Explainable Agency in Integrated Cognitive Systems

Advanced Course at ESSAI 2024

Mohan Sridharan
Chair in Robot Systems
School of Informatics, University of Edinburgh (UK)
m.sridharan@ed.ac.uk
`https://homepages.inf.ed.ac.uk/msridhar/`

July 15-19, 2024

Objectives

- **Advanced course** at the intersection of multiple topics.
  - Knowledge-based reasoning.
  - Data-driven learning.
  - Integrated cognitive systems.
  - Explainable agency.

- **Scope** of this course:
  - Non-monotonic logic, probability theory.
  - Machine learning, reinforcement learning, deep learning.
  - Robots that sense, reason, act, learn.
  - Relational descriptions of decisions; theory of mind.

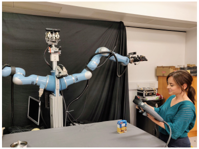
- **Format:** interactive, discussions, examples.

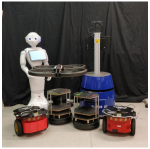
Tentative Outline

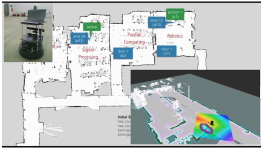
1. (L1) Knowledge representation and reasoning (KRR) I.
2. (L2) KRR II and learning.
3. (L3) KRR, learning, and control.
4. (L4) KRR, teamwork, and learning.
5. (L5) Explanations, integrated systems, closing the loop.

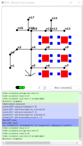
Illustrative Domain: Robot Assistants

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.

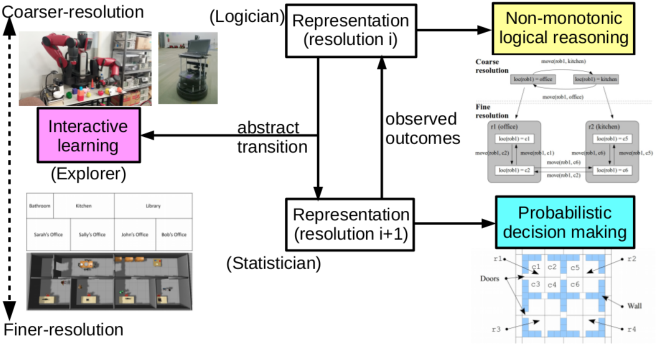
Shiqi Zhang and Mohan Sridharan. **A Survey of Knowledge-based Sequential Decision Making under Uncertainty**. Artificial Intelligence Magazine, 43(2):249-266, 2022.

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



Coarser-resolution

(Logician)
Representation (resolution i)

Non-monotonic logical reasoning

Interactive learning
(Explorer)

abstract transition

observed outcomes

Representation (resolution i+1)
(Statistician)

Probabilistic decision making

Finer-resolution

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**. In the 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**. In Journal of Artificial Intelligence Research, 65:87-180, May 2019.

Reasoning + Learning: Motivation

- Machine (deep?) learning widely used in AI and robotics.
- Limitations of deep network architectures:
    - 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; Answer questions (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.

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.

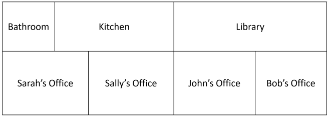
Gerd Gigerenzer. **What is Bounded Rationality?** Routledge Handbook of Bounded Rationality, Riccardo Viale (editor), Routledge, 2021.
Konstantinos Katsikopoulos, Ozgur Simsek, Marcus Buckmann and Gerd Gigerenzer. **Classification in the Wild: The Science and Art of Transparent Decision Making**. MIT 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

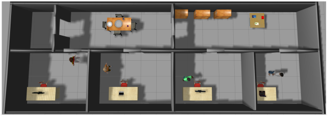
- Ecological rationality: satisficing with adaptation.
  - Algorithmic model of heuristics.
  - Competitive testing of predictions.
- Heuristics: ignore some information to make decision more quickly, frugally, and/or accurately.
  - One-reason (hiatus); sequential-search (take the best), tallying; fast 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!

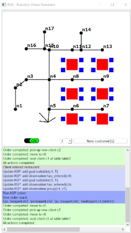
Ian N. Durbach, Simon Algorta, Dieudonne Kabongo Kantu, Konstantinos V. Katsikopoulos, and Ozgur Simsek. **Fast and Frugal Heuristics for Portfolio Decisions with Positive Project Interactions**. Decision Support Systems, 138, 2020.
Nadine Fleischhut and Gerd Gigerenzer. **Can Simple Heuristics Explain Moral Inconsistencies?** Simple Heuristics in a Social World, R. Hertwig, U. Hoffrage, and ABC group (eds.), Oxford University Press, 2013.

Scene Understanding + Planning

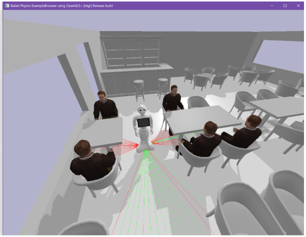


Bathroom
Kitchen
Library
Sarah’s Office
Sally’s Office
John’s Office
Bob’s Office





```
Order completed: pick-up new client c2
Order completed: move to n6
Order completed: seat client c1 at table table2
All actions completed
Client entered restaurant
Update ASP: add goal isattable(c4, T)
Update ASP: add observation has_entered(c4)
Update ASP: add goal isattable(c5, T)
Update ASP: add observation has_entered(c5)
Update ASP: add observation group(c4, c5)
Run ASP solver
New order stack:
['go_to(agent,n5)', 'pick(agent,c5)', 'go_to(agent,n6)', 'seat(agent,c5,table1)']
Order completed: move to n5
Order completed: pick-up new client c5
Order completed: move to n6
Order completed: seat client c5 at table table1
All actions completed
```



Bullet Physics ExampleBrowser using OpenGL3+ [btgl] Release build

Begin with a (very) brief overview of ML...

Detour: Machine Learning (Classification)

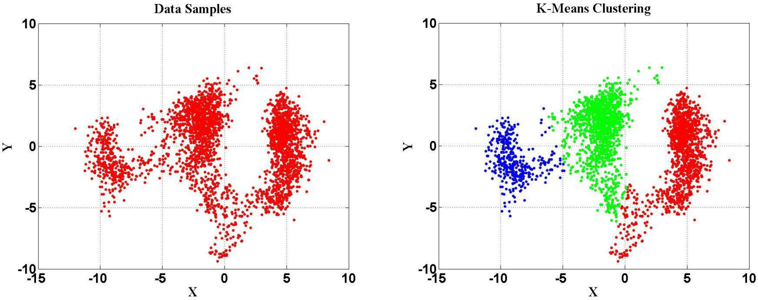
- Broad categories: supervised (labeled samples); unsupervised (no labeled samples).
- Group data based on similarity measures.

- Many sophisticated methods:
  - Supervised: decision trees, support vector machines, neural networks.
  - Unsupervised: nearest neighbors, clustering.

- Choice of classifier based on data and application.
- Probabilistic methods model the noise in input data; frequentist or Bayesian?

Clustering Data Samples

- Clustering of input data samples.
- Data grouped into three clusters.



Bayesian Classification

- Bayes' rule:

$$p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x)$$
$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)} = \frac{\text{likelihood . prior}}{\text{normalizer}}$$

- Classify based on Bayes decision rule:

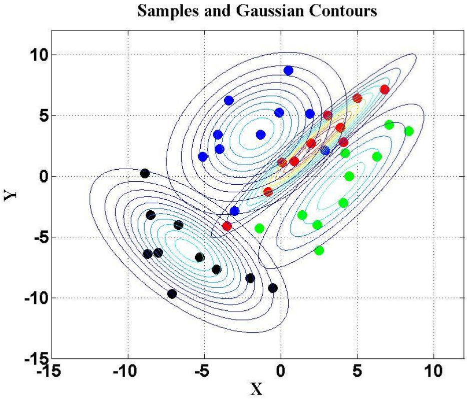
$$p(C_1|x) > p(C_2|x) \implies \text{choose } C_1\text{; else choose } C_2$$

- Decision rule extends to multiple classes:

$$p(C_i|x) > p(C_j|x) \;\; \forall j \neq i \implies \text{choose } C_i$$

Illustrative Example

- Four-class problem; ten training data samples per class.
- Model individual class likelihoods as Gaussians.



Illustrative Example: Modeling

- Compute Gaussian means and covariances:

$$\mu_1 = [2.16, 2.49]; \quad \mu_2 = [3.95, -0.84]$$
$$\mu_3 = [-1.57, 3.5]; \quad \mu_4 = [-6, -6.14]$$

$$\Sigma_1 = \begin{pmatrix} 9.32 & 10.12 \\ 10.12 & 11.85 \end{pmatrix}$$

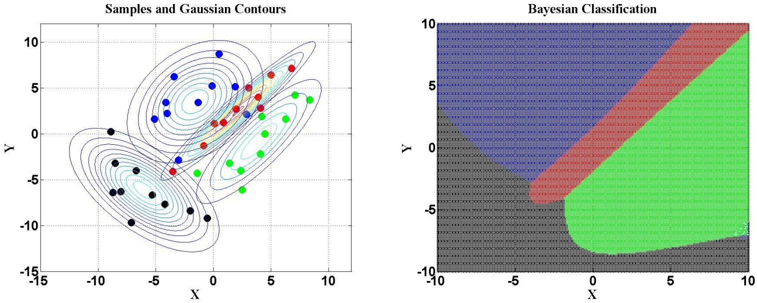
$$\Sigma_2 = \begin{pmatrix} 8.36 & 8.87 \\ 8.87 & 13.02 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 7.63 & 2.98 \\ 2.98 & 9.78 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 8.62 & -5.71 \\ -5.71 & 9.26 \end{pmatrix}$$

Illustrative Example: Classification Result

- Decision boundaries for all four classes:



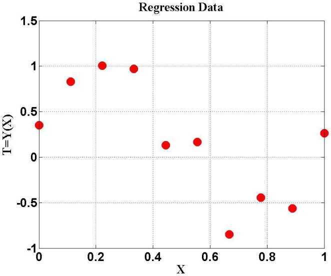
- What about real-valued outputs?

Detour: Machine Learning (Regression)

- Consider polynomial curve fitting of target variable t:

$$t = y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

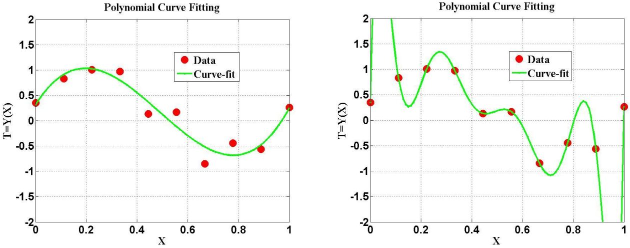
- Consider data sampled from a sinusoidal waveform:



- Can use polynomials of different degrees.

Illustrative Example

- Polynomial curve fitting: best performance for degree $= 3$.



- However *over-fitting* can lead to problems.
- Bayesian view of regression?

Regularization, Basis Functions

- Regularization in sum-of-squares error function:

$$\begin{aligned} E(\boldsymbol{w}) &= E_D(\boldsymbol{w}) + \lambda E_w(\boldsymbol{w}) \\ &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - y(x_n, \boldsymbol{w})\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \end{aligned}$$

- Regularization coefficient to minimize over-fitting; implemented in toolboxes.
- Model curve fitting using basis functions:

$$t = y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x})$$

Bayesian Regression

- Gaussian noise model:

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$
$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), 1/\beta)$$

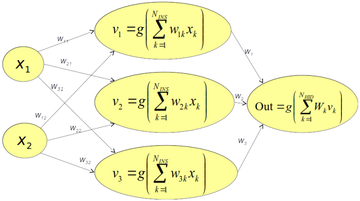
- Log likelihood and parameter estimation ($\boldsymbol{w}_{ML}$, β_{ML}):

$$ln\, p(\boldsymbol{t}|\boldsymbol{w}, \beta) = \frac{N}{2} ln(\beta) - \frac{N}{2} ln(2\pi) - \beta E_D(\boldsymbol{w})$$
$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{N} \{t_i - \boldsymbol{w}^T \phi(x_i)\}^2$$

- Frequentist (data determines model): maximize $p(\mathcal{D}|\boldsymbol{w})$.
- Bayesian (consider prior beliefs):

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})\, p(\boldsymbol{w})}{p(\mathcal{D})}$$

Detour: Neural Networks


x_1
x_2
w_{11}
w_{21}
w_{31}
w_{12}
w_{22}
w_{32}
$v_1 = g\left(\sum_{k=1}^{N_{INS}} w_{1k} x_k\right)$
$v_2 = g\left(\sum_{k=1}^{N_{INS}} w_{2k} x_k\right)$
$v_3 = g\left(\sum_{k=1}^{N_{INS}} w_{3k} x_k\right)$
W_1
W_2
W_3
$\text{Out} = g\left(\sum_{k=1}^{N_{HID}} W_k v_k\right)$

- Complex functions for classification, regression.

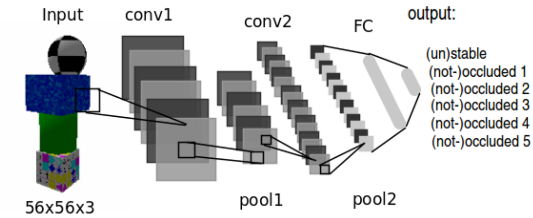
$$y = g\Big(\sum_j W_j g(\sum_k w_{jk} x_{jk})\Big)$$

- Forward projection of outputs; backpropagation of derivative of error to revise weights:

$$E = \sum_{k=1}^{N} (y_k - \mathbf{w}^T \mathbf{x}_k)^2$$

$$w_j = w_j - \eta \frac{\partial E}{\partial w_j} = w_j + 2\eta \sum_{k=1}^{N} \delta_k x_{kj}$$

Detour: Deep Neural Networks

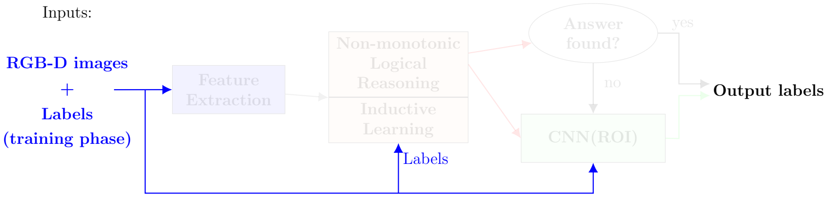


- Extends the notion of neural networks to many layers.
- Many architectures with different connections: CNN, LSTM, transformers, skip connections ...

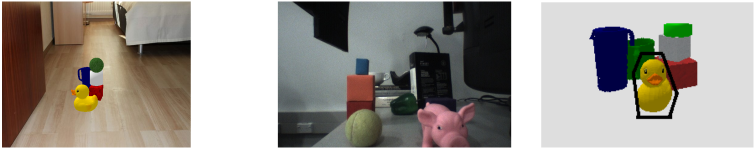
- Cognition is not just function approximation; it is a reasoning and learning problem!

Pat Langley. **The Central Role of Cognition in Learning**. Advances in Cognitive Systems, 4:3-12, 2016.

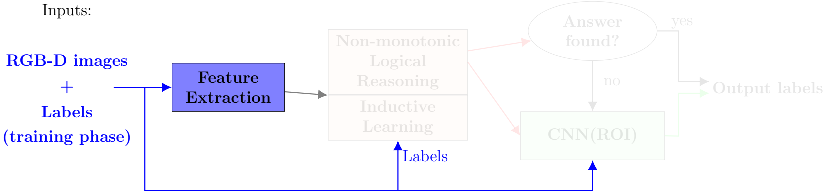
Architecture Components: Input



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



Architecture Components: Feature Extraction

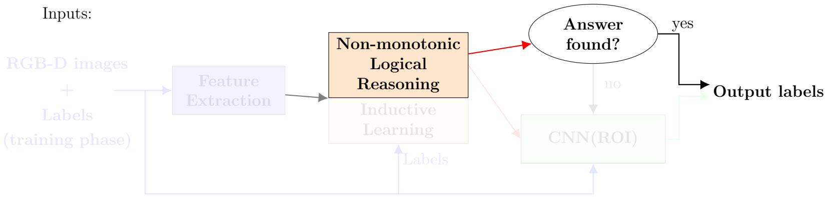


Geometric features extracted from images:

- 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

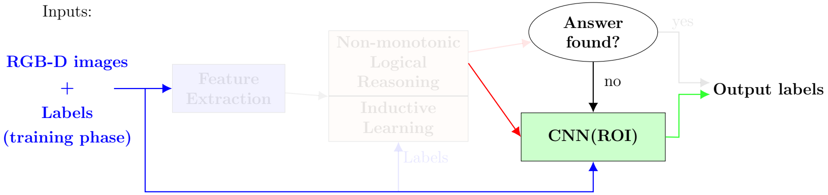


- Input: Extracted features, incomplete domain dynamics.
- ASP for non-monotonic logical reasoning.

$$stable(A) \leftarrow not\ obj\_rel(above, A, B)$$
$$\neg occurs(pickup(rob_1, O_1), I) \leftarrow holds(obj\_rel(below, O_1, O_2), I)$$

- Decision about input image if possible.

Architecture Components: CNN

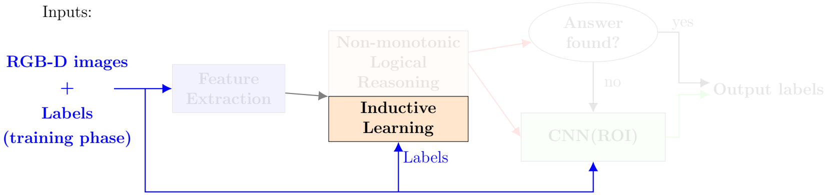


- Attention: ROI selection based on axioms.

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

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

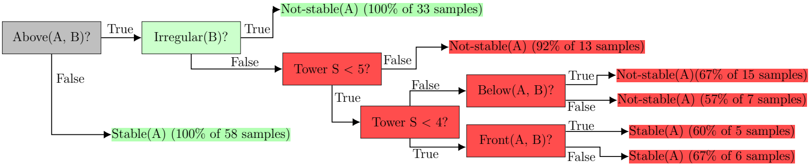
Architecture Components: Inductive Learning



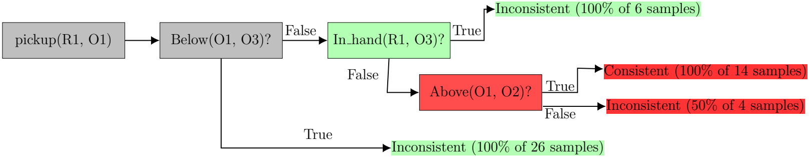
- Input: features and figure labels.
- Decision Tree: induction of rules (constraints, causal laws).
- Output: learned rules.

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



Above(A, B)?
True
Irregular(B)?
True
Not-stable(A) (100% of 33 samples)
False
Tower S < 5?
False
Not-stable(A) (92% of 13 samples)
True
Tower S < 4?
False
Below(A, B)?
True
Not-stable(A)(67% of 15 samples)
False
Not-stable(A) (57% of 7 samples)
True
Front(A, B)?
True
Stable(A) (60% of 5 samples)
False
Stable(A) (67% of 6 samples)
False
Stable(A) (100% of 58 samples)



pickup(R1, O1)
Below(O1, O3)?
False
In_hand(R1, O3)?
True
Inconsistent (100% of 6 samples)
False
Above(O1, O2)?
True
Consistent (100% of 14 samples)
False
Inconsistent (50% of 4 samples)
True
Inconsistent (100% of 26 samples)

$$\neg stable(A) \leftarrow obj\_rel(above, A, B),\ obj\_surface(B, irregular)$$

$$\neg pickup(R, O1) \leftarrow in\_hand(R, O2)$$

Mohan Sridharan and Tiago Mota. **Towards Combining Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning**. In Journal of Autonomous Agents and Multi-Agent Systems, 37(4), 2023.

Learn from Human Verbal Input

- **Assumptions:**
  - Humans provide accurate descriptions.
  - Other robots have same/similar capabilities.
  - Learner can process sensor inputs.
- Verbal cue: "that robot is labeling fairly big textbook":
  - Part-of-speech (POS) tagging; match with images.
- Construct causal laws.

$$label(rob_1, book_1)\ \ \textbf{causes}\ \ labeled(book1)$$

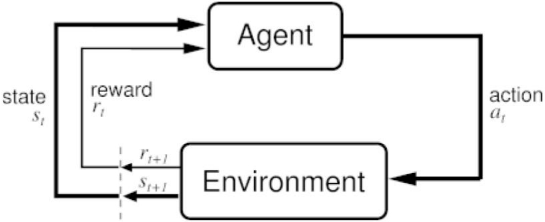
generalize ("robot labeled small fragile cup"):

$$label(R, O)\ \ \textbf{causes}\ \ labeled(O)$$

Learning from Experience

- Active exploration or unexpected (reactive) transition. Identify state action combinations.
- Formulate as reinforcement learning problem.
- Represent experiences relationally (binary decision tree); cumulative learning.
- Relevance and relational inference guide learning.
- Reason with experience to construct new axioms.
- A (very) brief overview of RL...

Brief Detour: Reinforcement Learning



- Underlying model is Markov Decision Process (MDP).
- Tuple $\langle S, A, T, R\rangle$; policy $\Pi^* : S \mapsto A$.
- Search problem: value iteration, policy iteration.
- RL problem: model (T, R) unknown; (s, a, s', r) examples.

RL Threads and Solutions

- Three threads of RL:
    - Trial and error: psychology.
    - Dynamic programming: stochastic optimal control.
    - Temporal difference methods: computer science.

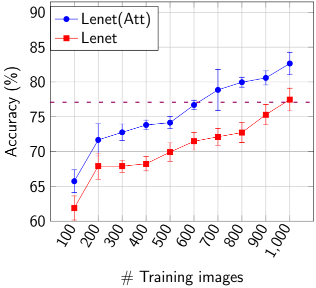
- Model-based learning: estimate T, R from examples, solve underlying MDP (probabilistic search problem).
- Model-free learning: directly compute values, policy from acquired experiences.

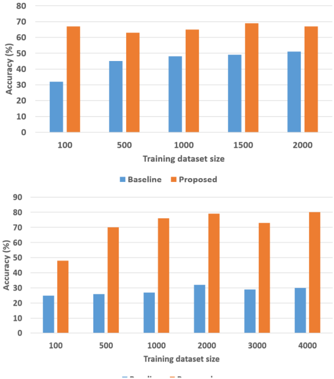
- Many challenges: credit assignment, reward shaping.
- Feature abstraction essential in practical problems.

Richard Sutton and Andy Barto. **Reinforcement Learning: An Introduction**. MIT Press, 2018.

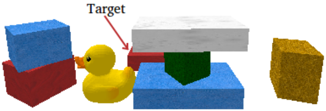
Return to Scene understanding

- Accuracy increases and training complexity decreases.



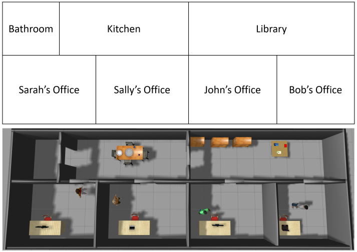


Experimental Results: VQA + Decision making



- **Initially**: 64 plans; most incorrect or sub-optimal.

- **Including learned axioms**: 3 correct plans.



- **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:

$$\begin{gathered} move(rob_1, kitchen), \;\; grasp(rob_1, cup_1) \\ move(rob_1, office), \;\; putdown(rob_1, cup_1) \end{gathered}$$

- 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):

$$
\begin{array}{ll}
move(rob_1, c_3) & \\
test(rob_1, loc(cup_1), c_3) & \%\ cup_1 \text{ not observed} \\
move(rob_1, c_5) & \\
test(rob_1, loc(cup_1), c_5) & \%\ cup_1 \text{ observed} \\
grasp(rob_1, cup_1) &
\end{array}
$$

- Grasping cup_1 fails; relational learning:

$$
\textbf{impossible}\ \ grasp(rob_1, C)\ \ \textbf{if}\ \ arm\_type(rob_1, electromagnetic),\ obj\_weight(C, heavy)
$$

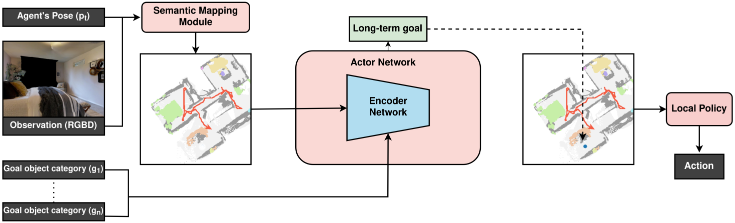
Robot Waiter: Reasoning (Video)

Example

Robot Waiter: Reasoning + Learning (Video)

Example

Using Learned Knowledge I: Overview



- Semantic mapping: extract more abstract concepts.
- Reward specification: non-procrastination, trade-offs.

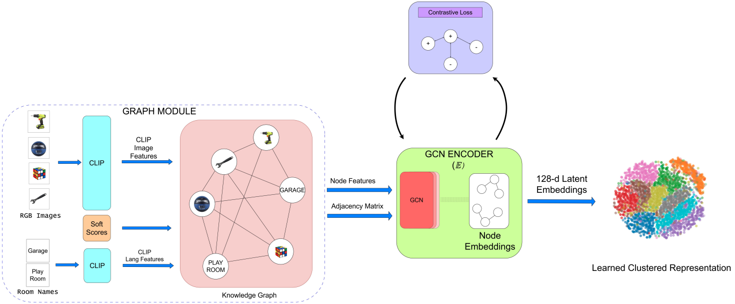
N. Gireesh, A. Agrawal, A. Datta, S. Banerjee, M. Sridharan, B. Bhowmick, and M. Krishna. **Sequence-Agnostic Multi-Object Navigation**. IEEE International Conference on Robotics and Automation (ICRA), May 2023.

Using Learned Knowledge I: 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.

Using Learned Knowledge II: Overview



- Multimodal inputs: extract concise embeddings.
- Knowledge graph: represent and use prior knowledge.

A. Agrawal, R. Arora, A. Datta, S. Banerjee, B. Bhowmick, K.M. Jatavallabhula, M. Sridharan, and M. Krishna. **CLIPGraphs: Multimodal Graph Networks to Infer Object-Room Affinities**. In the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), August 2023.

Using Learned Knowledge II: Video

CLIP Example

A. Agrawal, R. Arora, A. Datta, S. Banerjee, B. Bhowmick, K.M. Jatavallabhula, M. Sridharan, and M. Krishna. **CLIPGraphs: Multimodal Graph Networks to Infer Object-Room Affinities**. In the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), August 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**. In the IEEE International Conference on Robotics and Automation (ICRA), May 2024.

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?

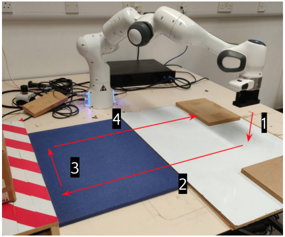
Changing-Contact Manipulation: Video

Changing contact manipulation

Saif Sidhik, Mohan Sridharan, and Dirk Ruiken. **Towards a Framework for Changing-Contact Robot Manipulation**. In the International Conference on Intelligent Robots and Systems (IROS), 2021.

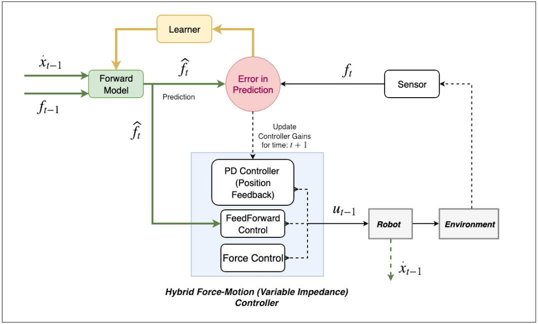
Michael Mathew, Saif Sidhik, Mohan Sridharan, Morteza Azad, Akinobu Hayashi, and Jeremy Wyatt. **Online Learning of Feed-Forward Models for Task-Space Variable Impedance Control**. In the International Conference on Humanoid Robots (Humanoids), 2019.

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.

$$\boldsymbol{u}_t = \boldsymbol{H}_t + \mathbf{K}_t^{\mathbf{p}} \Delta \boldsymbol{x}_t + \mathbf{K}_t^{\mathbf{d}} \Delta \dot{\boldsymbol{x}}_t + \boldsymbol{u}_t^{\mathbf{fc}} + \boldsymbol{u}_t^{\mathbf{ff}}$$

Changing-Contact Manipulation: Video

Changing contact manipulation

Task and Motion Planning with Dynamics

- Responding to discontinuous interaction dynamics; combine with task planning and motion planning.
- Task planning: discrete, abstract action; motion planning: continuous-space motion.
- Combine different representations and different update processes. Three strategies and open problems.
- What about multiple agents?

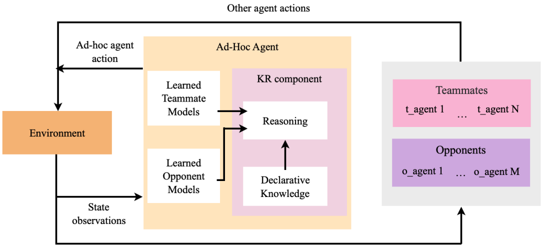
Collaboration without Prior Coordination (AHT)

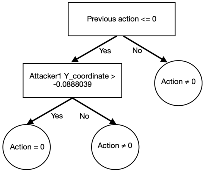
Example Example

- 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.

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.

$$move^*(Ag, X, Y)\ \textbf{causes}\ in^*(Ag, X, Y)$$
$$\neg in^*(Ag, X1, Y1)\ \textbf{if}\ in^*(Ag, X2, Y2),\ X1 \neq X2, Y1 \neq Y2$$
$$\textbf{impossible}\ shoot(Ag, Ago)\ \textbf{if}\ agent\_shot(Ago)$$
$$\textbf{initial default}\ spread\_attack(Ago)\ \ \textbf{if}\ \ attacker(Ago)$$
$$in(Ag, R)\ \ \textbf{if}\ \ in^*(Ag, X, Y),\ component(X, Y, R)$$

Hasra Dodampegama 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.
- Adaptation to different teammate and opponent types.
- Orders of magnitude fewer examples (5000 vs. 1M).
- 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 Dodampegama and Mohan Sridharan. **Knowledge-based Reasoning and Learning under Partial Observability in Ad Hoc Teamwork**. In Theory and Practice of Logic Programming, 2023.

AHT Results: “Embodied AI”

Embodied AI1

Architecture	**Steps**	**Time**
REACT	0.89 ± 0.11	0.90 ± 0.19
Baseline	1 ± 0.05	1 ± 0.04

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?

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.

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

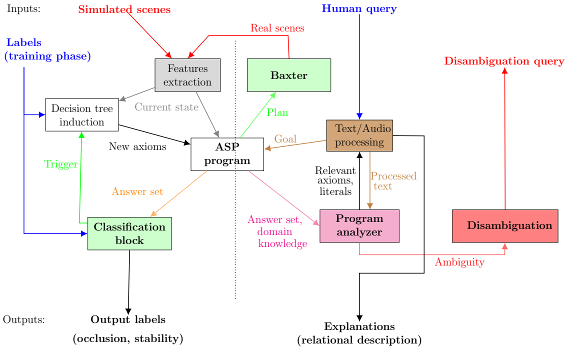
- Theory: claims, representation, processes.
- Claims about representing, reasoning with, learning knowledge; criteria for constructing descriptions.
- Three axes: abstraction of representation, explanation specificity, explanation verbosity.
- Methodology for constructing descriptions.

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.

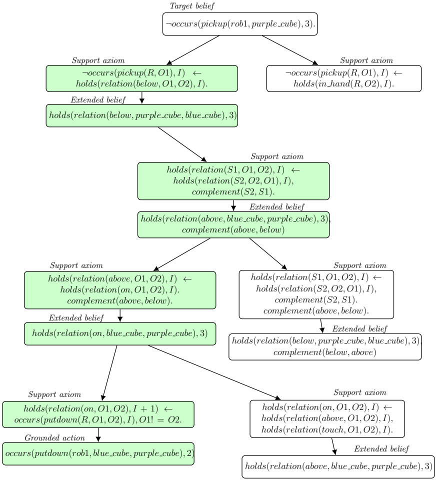
Reasoning + Learning + Explanation



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



Target belief

$\neg occurs(pickup(rob1, purple\_cube), 3).$

Support axiom

$\neg occurs(pickup(R, O1), I) \leftarrow holds(relation(below, O1, O2), I).$

Support axiom

$\neg occurs(pickup(R, O1), I) \leftarrow holds(in\_hand(R, O2), I).$

Extended belief

$holds(relation(below, purple\_cube, blue\_cube), 3)$

Support axiom

$holds(relation(S1, O1, O2), I) \leftarrow holds(relation(S2, O2, O1), I), complement(S2, S1).$

Extended belief

$holds(relation(above, blue\_cube, purple\_cube), 3), complement(above, below)$

Support axiom

$holds(relation(above, O1, O2), I) \leftarrow holds(relation(on, O1, O2), I). complement(above, below).$

Extended belief

$holds(relation(on, blue\_cube, purple\_cube), 3)$

Support axiom

$holds(relation(S1, O1, O2), I) \leftarrow holds(relation(S2, O2, O1), I), complement(S2, S1). complement(above, below).$

Extended belief

$holds(relation(below, purple\_cube, blue\_cube), 3), complement(below, above)$

Support axiom

$holds(relation(on, O1, O2), I + 1) \leftarrow occurs(putdown(R, O1, O2), I), O1 ! = O2.$

Grounded action

$occurs(putdown(rob1, blue\_cube, purple\_cube), 2)$

Support axiom

$holds(relation(on, O1, O2), I) \leftarrow holds(relation(above, O1, O2), I), holds(relation(touch, O1, O2), I).$

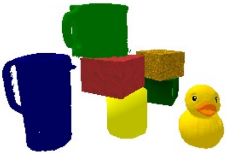
Extended belief

$holds(relation(above, blue\_cube, purple\_cube), 3)$

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

Constructing Disambiguation Queries

- Human command/query many be ambiguous: "move the yellow object on the green cube".

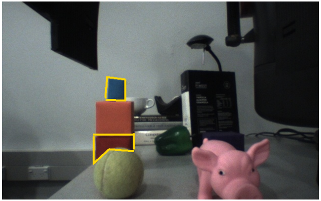


- Introduce heuristic measures of ambiguity, human confusion, attribute rank.
- Robot selects attributes to construct disambiguation queries, e.g., "are you referring to the yellow duck?"

Tiago Mota and Mohan Sridharan. **Answer me this: Constructing Disambiguation Queries for Explanation Generation in Robotics**. In the International Conference on Development and Learning, 2021.

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 *ofice*$_2$. 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 *ofice*$_2$. 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*$_2$."

Revisit Robot Waiter (Video)

Example

- Robot can be questioned about executed plan.
  Human: “why did you seat customers at *table*$_2$?”
  Pepper: “Customers wanted to sit together and *table*$_2$ was the closest available table.”

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.

	Precision		Recall	
Query Type	**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 Dodampegama 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.

Summary of Key Ideas

- 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.
- Non-monotonic logical reasoning, probabilistic reasoning, and interactive learning inform and guide each other.
- Predictive models provide run-time adaptation.
- Interactive explanations constructed efficiently on demand.

That’s all folks!