

Cognition and Control in Integrated Robot Systems

Introduction to Mobile Robotics Course

Mohan Sridharan

Chair (Professor) in Robot Systems

School of Informatics, University of Edinburgh (UK)

m.sridharan@ed.ac.uk

<https://homepages.inf.ed.ac.uk/msridhar/>

Spring 2026

Requirements of Integrated Robot Systems

- **Multistep, multilevel decisions** based on multimodal inputs: vision, speech, touch etc.
- Operate under **open world uncertainty**: optimal decisions unknowable, probabilities not always meaningful?
- Operate under **resource constraints**: computation, storage, training examples.
- Acquired/learned models **need to be revised over time**.
- **Transparency and understanding** in decision making.

Characteristics of “Modern AI” Methods for Robotics

- Deep networks, transformers: excellent **statistical predictors** for some tasks; may make arbitrary decisions.
- **Narrow set of representations and update processes**; what is being represented? what properties hold?
- **Resource hungry**: many **free parameters**, computation, storage, seen as necessity for generalization?
- **Batch learning**: rapid, incremental revisions difficult.
- **Opacity** has serious consequences; **transparency, reproducibility, sustainability**?

Back to Basics: Revisit Core Principles?

- Early pioneers deeply interested in **NI and AI**; **cognition and control in “natural systems”** very relevant to robotics.
- **Representation and processing commitments** determine capabilities (McCarthy-Hayes 1969); **attend** to suitable choices and abstractions (Turing 1952, Sloman 2012).
- Behavior a **joint function** of agent and environment; skills acquired **incrementally** and **compositionally** through **adaptive satisficing** (Simon 1956; Gigerenzer 2021).
- Skills evolved jointly over long periods for hard problems; **focusing on some may not lead to others** (Minsky 1986).

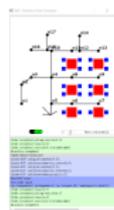
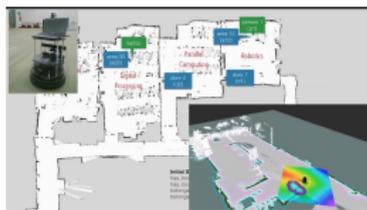
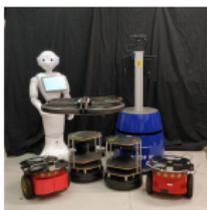
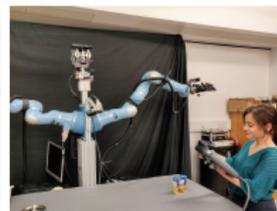
Mohan Sridharan. **Back to the Future of Integrated Robot Systems**. Senior Member Track of AAAI Conference on AI (AAAI), 2025.

Tentative Outline

- 1 (L1) Knowledge representation and reasoning (KRR).
- 2 (L2) Learning, control, and teamwork.
- 3 (L3) Explanations, integrated systems.

Illustrative Domain: Robot Assistants

Robot assistant finding and manipulating objects.



Integrated Cognitive Robot Systems: Desiderata

- Enable robots to **represent**, **reason**, and **act** with **different descriptions** of domain knowledge and uncertainty.
“Books are usually in the library”
“I am 90% certain the robotics book is in the library”
- Enable robots to **learn interactively** and **cumulatively** from sensor inputs and limited human feedback.
Learn actions, action capabilities, domain dynamics
“Robot with weak arm cannot lift heavy box”
- Enable designers to **understand** the robot's behavior and establish that it **satisfies desirable properties**.
Explainable agency, intentions, goals, measures
“What would happen if I dropped the spoon on the table?”

Inspiration and Core Ideas

- **Cognitive systems** inspired by human cognition, control.
- **Represent, reason, act, learn jointly** at **different abstractions** with **different schemes**.
- **Logician, statistician, creative explorer**; **formal coupling** not unified representation.
- Combine **knowledge-based** and **data-driven** reasoning and learning; **predictive, cumulative, interactive, relevant**.
- Explanations: **relational descriptions** of decisions, beliefs;
Questions: **descriptive, causal, contrastive, counterfactual**.

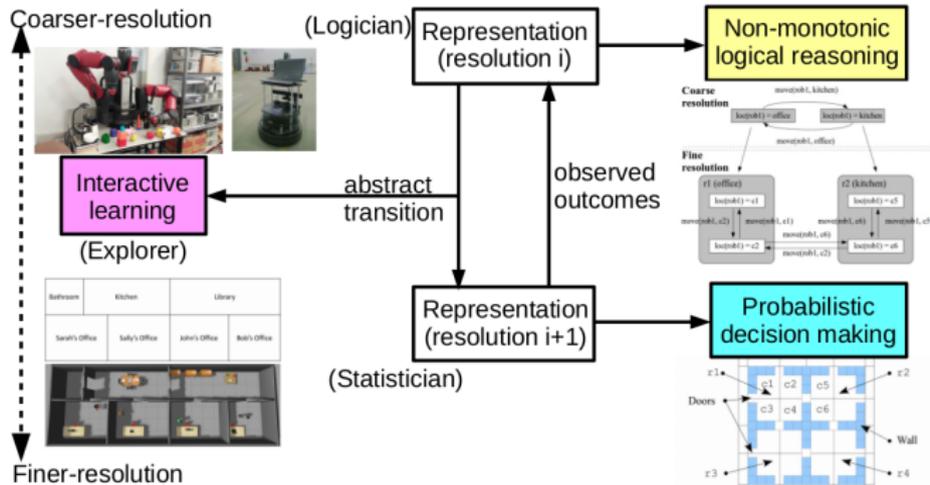
Shiqi Zhang and Mohan Sridharan. **A Survey of Knowledge-based Sequential Decision Making under Uncertainty**. Artificial Intelligence Magazine, 43(2):249-266, 2022.

Claims: Representation + Reasoning + Learning

- 1 **Distributed representation** of knowledge (commonsense, probabilistic) at **different coupled abstractions**.
- 2 **Separation of concerns** (domain-specific/independent knowledge, observations); **common methodology**.
- 3 Knowledge elements support **non-monotonic revision**; revise previously held conclusions.
- 4 “**Here and there**” reasoning; **satisfiability, stochastic policies**. Often focus on **rationality** and not on optimality!

Illustrative domains: **visual planning, scene understanding** and **manipulation** problems in robotics.

Refinement-Based Architecture: Overview



Exploit complementary strengths of **non-monotonic logical reasoning**, **probabilistic reasoning**, and **interactive learning**.

Mohan Sridharan. **REBA-KRL: Refinement-Based Architecture for Knowledge Representation, Explainable Reasoning, and Interactive Learning in Robotics**. European Conference on Artificial Intelligence, 2020.

Mohan Sridharan, Michael Gelfond, Shiqi Zhang and Jeremy Wyatt. **REBA: Refinement-based Architecture for Knowledge Representation and Reasoning in Robotics**. Journal of Artificial Intelligence Research, 65:87-180, 2019.

Tentative Outline

- 1 (L1) Knowledge representation and reasoning (KRR).
- 2 (L2) Learning, control, and teamwork.
- 3 (L3) Explanations, integrated systems.

Robot Waiter Example: Reasoning (Video)

Example

Action Language + Logician's System Description

- AL_d : formal description of transition diagrams.
- **System description** \mathcal{D}_C : sorted signature Σ_C and axioms as statements in AL_d .
- **Statics**: $next_to(place, place)$.
- **Fluents**: $loc : thing \rightarrow place$,
 $in_hand : robot \times object \rightarrow boolean$.
- **Actions**: $move(robot, place)$, $grasp(robot, object)$,
 $exo_move(object, place)$, $exo_lock(place)$.

Logician's System Description: Axioms

- Causal law, state constraint, executability condition.

- **Causal laws:**

$move(rob1, Pl)$ **causes** $loc(rob1) = Pl$

$grasp(rob1, Ob)$ **causes** $in_hand(rob1, Ob)$

$putdown(rob1, Ob)$ **causes** $\neg in_hand(rob1, Ob)$

- **State constraints:**

$loc(Ob) = Pl$ **if** $loc(rob1) = Pl$, $in_hand(rob1, Ob)$

$loc(Th) \neq Pl_1$ **if** $loc(Th) = Pl_2$, $Pl_1 \neq Pl_2$

- **Executability conditions:**

impossible $grasp(rob1, Ob)$ **if** $loc(rob1) \neq loc(Ob)$

impossible $grasp(rob1, Ob)$ **if** $in_hand(rob1, Ob)$

impossible $putdown(rob1, Ob)$ **if** *not* $in_hand(rob1, Ob)$

Histories with Defaults

- **History** contains records of observations and actions:

obs(fluent, boolean, step)

hpd(action, step)

- Expand to include **initial state defaults**:

initial default $loc(X) = library$ **if** $textbook(X)$

initial default $loc(X) = office$ **if** $textbook(X)$,
 $loc(X) \neq library$

- **Consistency-restoring rules** for recovery and **diagnostics**.

$loc(X) \neq library \stackrel{+}{\leftarrow} textbook(X)$

Modeling Intentions: What are they?

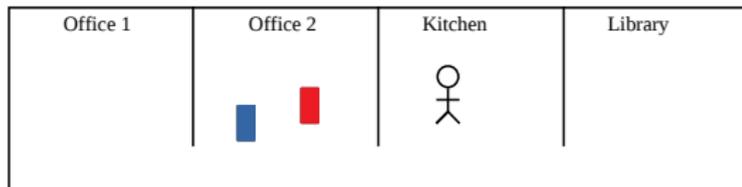
- Many different “definitions” proposed; also survey papers.
- Inferred from sensor inputs (gaze, gestures), features (tracked body pose and movement), or “meta” concepts.
- Intention as **joint high-level concept** defined over robot's beliefs and actions.

Tom Carlson and Yiannis Demiris. **Human-Wheelchair Collaboration Through Prediction of Intention and Adaptive Assistance**. International Conference on Robotics and Automation, 2008.

Adam Norton, Henny Admoni, Jacob Crandall, Tesca Fitzgerald, Alvika Gautam, Michael Goodrich, Amy Saretsky, Matthias Scheutz, Reid Simmons, Aaron Steinfeld, and Holly Yanco. **Metrics for Robot Proficiency Self-Assessment and Communication of Proficiency in Human-Robot Teams**. Transactions on Human-Robot Interaction, 11(3), 2022.

Modeling Intentions I

- Unexpected success and failure.



- Persistence, non-procrastination, relevance.
- Expand to $\Pi(\mathcal{D}'_C$ and $\mathcal{H}'_C)$; **activities**; **mental** fluents and **mental** actions.

Rocio Gomez, Mohan Sridharan, and Heather Riley. **What do you really want to do? Towards a Theory of Intentions for Human-Robot Collaboration.** In *Annals of Mathematics and Artificial Intelligence*, special issue on Commonsense Reasoning, 89(1): 179-208, February 2021.

Modeling Intentions II

- Expand Σ_H :
 - **Activity**: goal, plan, name.
 - **Mental** fluents and actions.
- **Expand axioms** to represent action effects, start/stop activity, generate intentional actions.
- Expand \mathcal{H} , e.g., to model **attempted actions**:

obs(fluent, boolean, step), hpd(action, step)
attempt(action, step), \neg hpd(action, step)

Modeling Affordances: What are they?

- Multiple interpretations and surveys, surveys of surveys?
- **Attribute** of object, agent, environment?
- Behavior based on agent's process and environment.
- Affordance as **joint attribute** of agent and object in the context of specific actions.

Keith S. Jones. **What is an Affordance?** Ecological Psychology, 15(2):104-114, 2003.

L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor. **Affordances in Psychology, Neuroscience and Robotics: A Survey.** IEEE Transactions on Cognitive and Developmental Systems, 2016.

P. Zech, S. Haller, S. R. Lakani, B. Ridge, E. Ugur, and J. Piater. **Computational Models of Affordance in Robotics: A Taxonomy and Systematic Classification.** Adaptive Behavior, 25(5): 235-271, 2017.

V. Sarathy and M. Scheutz. **A Logic-based Computational Framework for Inferring Cognitive Affordances.** IEEE Transactions on Cognitive and Developmental Systems, 10(1):26-43, 2018.

Modeling Affordances

- **Affordance** as combination of attributes of object(s) and agent(s) with reference to an action.
- Action can have **enabling** or **forbidding** affordances.
 - impossible** $\text{pickup}(R, O)$ **if** $\text{obj_weight}(O, \text{heavy})$,
 - $\text{not } \text{aff_enables}(id_1, \text{pickup}(R, O))$
 - $\text{aff_enables}(id_1, \text{pickup}(R, O))$ **if** $\text{strength}(R, \text{strong})$
 - impossible** A **if** $\text{aff_forbids}(ID, A)$
 - $\text{aff_forbids}(id_j, A)$ **if** ...
- **Distributed representation** supports information reuse.

Pat Langley, Mohan Sridharan, and Ben Meadows. **Representation, Use, and Acquisition of Affordances in Cognitive Systems**. AAAI Spring Symposium on Integrating Representation, Reasoning, Learning, and Execution for Goal Directed Autonomy, Stanford, USA, March 26-28, 2018.

Mohan Sridharan and Ben Meadows. **Knowledge Representation and Interactive Learning of Domain Knowledge for Human-Robot Collaboration**. Advances in Cognitive Systems Journal, 7:77-96, 2018.

Logician's Reasoning

- Logician's description:
 - **Input:** (a) \mathcal{D}_C and history \mathcal{H}_C ; (b) Goal.
 - **Output:** plan and next **transition** $T = \langle \sigma_1, a^C, \sigma_2 \rangle$ to execute.
 - Can translate to different formalisms for reasoning.
- **Answer Set Prolog** program $\Pi(\mathcal{D}_C, \mathcal{H}_C)$. Reason by computing **answer sets**. **Non-monotonic logical reasoning**.
- **Default negation** and **epistemic disjunction**.

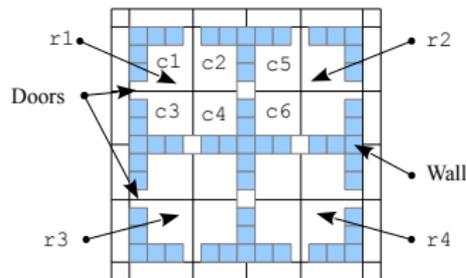
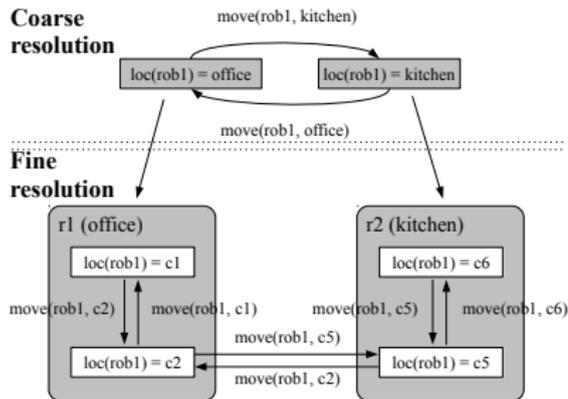
\neg I I is believed to be false

not I it is not believed that I is true

$p \vee \neg p$ is a tautology

$p \text{ or } \neg p$ is not tautological

Refinement: Overview



- **Refinement**: describe (\mathcal{D}_C) at finer resolution (\mathcal{D}_F).
- **Formal relationships**; add knowledge fluents and actions.

Mohan Sridharan, Michael Gelfond, Shiqi Zhang and Jeremy Wyatt. **REBA: Refinement-based Architecture for Knowledge Representation and Reasoning in Robotics**. In *Journal of Artificial Intelligence Research*, 65:87-180, May 2019.

Weak Refinement ($\mathcal{D}_{F,nobs}$) I

Refine signature Σ_F of τ_F :

- Inherit basic sorts and define s^* **counterparts**.

$$place = \{r_1, \dots, r_n\}, \quad place^* = \{c_1, \dots, c_m\}$$

$$cup = \{cup_1\}, \quad cup^* = \{cup_base_1, cup_handle_1\}$$

- Add new statics, fluents, and actions; define **component relationships**.

$$next_to^*(place^*, place^*)$$

$$loc^* : thing \rightarrow place^*, \quad cup \notin thing, \quad loc^* : cup^* \rightarrow place^*$$

$$move^*(robot, place^*), \quad grasp^*(robot, cup^*)$$

$$component(place^*, place), \quad component(cup^*, cup)$$

Weak Refinement ($\mathcal{D}_{F, nobs}$) II

- **Causal laws:**

$move^*(R, C)$ **causes** $loc^*(R) = C$

$grasp(R, O)$ **causes** $in_hand(R, O)$, $O \neq cup_1$

$putdown^*(R, O)$ **causes** $\neg in_hand^*(R, O)$, $O \in cup^*$

- **State constraints** (including **bridge axioms**):

$loc^*(O) = C$ **if** $loc^*(R) = C$, $in_hand(R, O)$

$next_to^*(C_2, C_1)$ **if** $next_to^*(C_1, C_2)$

$loc(Th) = P$ **if** $component(C, P)$, $loc^*(Th) = C$

$loc^*(O) = C$ **if** $loc^*(OPart) = C$, $component(OPart, O)$

- **Executability conditions:**

impossible $move^*(R, C_2)$ **if** $loc^*(R) = C_1$, $\neg next_to^*(C_1, C_2)$

impossible $grasp(R, O)$ **if** $loc^*(R) \neq loc^*(O)$

impossible $putdown(R, O)$ **if** $not\ in_hand(R, O)$

Strong Refinement (\mathcal{D}_F)

- Introduce **theory of observations**: knowledge fluents, knowledge-producing actions.
- Introduce new fluents, actions, and axioms to observe the environment.

$observed_f : robot \times dom(f) \times range(f) \rightarrow \{true, false, undet\}$

$test_f : robot \times dom(f) \times range(f) \rightarrow boolean$

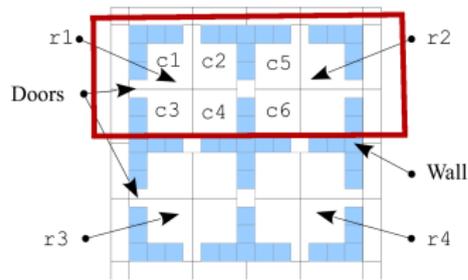
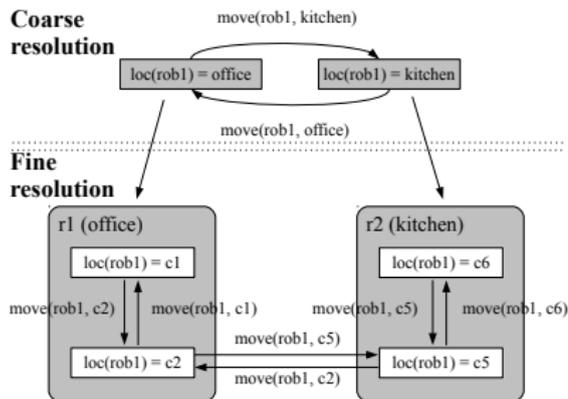
- Inherit axioms of \mathcal{D}_C ; expand as appropriate.

$test_{f^*}(R, \bar{X}, Y)$ **causes** $observed_{f^*}(R, \bar{X}, Y)$ **if** $f^*(\bar{X}) = Y$

$test_{f^*}(R, \bar{X}, Y)$ **causes** $\neg observed_{f^*}(R, \bar{X}, Y)$ **if** $f^*(\bar{X}) \neq Y$

impossible $test_{f^*}(R, \bar{X}, Y)$ **if** $\neg can_be_observed_{f^*}(R, \bar{X}, Y)$

Randomize and Zoom to $\mathcal{D}_{FR}(T)$

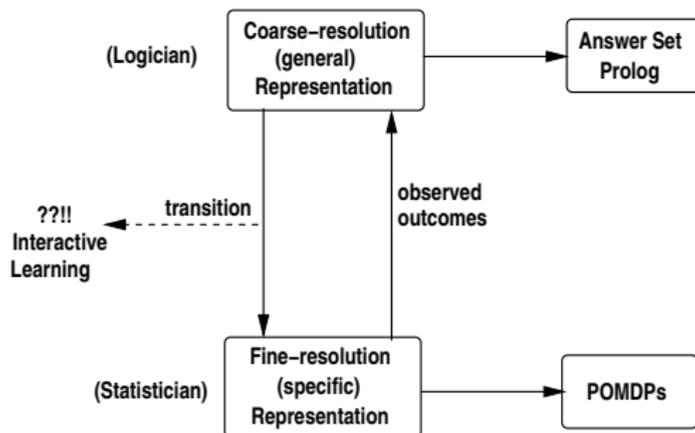


- **Randomization** to capture non-determinism (\mathcal{D}_{FR}).

$$\text{move}^*(R, C_2) \text{ causes } \text{loc}^*(R) = \{C : \text{range}(\text{loc}^*(R), C)\}$$

- Collect statistics to compute probabilities.
- Automatically **zoom** to $\mathcal{D}_{FR}(T)$ for $T = \langle \sigma_1, a^C, \sigma_2 \rangle$.

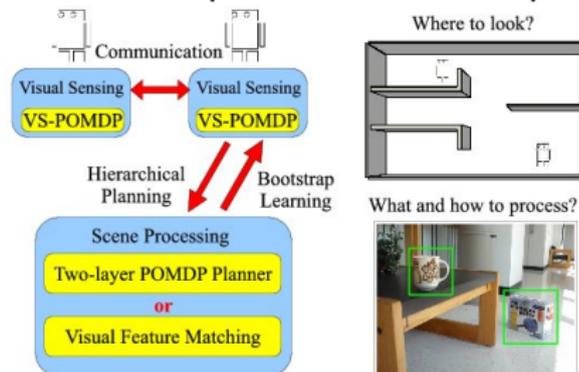
Statistician's task



- $\mathcal{D}_{FR}(T)$ and statistics to construct and solve **Partially Observable Markov Decision Process** (POMDP).
- Compute **policy** mapping belief states to actions. Invoke to execute sequence of actions.
- Add observed outcomes to \mathcal{H}_C to be used by logician.

Hierarchical POMDPs

- Tuple: $\langle S, A, Z, T, O, R \rangle$
- Where to look? What to process? How to process?



- **Policy kernels** and **adaptive observation functions**.
- **Automatic belief propagation** and **model generation** at all levels for reliable and efficient operation.

Shiqi Zhang, Mohan Sridharan and Jeremy Wyatt. **Mixed Logical Inference and Probabilistic Planning for Robots in Unreliable Worlds**. In the IEEE Transactions on Robotics, 31(3):699-713, June 2015.

Shiqi Zhang, Mohan Sridharan and Christian Washington. **Active Visual Planning for Mobile Robot Teams using Hierarchical POMDPs**. In the IEEE Transactions on Robotics, 29(4): 975-985, 2013.

Execution Trace: Reasoning

- **Goal:** some cup C has to be in the office:
 $loc(C) = office, \neg in_hand(rob_1, C)$.
- **Initial knowledge** (subset): $loc(rob_1, office)$,
 $obj_weight(cup_1, heavy)$, $arm_type(rob_1, electromagnetic)$.
- Based on **default**: $loc(cup_1) = kitchen$.
- One possible plan from ASP-based inference:
 $move(rob_1, kitchen), grasp(rob_1, cup_1)$
 $move(rob_1, office), putdown(rob_1, cup_1)$
- Assume rob_1 is in *kitchen*. Has to locate and grasp cup_1 .

Execution Trace: Reasoning

- Some **relevant** literals: $loc(rob_1) = c_i$, $loc(cup_1) = c_j$, where $c_i, c_j \in kitchen$.

- Possible action sequence (**executed probabilistically**):

move(rob₁, c₃)

test(rob₁, loc(cup₁), c₃) % *cup₁ not observed*

move(rob₁, c₅)

test(rob₁, loc(cup₁), c₅) % *cup₁ observed*

grasp(rob₁, cup₁)

- Proceed if grasping succeeds; what to do when it fails?

Robot Waiter Revisited: Reasoning (Video)

Example

Robot Waiter Video 2

Example

Advantages

- **Step-wise refinement** simplifies design and implementation.
- **Increases confidence** in behavior, **promotes scalability**.
- **Separation of concerns**: domain-independent and domain-specific knowledge.
- Designer follows pre-defined steps; **otherwise automated**.
- **Non-monotonic logical reasoning** and **probabilistic reasoning** inform and guide each other.

Tentative Outline

- 1 (L1) Knowledge representation and reasoning (KRR).
- 2 (L2) Learning, control, and teamwork.
- 3 (L3) Explanations, integrated systems.

Reasoning + Learning: Motivation

- Machine (deep?) learning widely used in AI and robotics.
 - Large **labeled datasets**; **computational/memory-heavy**; and
 - Representations and mechanisms **difficult to interpret**.
- Inspiration from **cognitive systems**:
 - Representation, reasoning, learning **inform each other**.
 - Scalability: **abstraction**, **relevance**, and **persistence**.
- **Experimental domains**:
 - Estimate object occlusion, stability; **Visual Question Answering (VQA)**.
 - Human-robot interaction; robot manipulation.

Pat Langley and Herbert A. Simon. **The Central Role of Learning in Cognition**. Cognitive skills and their acquisition, J. Anderson (ed.). Lawrence Erlbaum Associates, 1981.
Pat Langley. **The Central Role of Cognition in Learning**. Advances in Cognitive Systems, 4:3-12, 2016.

Bounded Rationality/Heuristic Methods: Three Views

- **Risk or uncertainty:** closed/small or open worlds.
- Herb Simon's definition of **Bounded Rationality**:
 - Study of human decision making under uncertainty.
 - **Focus on satisficing instead of optimization.**
 - Behavior function of cognition and environment.
- Definition hijacked and perverted by others:
 - Finance/Computer Science: **optimal search.**
 - Psychology: **heuristics-and-biases** program. Heuristics to explain human bias or irrationality.

Gerd Gigerenzer. **What is Bounded Rationality?** Routledge Handbook of Bounded Rationality, Riccardo Viale (editor), Routledge, 2021.

Konstantinos Katsikopoulos, Ozgur Simsek, Marcus Buckmann and Gerd Gigerenzer. **Classification in the Wild: The Science and Art of Transparent Decision Making.** MIT Press, 2021.

Jan Malte Lichtenberg and Ozgur Simsek. **Regularization in Directable Environments with Application to Tetris.** International Conference on Machine Learning, 2019.

Ecological Rationality Toolbox

- **Ecological rationality**: satisficing with adaptation.
 - Algorithmic model of heuristics.
 - Competitive testing of predictions.
- **Heuristics**: ignore some information to make decision more quickly, frugally, and/or accurately.
 - One-reason (hiatus); sequential-search (take the best), tallying; **fast and frugal trees**.
 - **Adaptive toolbox**: descriptive, prescriptive, engineering!
- Identify attributes, **learn predictive models** in many domains: medicine, legal, social decisions; **“optimization” driven by different principles!**

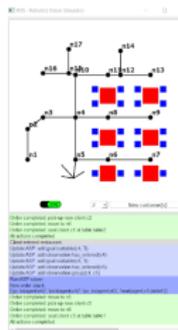
Ian N. Durbach, Simon Algorta, Dieudonne Kabongo Kantu, Konstantinos V. Katsikopoulos, and Ozgur Simsek. **Fast and Frugal Heuristics for Portfolio Decisions with Positive Project Interactions**. Decision Support Systems, 138, 2020.

Nadine Fleischhut and Gerd Gigerenzer. **Can Simple Heuristics Explain Moral Inconsistencies?** Simple Heuristics in a Social World, R. Hertwig, U. Hoffrage, and ABC group (eds.), Oxford University Press, 2013.

Scene Understanding + Planning

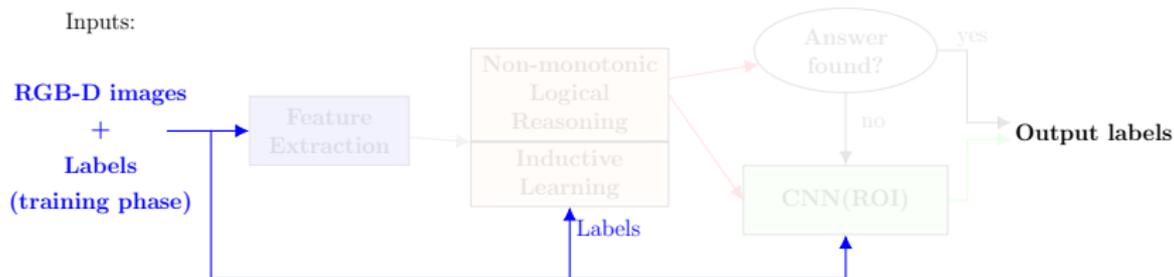


Bathroom	Kitchen	Library	
Sarah's Office	Sally's Office	John's Office	Bob's Office

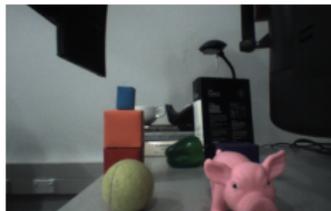


A (very) brief overview of ML for those who need it...

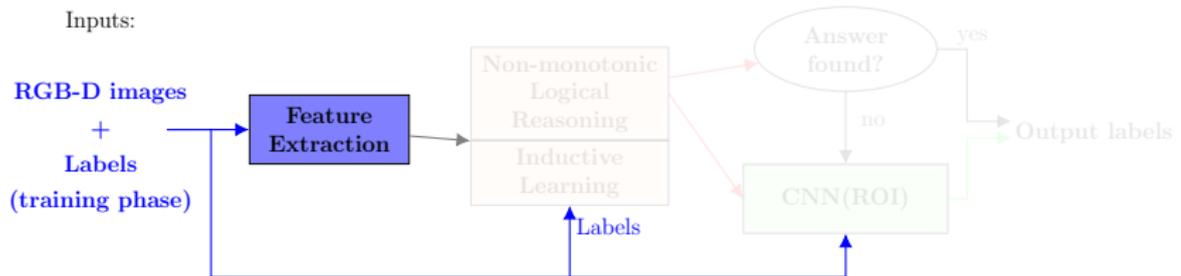
Architecture Components: Input



- **Images:** images of objects, scenes.
- **Labels:** object occlusion, stability of structures, answers.



Architecture Components: Feature Extraction

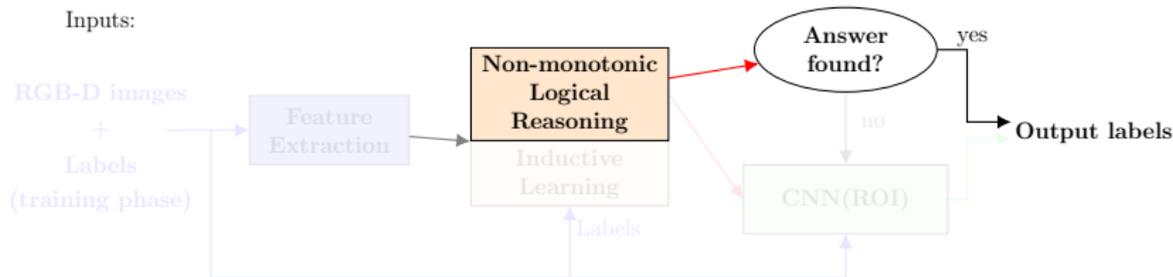


Geometric features extracted from images:

- **Spatial relations** between objects (above, behind, left of ...).
- **Color, shape, and size** of objects in the scene.
- Incremental grounding of prepositions for spatial relations.

Tiago Mota and Mohan Sridharan. **Incrementally Grounding Expressions for Spatial Relations between Objects**. In the International Joint Conference on Artificial Intelligence (IJCAI), July 13-19, 2018.

Architecture Components: Non-monotonic Logic



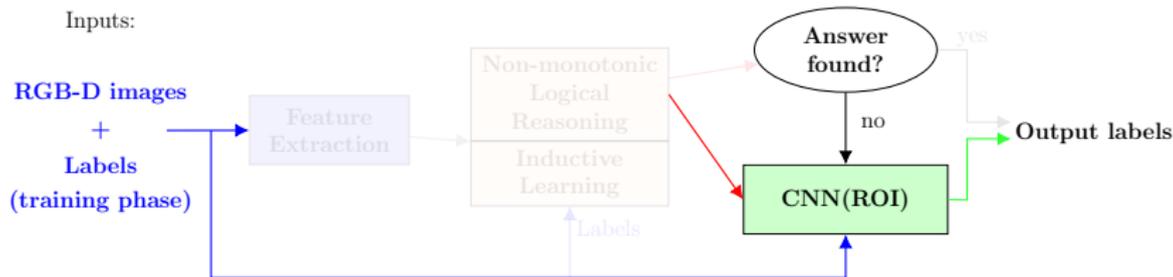
- **Input:** Extracted features, incomplete domain dynamics.
- **ASP** for non-monotonic logical reasoning.

$stable(A) \leftarrow not\ obj_rel(above, A, B)$

$\neg occurs(pickup(rob_1, O_1), I) \leftarrow holds(obj_rel(below, O_1, O_2), I)$

- Decision about input image if possible.

Architecture Components: CNN

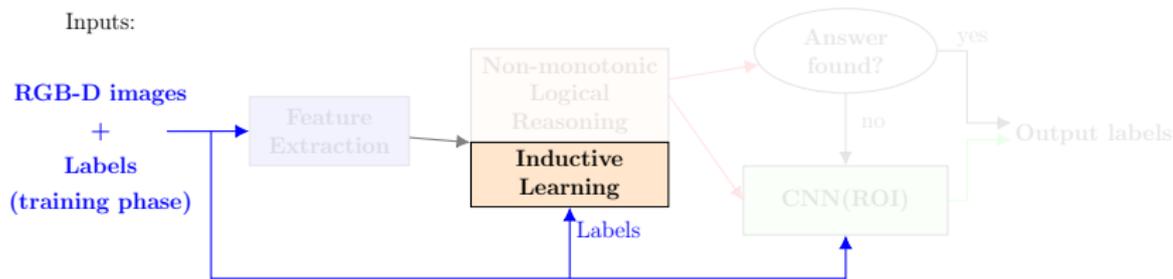


- **Attention**: ROI selection based on axioms.

$$\begin{aligned} \text{stable}(A) &\leftarrow \text{not } \text{obj_rel}(\text{above}, A, B) \\ \neg \text{stable}(A) &\leftarrow \text{obj_rel}(\text{above}, A, B), \text{size}(A, \text{large}) \\ &\quad \text{size}(B, \text{small}) \end{aligned}$$

- **CNN**: Convolutional Neural Network (Lenet and Alexnet).

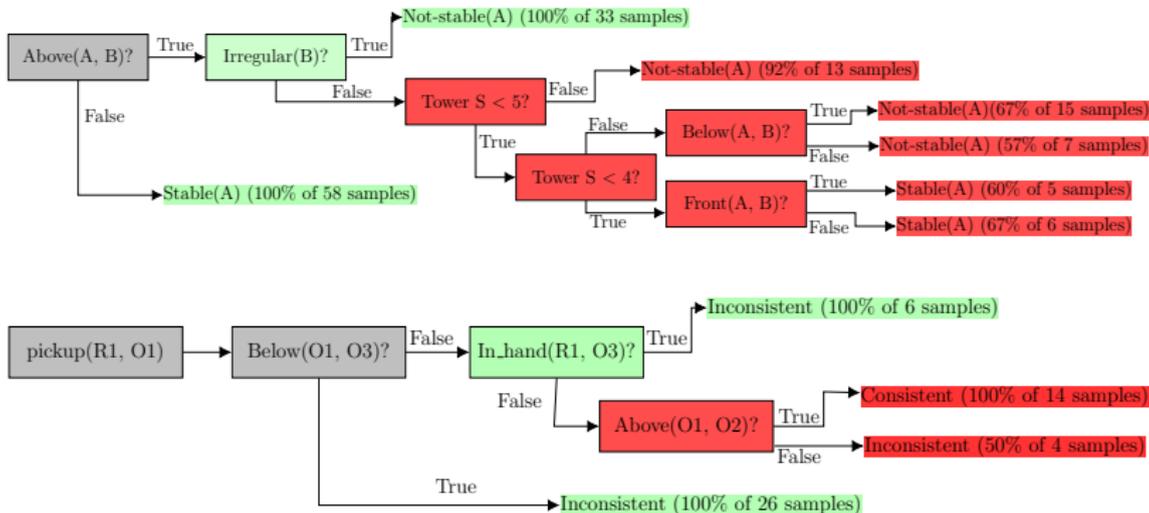
Architecture Components: Inductive Learning



- **Input:** features and figure labels.
- **Decision Tree:** induction of rules (constraints, causal laws).
- **Output:** learned rules.

Tiago Mota and Mohan Sridharan. **Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning on Robots**. In the Robotics Science and Systems Conference (RSS), Freiburg, Germany, June 22-26, 2019 (Best Paper Award Finalist)

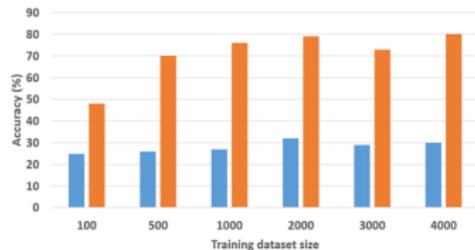
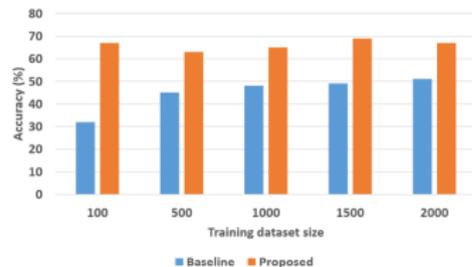
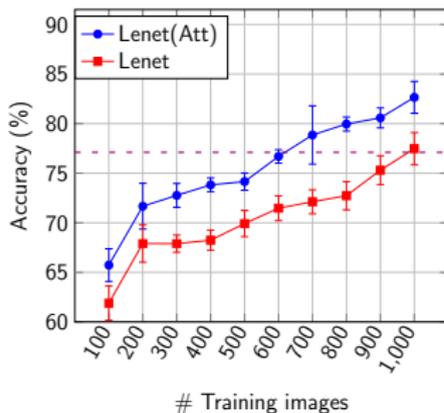
Architecture Components: Inductive Learning



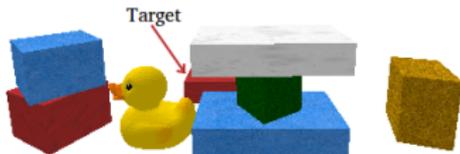
$\neg stable(A) \leftarrow obj_rel(above, A, B), obj_surface(B, irregular)$
 $\neg pickup(R, O1) \leftarrow in_hand(R, O2)$

Return to Scene understanding

- Accuracy increases and training complexity decreases.

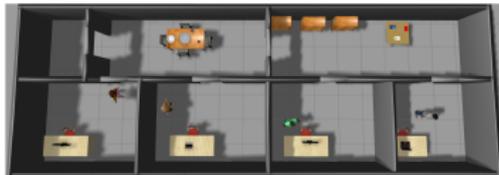


Experimental Results: VQA + Decision making



- **Initially:** 64 plans; most incorrect or sub-optimal.
- **Including learned axioms:** 3 correct plans.

Bathroom	Kitchen	Library	
Sarah's Office	Sally's Office	John's Office	Bob's Office



- **Without learned axioms:** four times as many plans; six times as much time per plan execution.

Heather Riley and Mohan Sridharan. **Integrating Non-monotonic Logical Reasoning and Inductive Learning With Deep Learning for Explainable Visual Question Answering.** In *Frontiers in Robotics and AI*, special issue on Combining Symbolic Reasoning and Data-Driven Learning for Decision-Making, Volume 6, December 2019.

Return to Execution Trace: Reasoning

- **Goal:** some cup C has to be in the office:
 $loc(C) = office, \neg in_hand(rob_1, C)$.
- **Initial knowledge** (subset): $loc(rob_1, office)$,
 $obj_weight(cup_1, heavy)$, $arm_type(rob_1, electromagnetic)$.
- Based on **default**: $loc(cup_1) = kitchen$.
- One possible plan from ASP-based inference:
 $move(rob_1, kitchen), grasp(rob_1, cup_1)$
 $move(rob_1, office), putdown(rob_1, cup_1)$
- Assume rob_1 is in *kitchen*. Has to locate and grasp cup_1 .

Execution Trace: Reasoning + Learning

- Some **relevant** literals: $loc(rob_1) = c_i$, $loc(cup_1) = c_j$, where $c_i, c_j \in kitchen$.

- Possible action sequence (**executed probabilistically**):

$move(rob_1, c_3)$

$test(rob_1, loc(cup_1), c_3)$ % cup_1 not observed

$move(rob_1, c_5)$

$test(rob_1, loc(cup_1), c_5)$ % cup_1 observed

$grasp(rob_1, cup_1)$

- Grasping cup_1 fails; **relational learning**:

impossible $grasp(rob_1, C)$ **if** $arm_type(rob_1, electromagnetic)$,
 $obj_weight(C, heavy)$

Using Learned Knowledge: Video

SAM Example

N. Gireesh, A. Agrawal, A. Datta, S. Banerjee, M. Sridharan, B. Bhowmick, and M. Krishna. **Sequence-Agnostic Multi-Object Navigation**. IEEE International Conference on Robotics and Automation (ICRA), May 2023.

Anticipate and Act: Video

LLM-PDDL Example

R. Arora, S. Singh, K. Swaminathan, S. Banerjee, B. Bhowmick, K. M. Jatavallabhula, M. Sridharan, and M. Krishna.
Anticipate & Act: Integrating LLMs and Classical Planning for Efficient Task Execution in Household Environments. IEEE International Conference on Robotics and Automation (ICRA), May 2024.

Anticipate and Collab: Human-in-the-loop?

LLM-KG Example

S. Singh, K. Swaminathan, N. Dash, R. Singh, S. Banerjee, M. Sridharan, and M. Krishna. **AdaptBot: Combining LLM with Knowledge Graphs and Human Input for Generic-to-Specific Task Decomposition and Knowledge Refinement**. IEEE International Conference on Robotics and Automation (ICRA), Atlanta, USA, May 2025.

Reasoning + Learning: Summary

- Many (if not most) robotics problems are reasoning and learning problems.
- Mistake to formulate as just reasoning or learning problem.
- Better approach: **reasoning and learning guide each other.**
- Focus on **representation and processing commitments!**
Ecological rationality for reliable and efficient operation.
- What about robot control and teamwork?

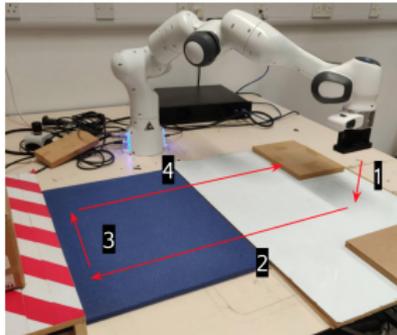
Changing-Contact Manipulation: Video

Changing contact manipulation

Saif Sidhik, Mohan Sridharan, and Dirk Ruiken. **An Adaptive Framework for Trajectory Following in Changing-contact Robot Manipulation Tasks**. In the Robotics and Autonomous Systems (RAS), 181:1-21, November 2024.

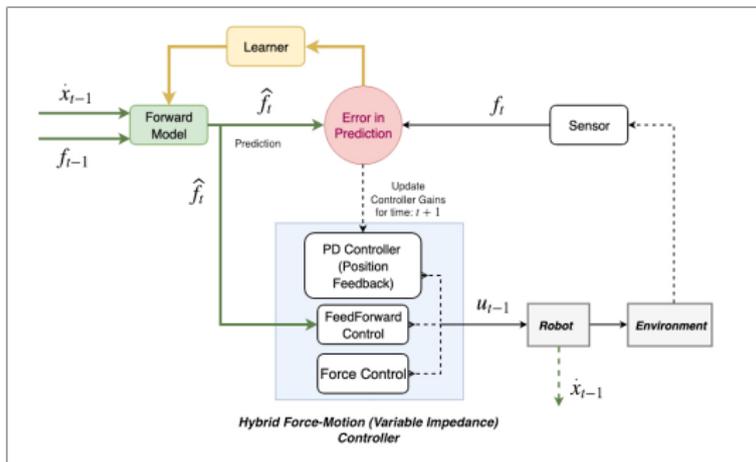
Michael Mathew, Saif Sidhik, Mohan Sridharan, Morteza Azad, Akinobu Hayashi, and Jeremy Wyatt. **Online Learning of Feed-Forward Models for Task-Space Variable Impedance Control**. In the International Conference on Humanoid Robots (Humanoids), 2019.

Changing Contact Manipulation: Problem



- Single demo of planned trajectory: **make, break contacts** with objects and surfaces; **discontinuous dynamics**.
- No visual sensors; limited knowledge of contact changes.
- Status quo: time dependence, learning/data complexity.

Changing Contact Manipulation: Approach



- **Forward models**; inspiration from human motor control.
- **Hybrid force-motion controller**; contact anticipation.

$$u_t = H_t + K_t^p \Delta x_t + K_t^d \Delta \dot{x}_t + u_t^{fc} + u_t^{ff}$$

Changing-Contact Manipulation: Video I

Changing contact manipulation

Michael Mathew, Saif Sidhik, Mohan Sridharan, Morteza Azad, Akinobu Hayashi, and Jeremy Wyatt. **Online Learning of Feed-Forward Models for Task-Space Variable Impedance Control**. In the International Conference on Humanoid Robots (Humanoids), 2019.

Changing-Contact Manipulation: Video II

Changing contact manipulation

Saif Sidhik, Mohan Sridharan, and Dirk Ruiken. **An Adaptive Framework for Trajectory Following in Changing-contact Robot Manipulation Tasks**. In the Robotics and Autonomous Systems (RAS), 181:1-21, November 2024.

Collaboration without Prior Coordination (AHT)

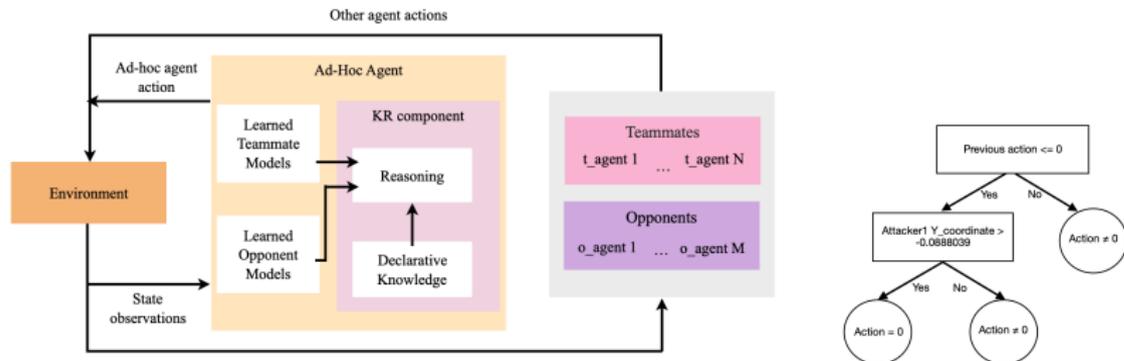
Example

- **Limited prior knowledge** of other agents/robots; observable state but **no (limited) communication**.
- **State of the art: data driven** methods.
 - Probabilistic and/or deep network-based models.
 - Estimate behavior of agent “types”, optimize actions using experience history.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, Stefano V Albrecht. **A Survey of Ad Hoc Teamwork: Definitions, Methods, and Open Problems**, arXiv:2202.10450, 2022.

Example

AHT Architecture: Overview



- Reason with domain knowledge and behavior prediction models **learned rapidly from limited data**.
- Ecological rationality**: match domain characteristics with properties of heuristic methods; **identify good features**.
- Ensemble of fast and frugal trees**: learn **predictive models**.

AHT Architecture: KRR as before

- Reason with domain knowledge at different levels of abstraction.

$move^*(Ag, X, Y)$ **causes** $in^*(Ag, X, Y)$

$\neg in^*(Ag, X1, Y1)$ **if** $in^*(Ag, X2, Y2)$, $X1 \neq X2$, $Y1 \neq Y2$

impossible $shoot(Ag, Ago)$ **if** $agent_shot(Ago)$

initial default $spread_attack(Ago)$ **if** $attacker(Ago)$

$in(Ag, R)$ **if** $in^*(Ag, X, Y)$, $component(X, Y, R)$

Hasra Dodamegama and Mohan Sridharan. **Toward a Hybrid Framework for Ad hoc Teamwork**. In the AAAI International Conference on AI (AAAI), February 7-14, 2023.

AHT Experimental Setup

- Train with simple policies, test on DNN/GNN policies.
Orders of magnitude fewer examples (5000 vs. 1M).
- **Adaptation** to different situations.
- Consider **partial observability** and **limited communication**.
- **Better performance** than data-driven systems.

Agent Type	Accuracy
Helios	86.0%
Gliders	66.4%
Cyrus	77.6%
Aut	67.7%
Axiom	73.6%
Agent2D	71.9%

Version	KAT (%)	PPAS (%)	PLAS (%)
Limited (2v2)	79	80	80
Full (4v5)	30	20	20

AHT Results: Videos

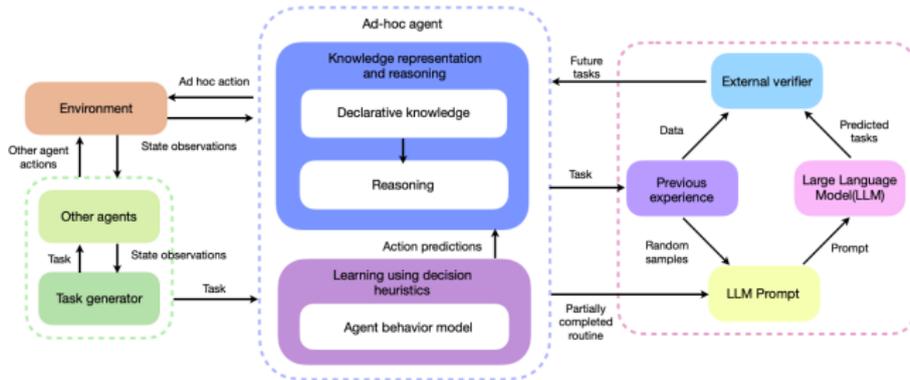
KAT FA

KAT HFO

Hasra Dodamegama and Mohan Sridharan. **Reasoning and Explanation Generation in Ad hoc Collaboration between Humans and Embodied AI**. In International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR), October 2024.

Hasra Dodamegama and Mohan Sridharan. **Knowledge-based Reasoning and Learning under Partial Observability in Ad Hoc Teamwork**. In Theory and Practice of Logic Programming, 2023.

AHT Architecture with LLM?



- Use LLM for **anticipating future tasks**.
- **Plan jointly** for current and upcoming tasks.
- Identify and use **contextual information** to prompt LLM.
- **External validator** to revise and correct LLM output.

Ad hoc Teamwork: “Embodied AI” Video (Scalability)

human

agent1

agent3

agent4

Control and Teamwork Summary

- Multiple open problems: often (incorrectly?) formulated as learning/optimization problems.
- **Choice of representation and processes still important!**
- Core principles (discussed earlier) still applicable; lead to reliable and efficient solutions.
- Often want to know why/how decisions were made: transparency, explainability, trust, safety?

Tentative Outline

- 1 (L1) Knowledge representation and reasoning (KRR).
- 2 (L2) Learning, control, and teamwork.
- 3 (L3) Explanations, integrated systems.

Explanation: Different Perspectives

- Long history: **many interpretations** across disciplines!
- Popular option: make existing “black box” models interpretable (**ML methods**); tracing decisions to features.
- **Explicability** in planning: choose options easier for humans to understand.
- Other methods: **transparency** in reasoning and learning.

Gerald Dejong and Raymond Mooney. **Explanation-Based Learning: An Alternative View**. Machine Learning, 1:145-176, 1986.

Raymond Reiter. **A Theory of Diagnosis from First Principles**. Artificial Intelligence, 32:57-95, 1987.

Tim Miller. **Explanations in Artificial Intelligence: Insights from the Social Sciences**. Artificial Intelligence, 267:1-38, 2019.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Framling. **Explainable agents and robots: Results from a systematic literature review**. AAMAS, 2019.

Ricards Marcinkevics and Julia E. Vogt. **Interpretable and Explainable Machine Learning: A Methods-Centric Overview with Concrete Examples**. WIREs Data Mining and Knowledge Discovery, 13(3), 2023.

Explanations: Important Considerations

- Important consideration: **who needs to understand?**
- Different “metrics”: simplicity, coherence, relevance.
- **Human in the loop**: provide feedback, introduce **cognitive biases** and **social expectations**.
- **Not just causal**; contrastive, counterfactual, selective, social (theory of mind).
- **Focus**: **explainable agency** in cognitive systems.

Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. **The Who in XAI: How AI Background Shapes Perceptions of AI Explanations**. CHI Conference on Human Factors in Computing Systems, 2024.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. **Interpretable machine learning: Fundamental principles and 10 grand challenges**. Statistics Surveys, 16:1-85, 2022.

Cynthia Rudin. **Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead**. Nature Machine Intelligence, 1:206-215, 2019.

Explainable Agency: Claims

- Provide **on-demand description/justification** of decisions, (beliefs, experiences).
- Before, during, after making and executing decisions.
- Consider, evaluate, and present **alternative choices** at **different abstractions**.
- Communicated information **makes contact with human concepts** such as beliefs and goals.

Pat Langley, Ben Meadows, Mohan Sridharan and Dongkyu Choi. **Explainable Agency for Intelligent Autonomous Systems**. In Innovative Applications of Artificial Intelligence, 2017.

Pat Langley. **Explainable, Normative, and Justified Agency**. AAAI Conference on Artificial Intelligence, 2019.

Theory of Explanations

- Theory: claims, representation, processes.
- Claims about representing, reasoning with, learning knowledge; **criteria** for constructing descriptions.
- Three axes: **abstraction** of representation, explanation **specificity**, explanation **verbosity**.
- Methodology for constructing descriptions.

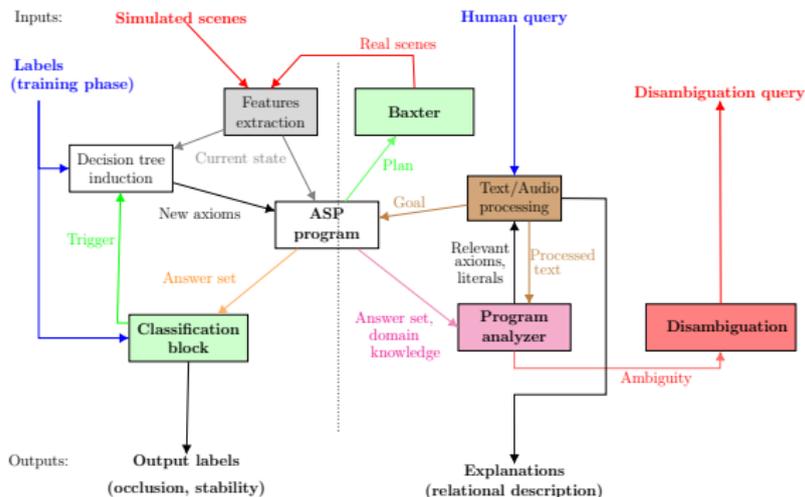
Mohan Sridharan. **Integrated Knowledge-based Reasoning and Data-driven Learning for Explainable Agency in Robotics**. In David Aha and Silvia Tulli (eds.), *Explainable Agency in Artificial Intelligence: Research and Practice*, CRC Press, 2024.

Mohan Sridharan and Ben Meadows. **Towards a Theory of Explanations for Human-Robot Collaboration**. In *Künstliche Intelligenz Journal*, 33(4):331-342, December 2019.

Explainable Agency: Approach

- **Focus:** KR tools for **transparent** reasoning and learning.
- Relational descriptions of **decisions**, **beliefs**, and **experiences** in terms of domain+agent attributes, actions.
- **Abstraction**, **specificity**, **verbosity**; provide **methodology** to identify and present **relevant** information **on-demand**.
- **Good explanation**: simplicity, coherence, coverage, convey contextual information?
- Exploit complementary strengths of **non-monotonic logical reasoning**, **probabilistic reasoning**, and **deep learning**.

Reasoning + Learning + Explanation



Relational descriptions as **explanations** of decisions and belief evolution. Construct **disambiguation questions** as needed.

- Questions (**descriptive, causal, contrastive, counterfactual**) about decisions, beliefs, experiences.
- Retrieve relevant information; **belief tracing**.

Tracing Example

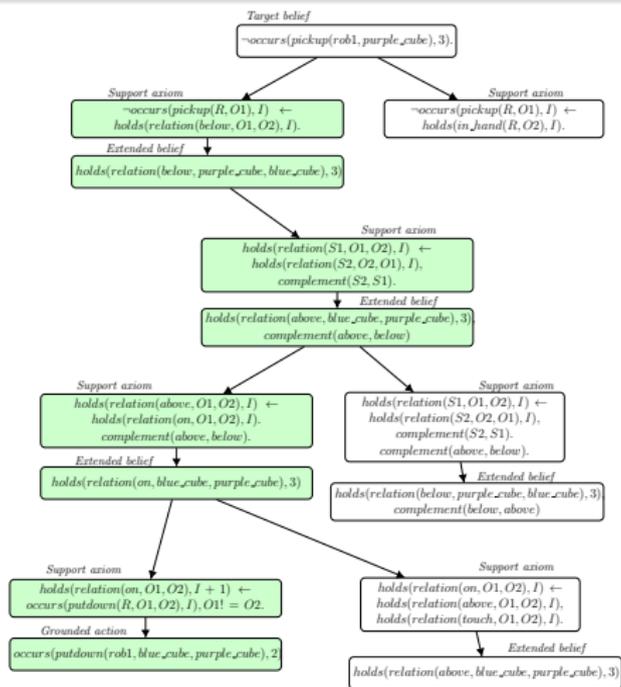
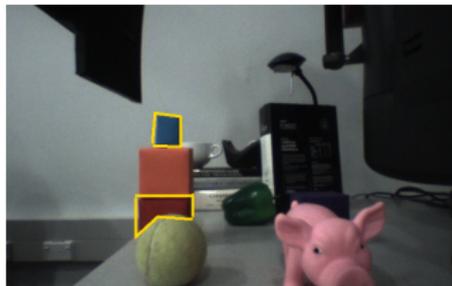


Figure: Example of belief tracing to explain non-selection of an action.

Execution Trace: Explanation

- **Goal:** red block on the top of orange block.



- **Human:** “Why did you pick up the blue block first?”;
- **Baxter:** “Because I had to pick up the red block, and it was below the blue block”;
- **Human:** “Why did you not pick up the orange block first?”;
- **Baxter:** “Because the blue block was on the orange block”;
- **Human:** “What would happen if the ball is pushed?”
- ...

Interactive Explanations (Robot Moving Books)

Human: Please describe the executed plan.

Robot: [...] I moved to the *library*. I put down the programming book on *table*₁. I picked up the robotics book from *table*₁. I moved to *office*₂. I gave the large robotics book to the *engineer*. I achieved the goal.

Human: Provide a more detailed description.

Robot: [...] I moved to cell *c*₂ in the *study*. [...] I moved to cell *c*₄ in the *library* and put down the small programming book with a hard cover on the table. I picked up the large robotics book with a soft cover by its spine from the table. [...] I moved to cell *c*₇ of the *office*₂. I gave the large robotics book to the *engineer*. I achieved the goal.

Human: Why did you move to cell *c*₄ in *library*?

Robot: I had to put the programming book on *table*₁. I know *table*₁ is in cell *c*₄ in *library*. I moved to *c*₄.

Human: Why do you believe the encyclopedia is in *office*₂?

Robot: At time 0, I assumed encyclopedia was in *library*. At time 4, I observed this was false. I inferred it was in *office*₂."

Learning + Explanations in Scene Understanding

Precision and recall for learning previously unknown axioms using decision tree induction.

Missing Axioms	Precision	Recall
Strict	69.2%	78.3%
Relaxed	96%	95.1%

(**Real scenes**) Precision and recall of **retrieving relevant literals** for explanations with and without the learned axioms for reasoning.

Query Type	Precision		Recall	
	Without	With	Without	With
Plan description	78.54%	100%	67.52%	100%
Why X?	76.29%	95.25%	66.75%	95.25%
Why not X?	96.61%	96.55%	64.04%	100%
Belief	96.67%	99.02%	95.6%	100%

Tiago Mota, Mohan Sridharan, and Ales Leonardis. **Integrated Commonsense Reasoning and Deep Learning for Transparent Decision Making in Robotics**. In Springer Nature Computer Science, 2(242), 2021

Tiago Mota and Mohan Sridharan. **Commonsense Reasoning and Deep Learning for Transparent Decision Making in Robotics**. In the European Conference on Multiagent Systems (EUMAS), Thessaloniki, Greece, September 14-15, 2020.

Explanations in AHT

- **Scenario:** bread slice inside toaster; cutlets on counter; poundcake on kitchen table; water glass in bedroom; microwave switched off; frying pan on stove (switched off); and human and ad hoc agent in kitchen.
- **Goal:** **prepare breakfast**. Plan with 23 actions; humans expected to complete some intermediate steps.
- Different types of questions posed after plan execution: descriptive, contrastive, counterfactual.

Hasra Dodamegama and Mohan Sridharan. [Explanation and Knowledge Acquisition in Ad Hoc Teamwork](#). International Symposium on Practical Aspects of Declarative Languages (PADL) at POPL, 2024.

Explanations in AHT: Interaction

Question: Why did you find bread slice in step 0?

Ad hoc Agent: Because I had not found the bread slice yet and I wanted to grab it in step 1.

Response highlights action as requirement for subsequent action.

Question: Why did you not find water glass in step 0?

Ad hoc Agent: Because I predicted human will find water glass in 0.

Agent may be asked about the human's (future) action choices.

Question: What will human do in step 1?

Ad hoc Agent: Human will grab water glass in step 1.

Question: Why will human grab water glass in step 1?

Ad hoc Agent: Because I think the human wants to bring glass to the table.

Other Considerations

- Ethics, norms, legality: can vary with **context**.
- Can model **well-defined concepts** computationally.
- Explored in different disciplines over many years.
- **AI industry benefits from subsidies based on public funds!**
- Need **regulation** and **rigour** in the design and use of AI (robot) systems.

What are Trust and Ethics?

- Many different definitions, models, surveys!
- Trust: psychological state; accept **vulnerability** based on **positive expectation** of the intention/behavior of another.
- Ethics: knowledge describing **how a person should act** in particular situations.
- Society's **ethical principles** determine what it means to be **trustworthy**; determine whom or what to trust.
- Trust enables **cooperation** and supports social norms, leading to **societal benefits**.

Benjamin J. Kuipers. **Trust and Cooperation**. *Frontiers in Robotics and AI*, 9:1-18, 2022.

Benjamin J. Kuipers. **How Can We Trust a Robot?** *Communications of the ACM*, 61(3):86-95, 2018.

The Danger of Bad Models!

- **Prisoner's Dilemma**: cooperate or defect?

	C	D
C	-1, -1	-4, 0
D	0, -4	-3, -3

- More recent blunder: the **Moral Machine** experiment.
- Design (representation, processing) choices crucial!
- Not modeling **important factors** can lead to **arbitrary predictions** in unusual situations.
- Existential threats: Climate change? AGI? Human misuse?

Benjamin J. Kuipers. **Trust and Cooperation**. *Frontiers in Robotics and AI*, 9:1-18, 2022.

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, et al. **The Moral Machine Experiment**. *Nature* 563, 59-64, 2018.

Summary of Key Ideas

- Leverage fundamental principles: **refinement, ecological rationality, abstraction, interactive learning, explainable agency.**
- **Step-wise refinement** simplifies design and implementation, **increases confidence** in behavior, **promotes scalability.**
- **Separation** of domain-independent/specific knowledge. Designer follows pre-defined steps; **otherwise automated.**
- Represent, reason, act, learn **jointly** at **different abstractions** using **different processes.**
- **Ecological rationality** and **explainable agency** for runtime adaptation, transparency, and trust.

That's all folks!