Towards Generalizable Autonomy
Structure in Reinforcement Learning for Robotics

Animesh Garg
Generalizable Autonomy: Computer Vision & Language

Structured Models + Data + Compute → Performance

Open Images Dataset

SQuAD
The Stanford Question Answering Dataset

Common Crawl

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Rajpurkar &amp; Jia et al. '18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE-Net (ensemble) RICOH_SRCB_DML</td>
<td>90.939</td>
<td>93.214</td>
</tr>
</tbody>
</table>
Generalizable Autonomy: Computer Vision & Language

Ingredients of Modern Machine Learning & Applications

Large Structured Models
- Over-parameterized
- Structured Biases

IID Data & Datasets
- Concise problem Definition
- IID Data, easier to label

Distributed Deployment
- Large Scale Compute
- Distributed Deployment

Visual Perception
Natural Language
Passive Offline Decisions
Intelligent Robotics
Embodied
Generalizable Autonomy: Duality of Discovery & Bias

Domain Expertise

Data Driven

One problem, One solution!

Structured Environments

Input: Image

Output: Action

Independent Processes

\[\mathcal{O} \rightarrow \text{Perception} \rightarrow \text{State} \rightarrow \text{Planning} \rightarrow \text{Plan} \rightarrow \text{Control} \rightarrow \mathcal{A}\]
Generalizable Autonomy: Duality of Discovery & Bias

Just add data…

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

\[ \mathcal{O} \xrightarrow{\text{End-to-End Policy } \pi} \mathcal{A} \]

Input: Image

Output: Action
Generalizable Autonomy: Duality of Discovery & Bias

Domain Expertise

One problem, One solution!

✗ Need for experts
✗ Limited applicability
✗ Perf vs Flexibility

Just add data…

✗ Computational sustainability
✗ Data accessibility
✗ Out-of-distribution errors

Neither achieves Generality at Scale
...make the **inductive** leap necessary to classify instances beyond observed...

...other sources of information, or **biases** for choosing one generalization over the other...

No Generalization without Structure!
Generalizable Autonomy: Duality of Discovery & Bias

Generalizable Autonomy

Structure + Data

- Domain knowledge,
- Inductive bias,
- Symmetries,
- Priors
- ...
- Online & Offline,
- Simulation & Real,
- Labelled & self-supervised
- Human in the loop
- ...

Domain Expertise

Data Driven
Structured Representations: Vision & Language

Insight: Structure makes learning possible

- Relational Inductive Biases
  - Independence
  - Locality
  - Specified
  - Sequentiality

- Attention Mechanisms
  - Unsupervised Pretraining
  - Supervised Pretraining

- Transfer Learning

- Contrastive Methods
  - Align/Repel

Explicit Capacity-Focused

Implicit Task-Focused

Sam Finlayson, 2020
Structure in Reinforcement Learning

Markov Decision Process
\[ \mathcal{M} = \langle S, A, P(\cdot;\cdot), R(\cdot;\cdot), H \rangle \]

Goal: Find Optimal Policy
\[ \pi^*: S \to A \]
Structure in Skill Learning: or the lack of it

Slow and Narrow:
- Specific tasks (Grasp/Stack)
- Often Supervised and rigid!

Learning Fast: Elephants Learning to use trunk

Learning Broad: Human infant learns to interact

[Kalashnikov et al. (2018), Cabi et al. (2019)]
Structure for Reinforcement Learning

Markov Decision Process

\[ \mathcal{M} = \langle S, A, P(\cdot;\cdot), R(\cdot;\cdot), H \rangle \]

Goal: Find Optimal Policy

\[ \pi^*: S \rightarrow A \]

Which structured biases enable generalizable autonomy in decision-making?
Anatomy of an RL agent

- Observation ($O$) to Observation
- Abstract State ($S$)
- Previous Action
- Agent
- Transition Model
- Reactive Policy
- Value Function
- Learning
- Planning
- Abstract Action ($A$) to Action Decoder
- Control ($U$)
- Environment
- Reward

Update Rule
Structure for RL: Off-policy RL

\[
\min_Q \mathbb{E}_{a \sim \pi_{Behavior}} [\mathcal{L}(Q(s, a) - Q_{Target})]
\]

\[Q(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{a' \sim \pi_{new}} [Q(s', a')]\]

\[
\pi_{new} = \begin{cases} 
1 - \epsilon, & \text{if } a_t = \arg\max_a Q(s, a) \\
\frac{\epsilon}{|\mathcal{A}| - 1} & \text{if } a_t \neq \arg\max_a Q(s, a)
\end{cases}
\]
Structure for RL: Off-policy RL

- **Replay Buffer**
  - Sample Minibatch \( B \)
  - Policy Rollout (Generate Samples with Interaction)

**Update Rule**

- \( \min_q \mathbb{E}_{a \sim \pi_{Behavior}} [L(Q(s, a) - Q_{Target})] \)
- \( Q(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{a' \sim \pi_{new}} [Q(s', a')] \)

**Policy Improvement**

- \( \pi_{new} = \begin{cases} a_t & \text{max} (Q(s, a)) \\ \epsilon/(|\mathcal{A}| - 1) & \text{otherwise} \end{cases} \)

Where is this batch of data coming from? And can we use it more efficiently?
Data in Robotics

**Manipulation**
- Mason & Salisbury 1985
- Srivinasa et al 2010
- Berenson 2013
- Odhner1 et al 2014
- Chavan-Dafle et al 2014
- Yamaguchi, et. al, 2015
- Li , Allen et al. 2015
- Li, Allen et al. 2015
- Yahya et al, 2016
- Schenck et al. 2017
- Mar et al. 2017
- Laskey et al 2017
- Quispe et al 2018
- ... (additional references)

**Grasping**
- Mishra et al 1987
- Ferrari & Canny, 1992
- Ciocarlie & Allen, 2009
- Dogar & Srivinasa, 2011
- Rodríguez et al. 2012
- Bohg et al 2014
- Pinto & Gupta, 2016
- Levine et al 2016
- Mahler et al 2017
- Jang et al 2017
- Viereck et al 2017
- ... (additional references)

**Imitation**
- Abbeel et al, 2004
- Ratliff et al 2006,
- Ziebart et al, 2009
- Argall et al, 2009,
- Boularias et al., 2011
- Montfort et al 2015,
- Wulfmeier et al 2015,
- Krishnan et al 2017
- Finn et al. 2017
- Vecerik et al. 2017
- Rajeswaran et al 2018
- Zhu et al 2018
- Ravichandar et al 2020...
Data in Robotics

Robot Data is Expensive!

- Short-Horizon skills
- Skill Specific learning

Manipulation
- Mason & Salisbury 1985
- Srinivasa et al 2010
- Berenson 2013
- Odhner et al 2014
- Chavan-Dafle et al 2014
- Yamaguchi et al 2015
...

Grasping
- Li, Allen et al. 2015
- Yahya et al 2016
- Schenck et al. 2017
- Mar et al 2017
- Laskey et al 2017
- Quispe et al 2018
...
- Mishra et al 1987
- Ferrari & Canny, 1992
- Ciocarlie & Allen, 2009
- Dogar & Srinivasa, 2011
- Rodriguez et al 2012
- Bohg et al 2014
- Pinto & Gupta, 2016
- Levine et al 2016
- Mahler et al 2017
- Jang et al 2017
- Viereck et al 2017
...

Imitation
- Abbeel et al, 2004
- Ratliff et al 2006,
- Ziebart et al, 2009
- Argall et al, 2009,
- Boularias et al., 2011
- Montfort et al 2015,
- Wulfmeier et al 2015,
- Krishnan et al 2017
- Finn et al. 2017
- Vecerik et al. 2017
- Rajeswaran et al 2018
- Zhu et al 2018
- Ravichandar et al 2020...

- Platform dependent data
- Scaling to other skills?
- Small datasets (minutes)
- Low diversity?

✗ Short-Horizon skills
✗ Skill Specific learning
✗ Platform dependent data
✗ Scaling to other skills?
✗ Small datasets (minutes)
✗ Low diversity?
Data Augmentation in RL
How to do this Algorithmically

• Substantial performance boosts!

Goal relabeling (e.g., HER)

Visual Input relabeling (e.g., RAD)

Data Augmentation in RL
Unified View

**Insight**
Exploit the independence of the causal mechanisms guiding transition

**Counterfactual reasoning** to generate new, causally valid (counterfactual) data!

Goal is independent of State/Action Dynamics

Visual characteristics (e.g., crop) are independent of physical dynamics

Given two independent mechanisms, Relabel one (conditional independence!)

Data Augmentation in RL
Do more with the same data

Scenario 1

Scenario 2

Which of the following is possible (only based on observed data)

$O_1 + B_2$

Independent
Compositional Generalization

$B_1 + O_2$

Not-Independent (!)
Hence need evidence of possibility
Data Augmentation in RL
Do more with the same data

Left Arm Pick and Place
Right Arm Pick and Place

Which of the following is possible (only based on observed data)

✓ Independent Compositional Generalization
✗ Not-Independent (!) Hence need evidence of possibility
Counterfactual Data Augmentation

Counterfactual reasoning to generate new, causally valid (counterfactual) data!

- Model-Free relabelling
- But Causal Independence is not Global

Given two independent mechanisms, Relabel one (conditional independence!)

- For the most part, entities behave independently, and we can use CoDA
- But entities are not always independent, so this can also produce nonsense
Counterfactual Data Augmentation

Local Causal Model

Global Model

Local Model

Structural Causal Model (SCM) that marginalizes across all possible transitions

Local Causal Model (LCM) that behaves like the global SCM in local subspace $\mathcal{L}$

Where do local models come from?
Counterfactual Data Augmentation
Learning Local Causal Model

- Input: 2 balls, each with 4 features: $[x, y, \dot{x}, \dot{y}]$
  - $[[1.23, -0.73, 1.31, 1.07],
    [-0.6, 2.51, -1.51, -0.89]]$
- Output: Adjacency matrix $M$ of the causal graph (between $x_t$ and $x_{t+1}$)
- (intuition) $M$: the input-output Jacobian is non-zero
CoDA: Goal-Conditioned (Online) RL

Fetch-Push-v1
state space: [Robot and 1 object]

CoDA Heuristic for independence:
Obj \perp\ Gripper \mid if |x_g - x_o| \geq 10 \text{ cm}

Fetch-Slide2
state space: [Robot and 2 objects]
Harder task (30x more samples)!
Structure for RL: Off-policy RL

Replay Buffer → Sample Minibatch \( \mathcal{B} \)

**Update Rule**

\[
\min_q E_{a \sim \pi_{behavior}} [\mathcal{L}(Q(s, a) - Q_{Target})]
\]

\[
Q(s, a) \leftarrow r(s, a) + \gamma E_{a' \sim \pi_{new}} [Q(s', a')]
\]

**Counterfactual predictions help Data Efficiency**

\[
\pi_{new} = \begin{cases} 
1 - \epsilon, & \text{if } a_t = \arg\max_a Q(s, a) \\
\epsilon & \frac{1}{|\mathcal{A}| - 1}
\end{cases}
\]

Policy Rollout
(Generate Samples with Interaction)
Structure for RL: **Off-policy RL**

Counterfactual Data Augmentation helps How to learn this structure automatically?

Discover Causal Dynamics Structure from Visual Data

(V-CDN, Neurips 2020)
Structure for RL: **Off-policy RL**

**Does the choice of architecture matter?**

\[ Q(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{a' \sim \pi_{\text{new}}} [Q(s', a')] \]

Using Dense connections in Policy/Value improves sample efficiency

[D2RL, 2020], [C-Learning, ICLR 2021]
Structure for RL: Off-policy RL

Can we use better Utility Functions?

Learning Cumulative Accessibility $C(s, a, h)$ is better than $Q(s, a)$
Can represent multimodal, multi-goal, horizon-aware solutions as well as reachability

$$Q(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{a' \sim \pi_{new}} [Q(s', a')]$$

[D2RL, 2020], [C-Learning, ICLR 2021]
Structure for RL: Off-policy RL

\[
\min_Q \mathbb{E}_{a \sim \pi_{\text{Behavior}}} [\mathcal{L}(Q(s, a) - Q_{\text{Target}})] \\
Q(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{a' \sim \pi_{\text{new}}} [Q(s', a')]
\]

Policy Improvement

\[
\pi_{\text{new}} = \left\{ \begin{array}{ll}
1 - \epsilon, & \text{if } a_t = \text{argmax}_a Q(s, a) \\
\epsilon & \frac{a_t}{|\mathcal{A}| - 1}
\end{array} \right.
\]
Structure for Reinforcement Learning

\[
Q(s, a) \leftarrow r(s, a) + \gamma E_{a \sim \pi_{new}}[Q(s', a')]
\]

Not the same Policies

Update Rule

\[
\min_{Q} \mathbb{E}_{a \sim \pi_{Behavior}}[\mathcal{L}(Q(s, a) - Q_{Target})]
\]
Structure for Reinforcement Learning

Structures Models for hierarchy?

- choose $c \sim p(c)$
- choose $z \sim p(z)$

- meta-dynamics
- action generator
- dynamics

- Transition Model
- Planning

Not the same Policies

Update Rule

Structure for hierarchy?
Structure for Reinforcement Learning

Dynamics Prediction, Correct objective?

Task-Aware Objective helps efficient and targeted Model-Learning
Structure for Reinforcement Learning

How can better state representations capture multimodal data?

Generalizable Multi-modal State Representations
- Learn a joint Visuo-Tactile representation for Peg Transfer
- Representation transfers to new task, while Policy doesn’t
**Structure for Reinforcement Learning**

Variable Impedance Controller in End-Effector Space

- Efficiency in RL
- Ease of Sim2Real Transfer

\[ \tau = f_{\text{Sim}}(\pi(o_t)) \]

\[ \tau = f_{\text{Real}}(\pi(o_t)) \]

How can better action representations result in generalization?

\[ \pi(o_t) = a: [x_d, \dot{x}_d, K_p, K_v] \]

\[ \tau = f(x_d, \dot{x}_d, K_p, K_v) \]

VICES IROS 2020
Structure for Reinforcement Learning

Learned Action Representations for Shared Autonomy

Learned Action Space
Easier to control high-dimensional robots by embedding the robot’s actions into a low-dimensional latent space

Controlling Assistive Robots with Learned Latent Actions, ICRA 2020, AURO 2021
Structure for Reinforcement Learning

Centroidal Task Space
- Easier for learning
- RL + Optimal Control

Action representations for Legged Locomotion?

GLiDE: Generalizable Quadrupedal Locomotion, under review 2022
Structure for Reinforcement Learning

Data Augmentation

Abstract State

Previous Action

Model Architecture

Utility/Value Function

Q(s, a) ← r(s, a) + γ⋅max_{a'} Q_{Target}(s', a')

min_{Q} E_{a ~ π} [Q(s, a) - Q_{Target}] + Q_{Target}

Not the same Policies

Update Rule

Learned Action Decoder

Learned State Encoder

Not the same Policies
Structure for Reinforcement Learning

- **Observation**: $O$
- **Abstract State**: $S$
- **Previous Action**: $\pi_{\text{Behavior}}$
- **Abstract Action**: $A$
- **Control**: $U$

**Agent**
- **Perception**
- **Transition Model**
- **Reactive Policy**
- **Value Function**

**Update Rule**
$$Q(s, a) \leftarrow r(s, a) + \gamma E_{\pi_{\text{new}}} Q(s', a')$$

**Minimization**
$$\min Q E_{\pi_{\text{Behavior}}} [L(Q(s, a) - Q_{\text{Target}})]$$

**Not the same Policies**
Structure in Compositional Planning

Visuo-Motor Skills
- Grasping
- Picking
- Wiping
- Pushing
- Open door
- Sweeping
- Hammering
- Cutting

Compositional Planning
- Laundry
- Put
- Soap
- Put in Washer
- Open
- Pick
- Push

Do Laundry
- Stack
- Pick
- Move
- Place
- Grasp
- Release

Visuo-Motor Skills
- Hammering
- Cutting
- Sweeping

Compositional Planning
- Laundry
- Put
- Soap
- Put in Washer
- Open
- Pick
- Push
Structure in Compositional Planning

**Imitation**: But at which level? What should I copy?
What does it mean to “open” a “door”? “open” a “jar”? “open” less-than symbol? 

Causal Generative Model

- Learn to predict the “effect” of “action”
- Compositional & Counterfactual
- Multi-step Semantic consistency
- Pre-trainable over large problem settings

Structure in Compositional Planning
Structure in Compositional Planning

State $s_0$ → Feed-Forward Model → Latent State $s$ or $z$ → Reference or Implicit Reward → Feedback Controller → Measured State $s$ or $z$

Object-Centric Generative Model of Action Concepts

Agent-Centric Policy/Planner

Input

→ “Take” “Jug”
→ “Open” “Fridge”
→ “Put” “Jug” in “Fridge”

Goal Generation

Goal-conditioned Reactive controller

Solvable online for different agents
Structure in Compositional Planning

State $s_0$ → Feed-Forward Model → Latent State $s$ or $z$

Object-Centric Generative Model of Action Concepts

Reference or Implicit Reward

Agent-Centric Policy/Planner

Feedback Controller

Measured State $s$ or $z$

Input

→ “Take” ”Jug”
→ “Open” ”Fridge”
→ “Put” ”Jug” in ”Fridge”

Goal generation

Goal-conditioned Reactive controller

Solvable online for different agents
Object-Centric Causal Generative Model
Semantic + Action-Conditional

Semantic action-conditional video prediction

Self-Supervised Modular Object Representation
Long-term Semantically Consistent Predication
No bounding box or object level supervision.
Prompt: Sequential Language Instruction
Object-Centric Causal Generative Model
Modular Action Concepts

Input: t=1

→ “Take” “Jug”
→ “Open” “Fridge”
→ “Put” “Jug” in “Fridge”

→ “Pick up” “Green Bowl”
→ “Place in” “large pink bowl”

Ground Truth Instruction
Ground Truth
MAC Prediction
Object-Centric Causal Generative Model
Systematic Generalization: Out of Distribution

All **Red** cubes are removed from training data

Testing: Concurrent actions
Training: Single action
Structure in Compositional Planning

State $s_0$ → Feed-Forward Model → Latent State $s$ or $z$

Object-Centric Generative Model of Action Concepts
Reference or Implicit Reward

Feedback Controller → Measured State $s$ or $z$

Agent-Centric Policy/Planner

Input

→ “Take” ”Jug”
→ “Open” ”Fridge”
→ “Put” ”Jug” in ”Fridge”

Goal Generation

Goal-conditioned Reactive controller

Solvable online for different agents
Input

→ “Take” "Jug"
→ “Open” "Fridge"
→ “Put” "Jug" in "Fridge"

Goal Generation

Goal-conditioned Reactive controller

Solvable online for different agents

IROS 2022 (under review)
Agent-centric Planner

Object-centric Planner

Scene Interpreter
Grasp Pose
Articulation

Keyframes Generator

Proprioception Data

RGB-D
Semantic Mask

Applied Torque

Semantic Mask
Agent-centric Planner

Object-centric Planner

Scene Interpreter
- RGB-D
- Semantic Mask
- Proprioception Data
- Grasp Pose
- Articulation

Keyframes Generator

Agent-centric Planner

Applied Torque

IROS 2022 (under review)
Agent-centric Planner

Object-centric Planner

Scene Interpreter

Keyframes Generator

Proprioception Data

RGB-D

Semantic Mask

Grasp Pose

Articulation

Agent-centric Planner

Applied Torque

IROS 2022 (under review)
Agent-centric Planner

Object-centric Planner

Scene Interpreter

Keyframes Generator

Proprioception Data

RGB-D

Semantic Mask

Grasp Pose

Articulation

Agent-centric Planner

Mapping

optimization

model

SLQ-MPC

Inverse Dynamics Controller

Applied Torque

IROS 2022 (under review)
Structure in Compositional Planning: Setup

Different kitchen layouts designed on NVIDIA Isaac Sim using PartNet-Mobility dataset

(a) Drawers
(b) Ovens
(c) Washing Machines
Static Scene: novel instances of known articulated object category

Simulation: Wheel-base
Hardware: Legged-base
Structure in Compositional Planning

State $s_0$ → Feed-Forward Model → Latent State $s$ or $z$ → Feedback Controller → Measured State $s$ or $z$

Object-Centric Generative Model of Action Concepts

Reference or Implicit Reward

Agent-Centric Policy/Planner

Input

$→$ “Take” ”Jug”
$→$ “Open” ”Fridge”
$→$ “Put” ”Jug” in ”Fridge”

Goal Generation

Goal-conditioned Reactive controller

Solvable online for different agents
Representations for Planning

What model structure enables longer term planning?

Program Induction provides a very efficient model of compositional generalization.
Structured Biases improve both efficiency & generalization
Robot Learning needs new ones!
Towards Generalizable Autonomy: Structure in Reinforcement Learning for Robotics

Animesh Garg

garg@cs.toronto.edu

@animesh_garg