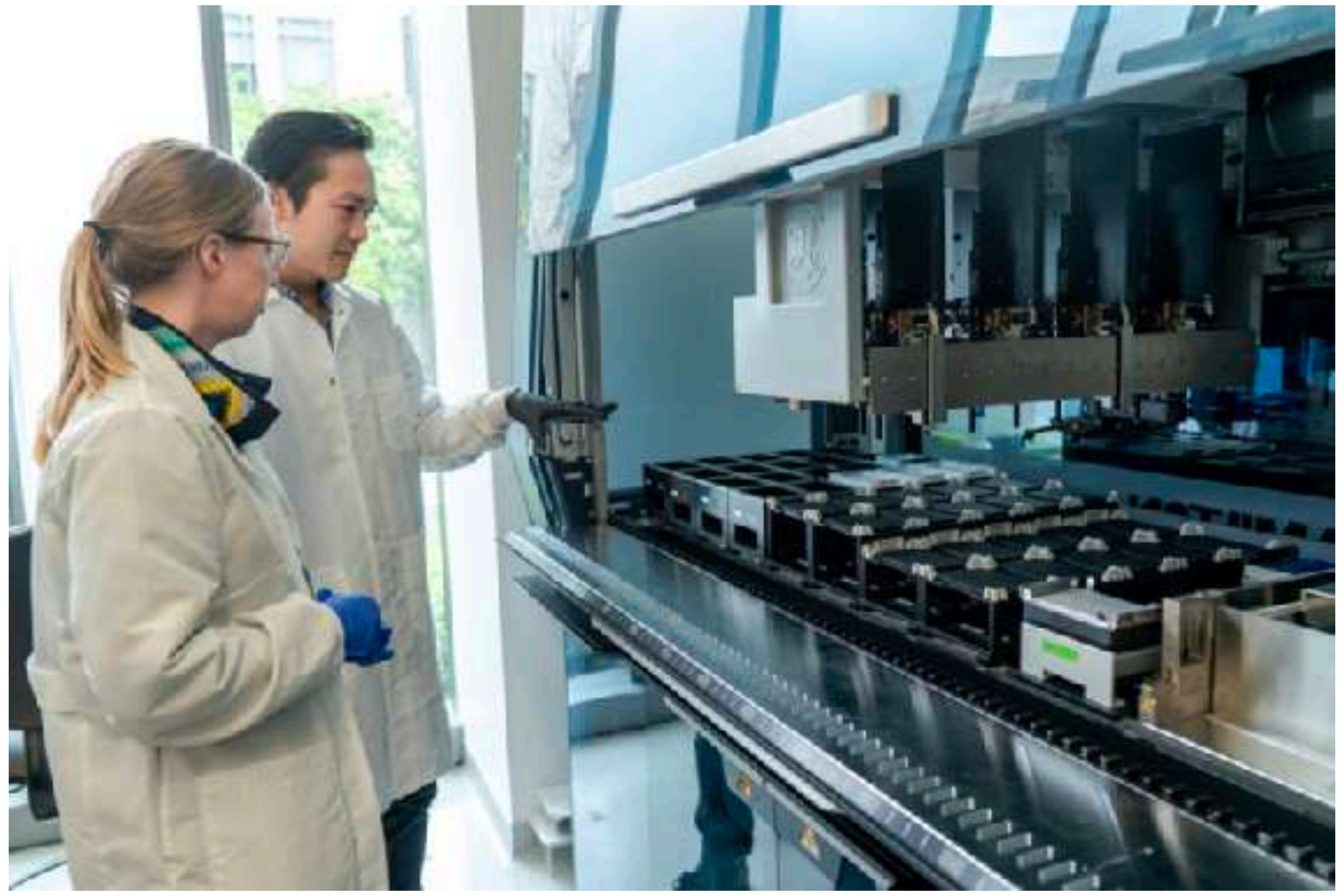
Holistic approaches taking into account social aspects…

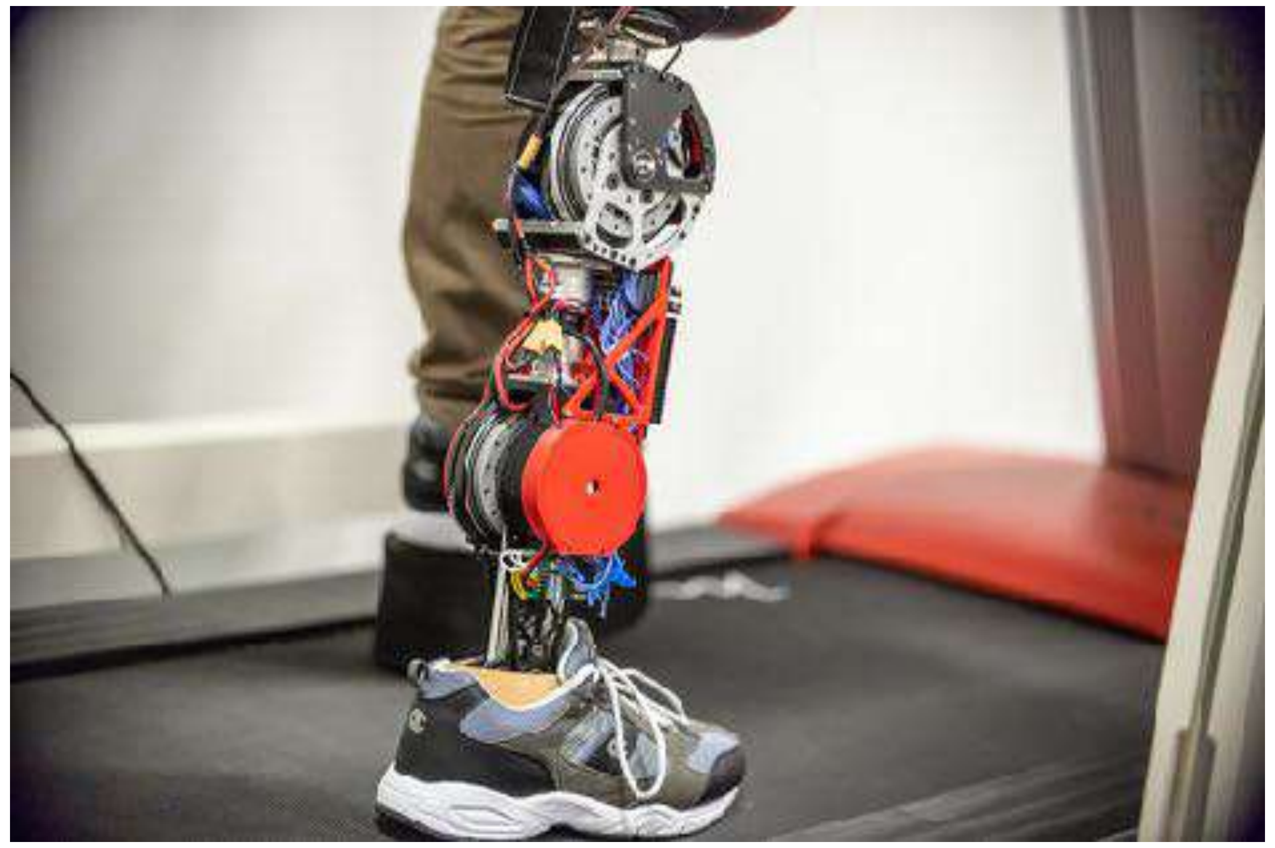
=> engineering cannot be done in isolation anymore

=> work with social scientists, ethicists, etc

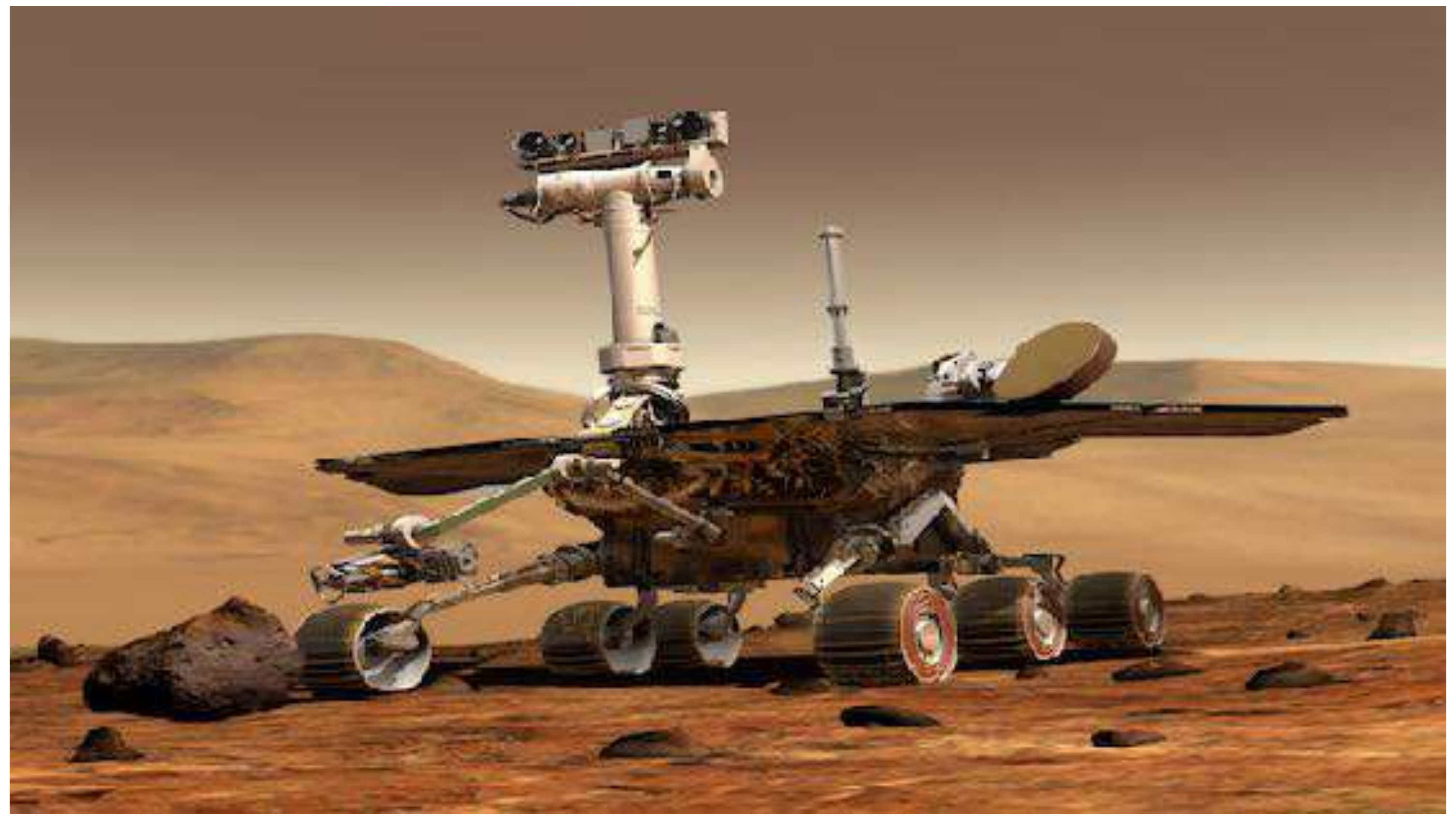
What do we do with our robots?

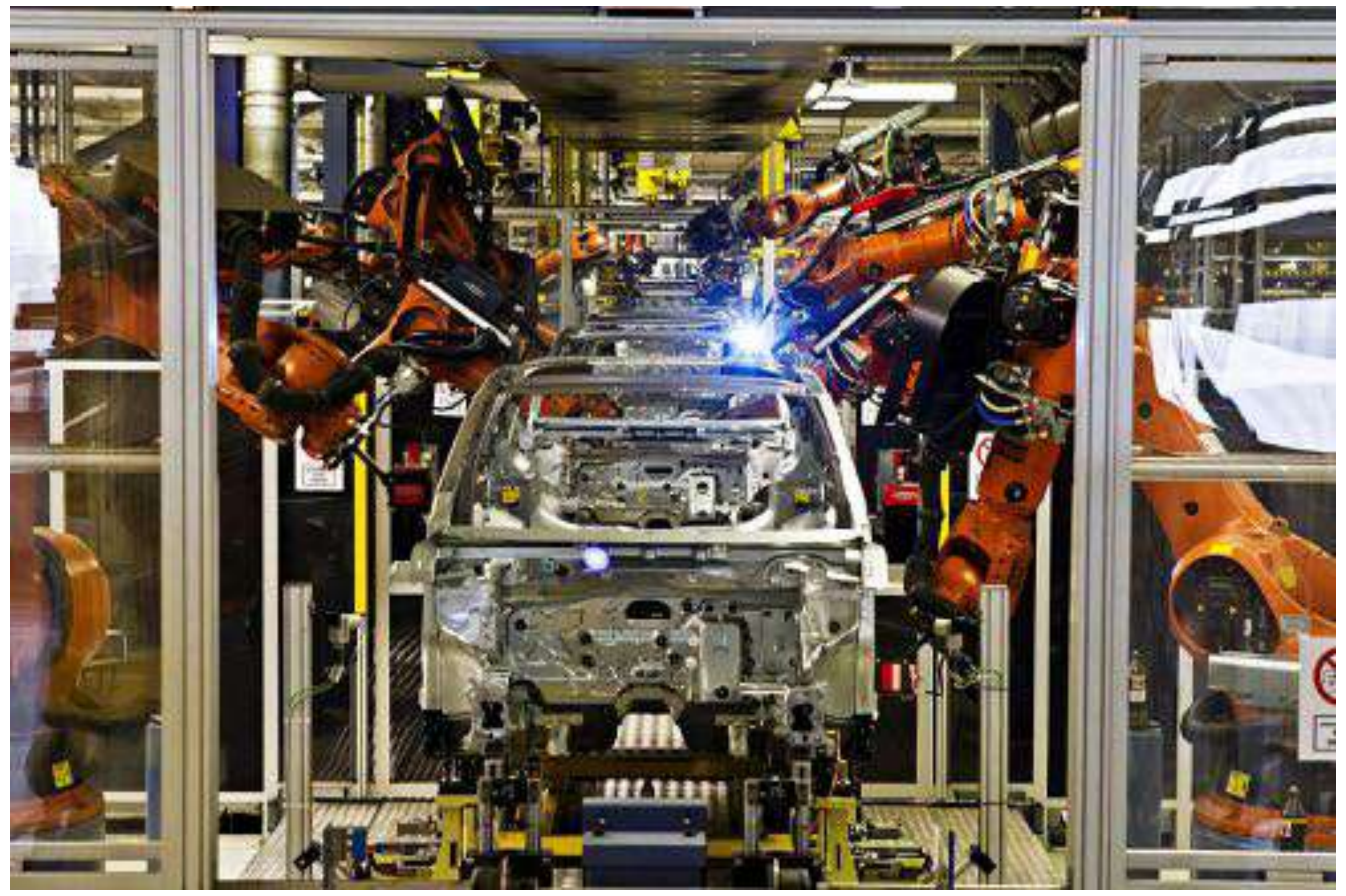
What do we do with our robots?





What do we do with our robots?





Enjoy a well-deserved break!

Congratulations to everyone graduating this semester!