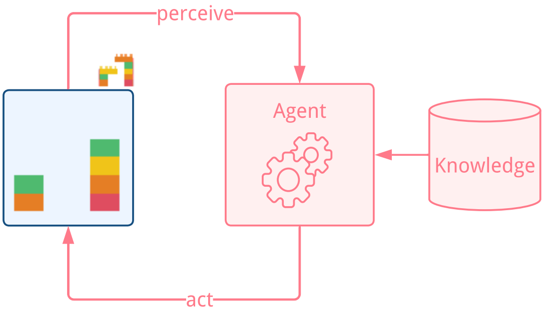
Radical Agents that Reason, Learn, & Collaborate

Shiwali Mohan
shiwali.mohan@gmail.com

February 14, 2025

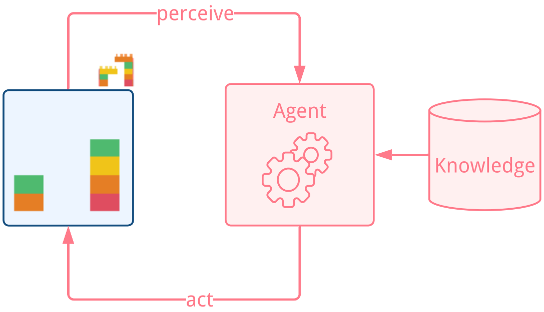
Classical Agents: A Primer



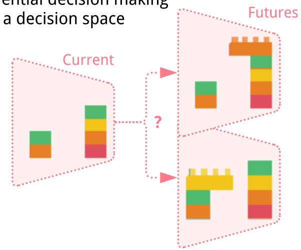
Challenges: brittle logic-based inference, machine language interaction

Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Pearson.

Classical Agents: A Primer



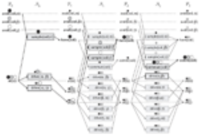
sequential decision making
over a decision space



Rule-based

```
1 if if condition then
2 | something if ;
3 else if elseif condition then
4 | something elseif ;
5 else
6 | something else ;
```

Model-based



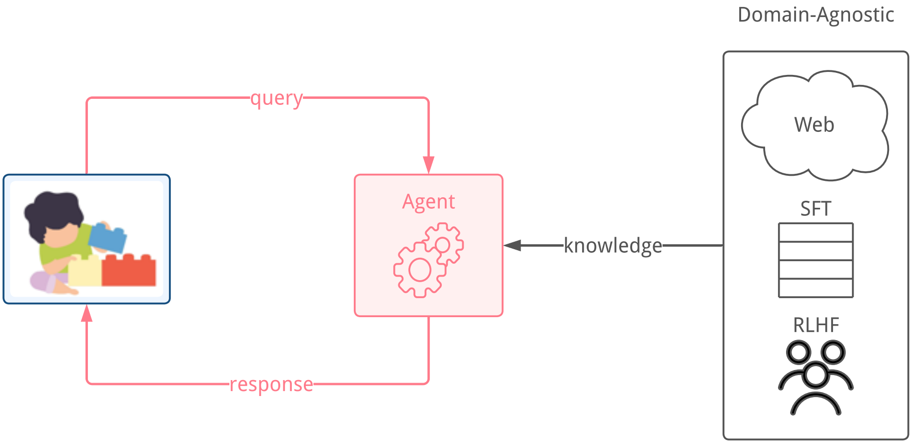
RL

$\pi^*(s) = \arg\max_a Q^*(s, a)$

Challenges: brittle logic-based inference, machine language interaction

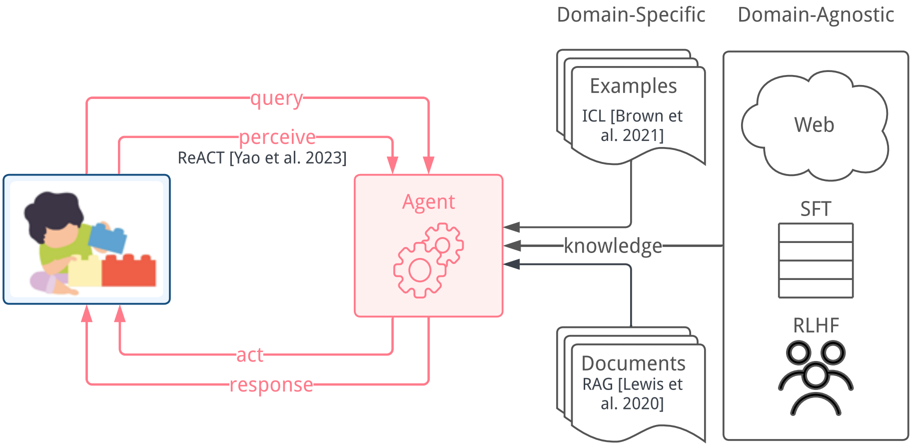
Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Pearson.

Generative AI Agents*



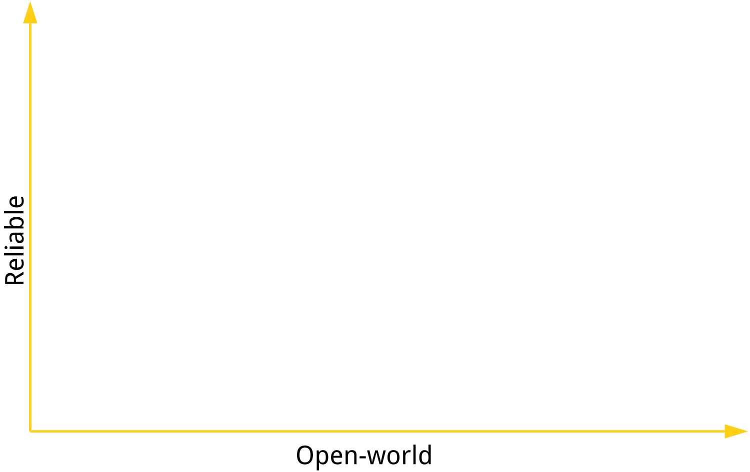
*LLM-only; may not aply to LVLMs [LLaVa: Liu et al. 2023] and LAMs [Wang et al. 2025]

Generative AI Agents*



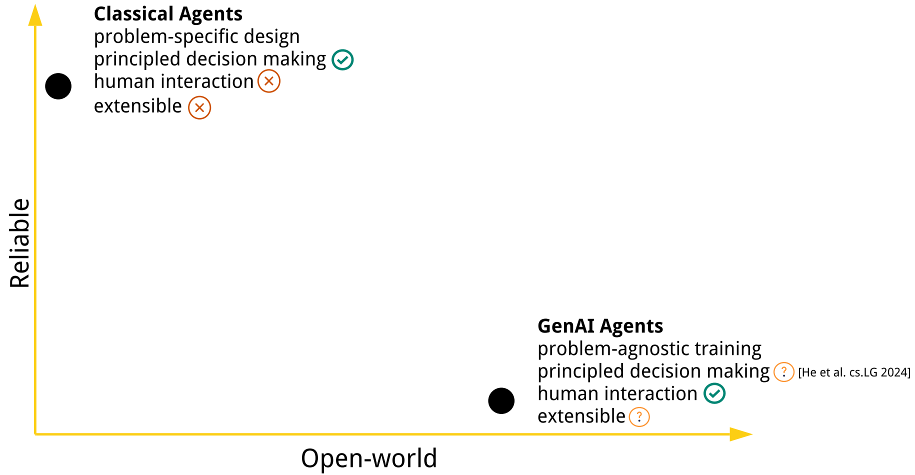
*LLM-only; may not aply to LVLMs [LLaVa: Liu et al. 2023] and LAMs [Wang et al. 2025]

Tradeoffs in Agent Design Space

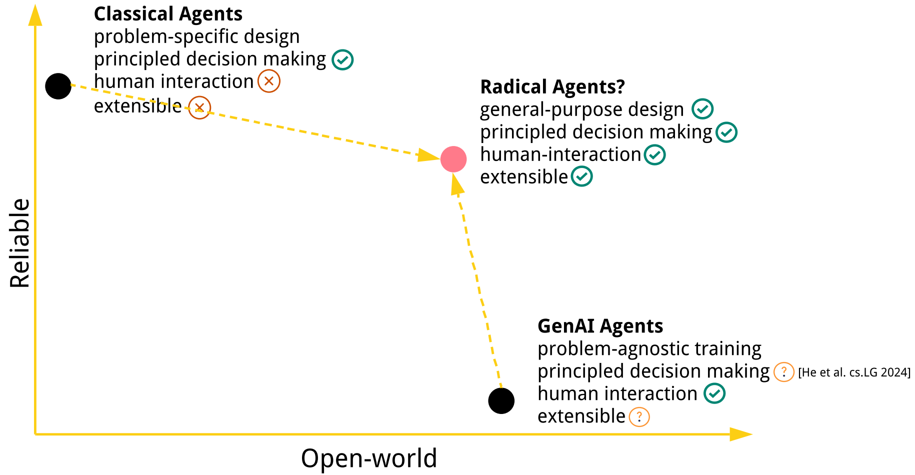


Tradeoffs in Agent Design Space

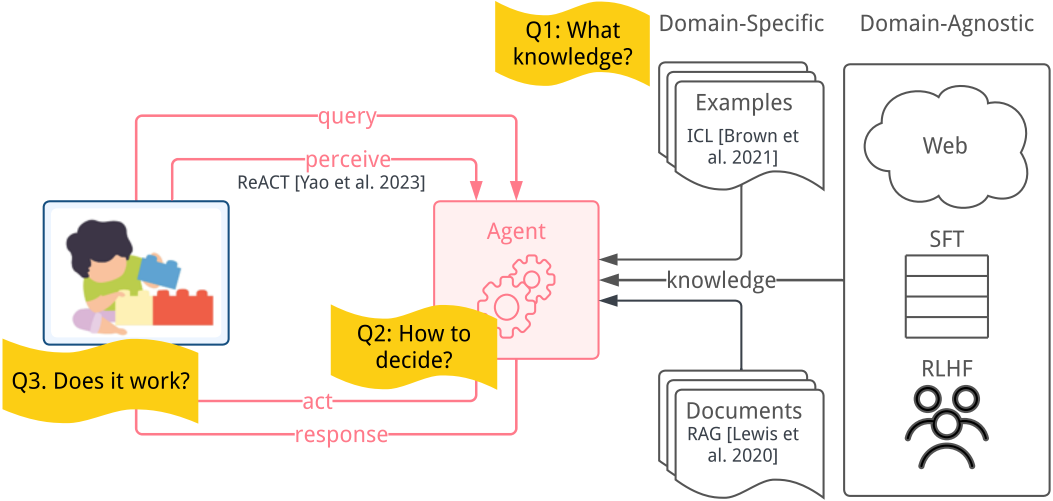
Tradeoffs in Agent Design Space



Tradeoffs in Agent Design Space



Approach

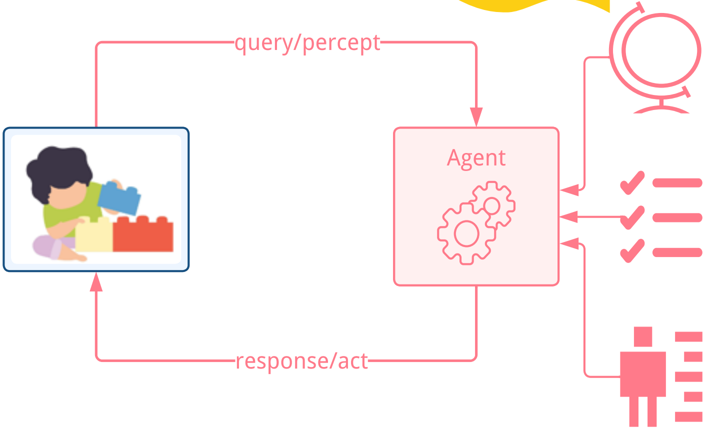

Q1: What knowledge?
Domain-Specific
Domain-Agnostic
Examples
ICL [Brown et al. 2021]
Web
query
perceive
ReACT [Yao et al. 2023]
Agent
SFT
knowledge
Q2: How to decide?
Q3. Does it work?
act
response
RLHF
Documents
RAG [Lewis et al. 2020]

Approach

Q1: What knowledge?

*not neural network models

Sets up the decision space



World models

What futures are possible?

$o^{t} \rightarrow s^{t}, s^{t}, e^{t} \rightarrow st+1$

[Johnson-Laird CS1980]
[Garrido et al. CVPR2024]
[LLaVa: Liu et al. 2023]
[LAM: Wang et al. 2025]

Task models

Which futures can/should I achieve?

parameters, goals, soft & hard constraints, rewards

[Newell AP1979]
[Weld AIM1994]
[Kamar and Horvitz, AAMAS2015]

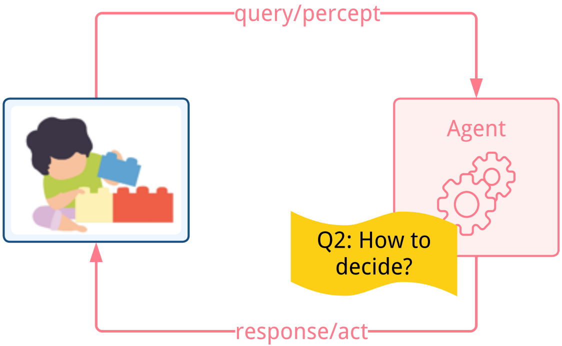
Human models

What is desirable to the user?

needs, beliefs, desires, prefrences,

[Grosz and Sidner, CL1986]
[Kamar et al., AAMAS2012]
[Wilder et al., IJCAI 2020]
[MAIA: Tange et al., NeurIPS 2024]

Approach



Reasoning Models
Orca-2 [Mitra et al. 2023]
DeepSeek-r1 [Guo et al. 2025]
..

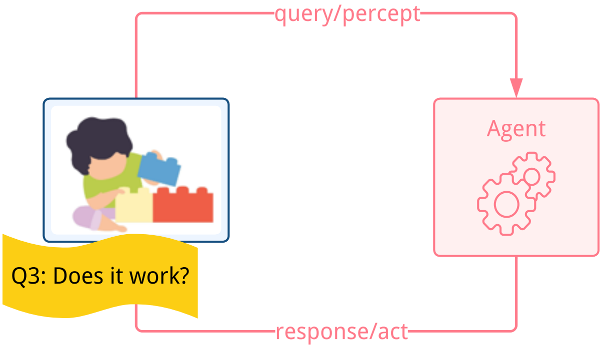
Post-training Control
Chain-of-thought [Wei et al. NIPS 2022]
Reflexion [Shinn et al. NIPS 2024]
....

Math word problems

Sequential decision making?
- generate space of futures
- constrain based on capabilities, action properties ...
- bias by human beliefs, needs, preferences

More research needed
1. Validate [Verma et al. 2024]
2. Augment [LLM-modulo Kambhampati et al. 2024]
3. Blend
4. Architect

Approach



Beyond measuring accuracy on generic benchmarks

A Rigorous Evaluation Paradigm

0. Human-centric metrics [Bansal et al. AAAI2021]
1. Realistic benchmarks [Sachdeva et al. 2024]
2. Acceptability [Li et al. ToCHI2023]
3. Alignment
4. Impact measurement
    a. Observational study [Mozannar, Chen et al. 2024]
    b. Randomized control trial

Experience

World models: Piotrowski et al. ICAPS2023, Piotrowski et al. ICAPS2021
Task models: Grover and Mohan, ICAPS-D2024, Mohan et al. IUI-W2019, Mohan and Laird AAAI2011
Human models: Ramaraj et al. ROMAN2021, Mohan TiiS2021, Mohan et al. TiiS2020, Mohan et al. JAIR2019, Mohan et al. 2017

Q1: What knowledge?

query/percept

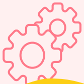
Realistic benchmarks
Rajagopal et al. HealthIUI2025

Observational/Choice Studies
Mohan TiiS2020
Mohan et al. JAIR2019
Mohan et al. AIES2019

RCT partial-factorial design
Mohan TiiS2021
Springer et al. JMIR2017
Pirolli et al. JMIR2017

Q3. Does it work?

Agent



Q2: How to decide?

response/act

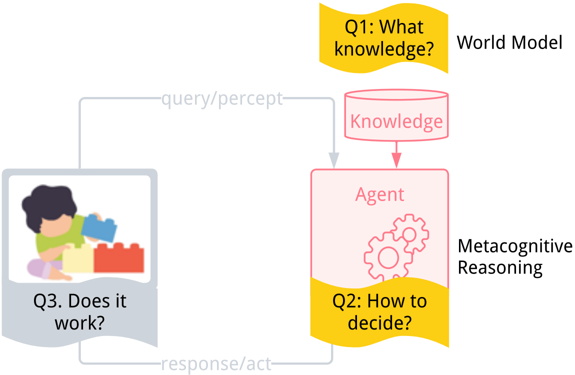
Learning Fast & Slow
Laird and Mohan AAAI2018, Blue Sky Award

Open-world Learning
Mohan et al. AIJ2024
Piotrowski et al. ICAPS-D2024
Piotrowski et al. AAMAS2023

Interactive Task Learning
Mohan et al. ACS2020
Mohan and Laird AAAI2014

Analogical Generalization
Hancock et al. JAIR2025 (in-review)

Open World Learning



Agents in Open Worlds

Agents are built with design assumptions capturing the nature of deployment

- Model-based: representation, decision process
- ML: datasets, training regime, simulations
- Deployment can diverge or evolve from design assumptions
- Resource intensive redesign or retraining

DARPA SAIL-ON with NIWC/US Navy

- 5 scientists, 2 faculty members (Penn State and Ben Gurion), students and interns
- Publications
  - Open-world learning: AIJ2024, ICAPS-D2024, AAMAS2023, AAMAS-W2023, AAAI2021, ACS2020
  - Planning: ICAPS2023, SOCS2023, AAAI2023, AAAI2022, ICAPS-D2021
  - Machine learning: CoLLA2022
- System-level invention submission
- Only team (of 12) to transition technology to US Navy/NIWC

Setup

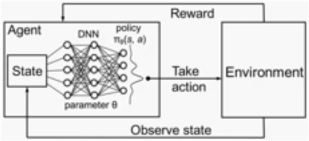
Novelty: a meaningful change in the world, a significant shift in the distribution. Examples: a new object, a new skill, a new goal, a new constraint

State of Art: deep reinforcement learning [Mnih et al. NeurIPS2013]

- Represents knowledge as undifferentiated network weights
- Fails drastically when novelty presents itself
- Requires thorough retraining

Ideal behavior: life-long, continual learning

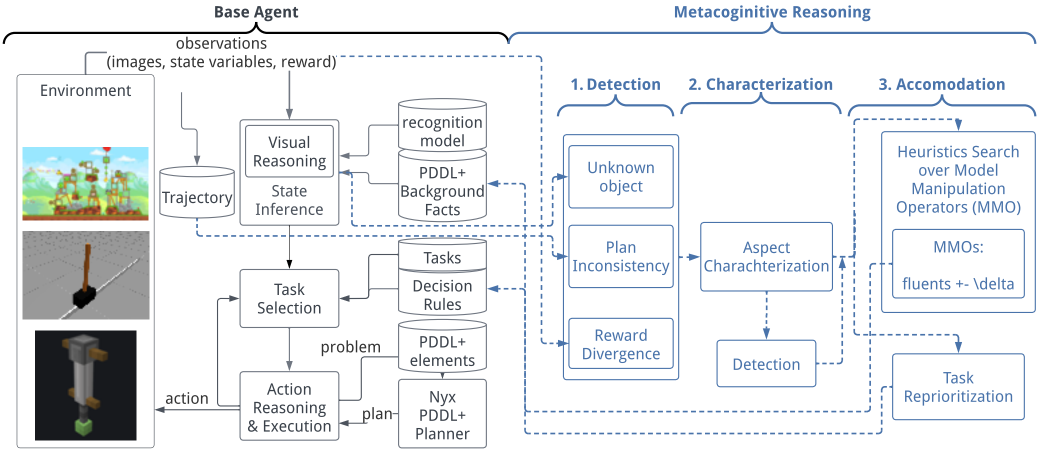
- Autonomous; require no human intervention in redesign or retraining
- Online; learn post-design, during performance
- Efficient; build upon what was known previously



HYDRA

Mohan et al. AIJ 2024; Piotrowski et al. AAMAS 2023

Integrated agent system: computer vision, planning, deep ML, goal reasoning, knowledge diagnosis & repair
Key innovations: an explicit world model and metacognitive reasoning



A Demonstration

Open world learning in Angry Birds

How does it work?

1. Explicit event model

```
(:process flying
 :parameters (?b - bird)
 :precondition (and (bird_released ?b)
                    (= (active_bird) (bird_id ?b))
                    (> (y_bird ?b) 0))
 :effect (and (decrease  (vy_bird ?b) (* #t (gravity))
              (increase   (y_bird  ?b) (* # (vy_bird ?b))
              (increase   (x_bird ?b) (* #t (vx_bird ?b)))))
```

2. Definition of inconsistency

$$C(\pi, D, \tau) = \frac{1}{|\tau|} \sum_i \underbrace{\gamma^i}_{\text{Discount factor}} \cdot ||\overbrace{S(\tau)[i]}^{\text{Observation}} - \overbrace{S(\pi, D)[i]}^{\text{Expectation}}||$$

3. Space of model design

$$\text{repairable fluents} = \{x_1, x_2, x_3, \dots, x_n\} \subseteq X \in D$$
$$\text{deltas} = \{x_1: 1, x_2: 0.2, x_3: 0.1, \dots, x_n: \Delta_n\} \in \mathbb{R}^n$$

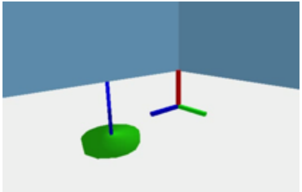
4. Search to minimize inconsistency

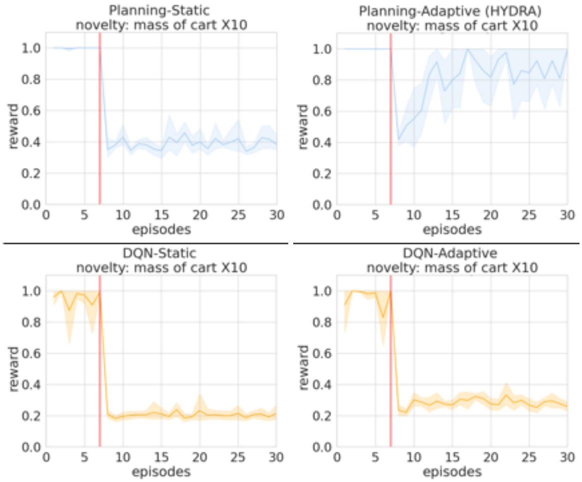
Empirical Results

- Resilient
- Fast
- Interpretable by design

Repair

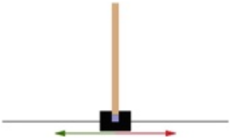
```
[mcart: 9.0, lpole: 0, mpole:
0, forcemag: 0, gravity: 0,
...], resulting inconsistency:
0.0067561
```

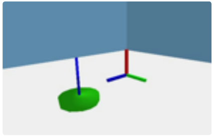




A Domain-Independent Framework

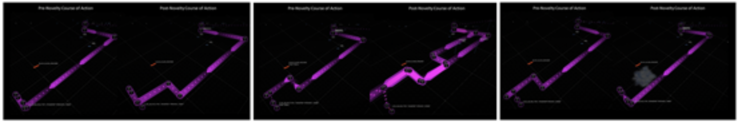
CartPole 2D/3D: continuous state and action; continuous control problem





AngryBirds: continuous state and action; initial condition problem

MineCraft: discrete state and action; goal achievement problem



UAV: continuous state and action; mission flying

ID	Type	Description	Evidence					
			CartPole++		ScienceBirds		PogoStick	
			D	A	D	A	D	A
1	Attribute	New attribute of a known object or entity	✓	✓	✓	✓	✓	✓
2	Class	New type of object or entity	✓	.	✓	.	✓	.
3	Action	New type of agent behavior/control	*	*	*	*	*	*
4	Interaction	New relevant interactions of agent, objects, entities	✓	.	✓	.	✓	✓
5	Activity	Objects and entities operate under new dynamics/rules	✓	.	✓	✓	.	.
6	Constraints	Global changes that impact all entities	✓	✓	✓	✓	✓	✓
7	Goals	Purpose of the agent changes	*	*	*	*	*	*
8	Processes	New type of state evolution not as a direct result of agent or entity action	✓	.	.	.	.	.

Blending GenAI and Model-based Reasoning

1. Explicit event model

```
(:process flying
 :parameters (?b - bird)
 :precondition (and (bird_released ?b)
                    (= (active_bird) (bird_id ?b))
                    (> (y_bird ?b) 0))
 :effect (and (decrease  (vy_bird ?b) (* #t (gravity))
              (increase    (y_bird  ?b) (* # (vy_bird ?b))
              (increase    (x_bird ?b) (* #t (vx_bird ?b)))))
```

2. Definition of inconsistency

$$C(\pi, D, \tau) = \frac{1}{|\tau|} \sum_i \underbrace{\gamma^i}_{\text{Discount factor}} \cdot \overbrace{||S(\tau)[i]}^{\text{Observation}} - \overbrace{S(\pi, D)[i]||}^{\text{Expectation}}$$

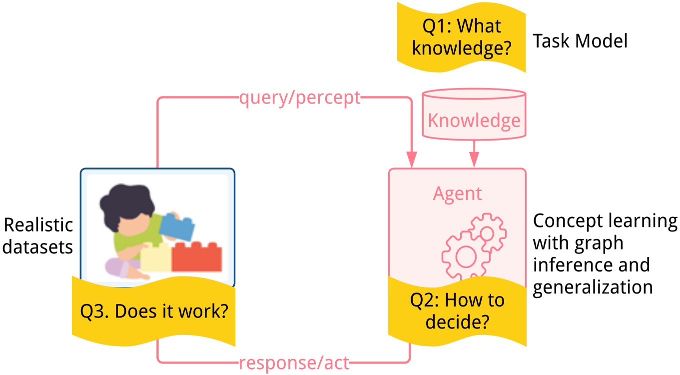
3. Space of model design with GenAI

repairable fluents = $\{x_1, x_2, x_3, \ldots, x_n\} \subseteq X \in D$

deltas = $\{x_1\text{: } 1, x_2\text{: } 0.2, x_3\text{: } 0.1, \ldots, x_n\text{: } \Delta_n\} \in \mathbb{R}^n$

4. Search to minimize inconsistency

Interactive Task Learning



Agents for Unknown Tasks

Agents are designed/trained to perform specific tasks

- All tasks cannot be predicted at design time
- General agent design/training
- Generally-learning agent

DARPA GAILA with Xerox

- How can agents learn new concepts and tasks like human children?
- 5 scientists & engineers, interns
- Human-centered approach to agent design
- 5 patents on NL interaction with physical machines
- JAIR 2025 under review, IEEE RO-MAN 2021, ACS 2020; 2 theses UM, Northwestern
- Contributes to a 10 year legacy of Interactive Task Learning research

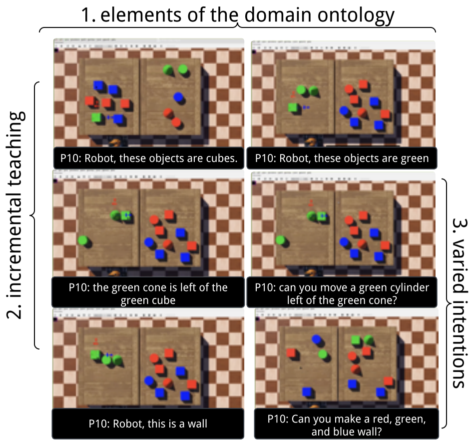
Children Learn in Social Constructs

Earliest learning occurs through caregiver-guided interactions with the world - guided, embodied learning

How Do Humans Teach?

Ramaraj, Ortiz, and Mohan IEEE ROMAN 2021

- N = 10, teach the robot how to build a multi-colored wall
- Video recording of teachers, inductive thematic analysis
- People taught
  1. Compositional concepts; motivates factored task models
  2. Incrementally; motivates incremental learning
  3. Expressing varied intentions; motivates interactive learning
  4. In a structured curriculum; motivates simplifying assumptions



Video

Human-Teaching Inspired Curriculum

Teach | **Measure generality (true positive score)** | **Measure specificity (true negative score)**

inform: green cone

verify: green cone

verify: green cone

inform: yellow cone left of red cylinder

verify: blue cylinder left of red sphere

verify: green cube left of blue cone

inform: move blue cylinder to left of red cube

react: move red cylinder to left of red cone

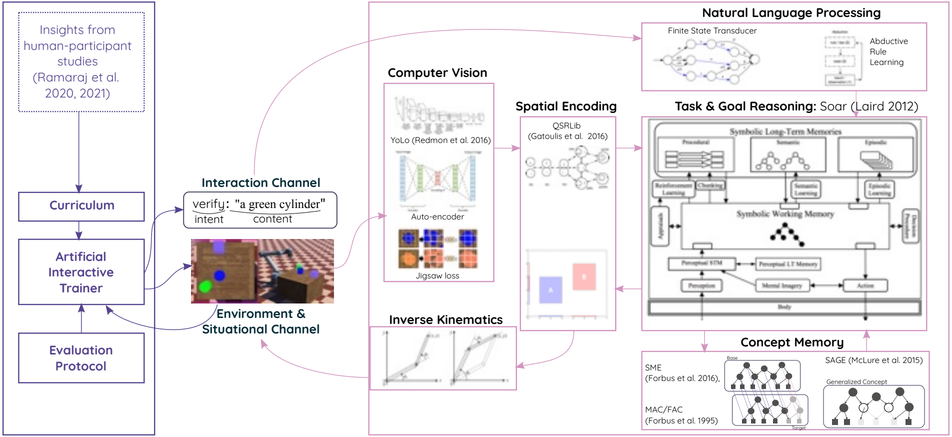
react: move red cube to left of blue cylinder

incremental curriculum

AILEEN

Mohan et al. ACS 2020

Integrated agent system: computer vision, spatial reasoning, task & goal reasoning, planning, analogical reasoning & generalization, inverse kinematics



Factored Task Models

Task **T**

parameters: plate, bread, knife, toaster [$T(o_p, o_b, o_k, o_t)$]
predicates: state [$\text{toasted}(o_b)$], configuration [$\text{on}(o_b, o_p)$]
availability: bread exists, knife exists, plate exists [$o_p, o_b, o_k, o_t \rightarrow \text{propose}(T)$]
children tasks: go-to, slice, retrieve
policy: if holding bread and not sliced, slice bread [$\text{holds}(o_b) \wedge \neg\text{sliced}(o_b) \rightarrow \text{slice}(o_b)$]
termination: bread is toasted, bread is on plate [$\text{toasted}(o_b) \wedge \text{on}(o_b, o_p)$]
model: [$o_p, o_b, o_k, o_t \rightarrow \text{toasted}(o_b) \wedge \text{on}(o_b, o_p)$]
performance criterion: shortest distance

Padmakumar et al. 2022: 'Make a plate of toast"

- Benefits vis-a-vis end-to-end representations: composable, incrementally-learnable, hierarchical
- Mohan and Laird AAAI 2014: availability, policy, termination, model
- Kirk and Laird, IJCAI 2019: games, Mininger and Laird, AAAI 2022: complex task hierarchy
- Mohan et al. ACS 2020, Hancock et al. JAIR 2025 (in review): predicates grounded in visuo-spatial information

Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

Grounded definition of $\text{on}(o_1, o_2)$

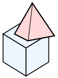
Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

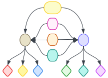
Cognitive theory of structure mapping (Gentner AP 1987)

$\text{sim}(G_s, G_c) = \sum_e w(e) \times \text{corr}(e, G_c)$

Scene **Teacher** **Scene Graph** **Task Control** **Concept Memory**

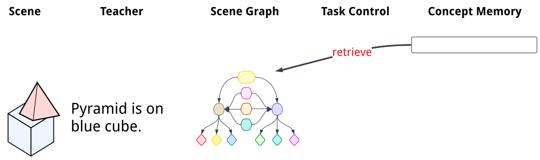


Pyramid is on blue cube.



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)



Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

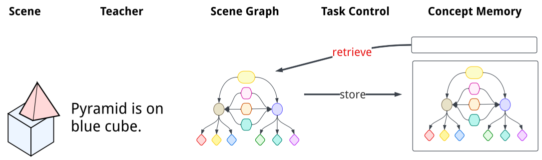
CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$

Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)



Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$

Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

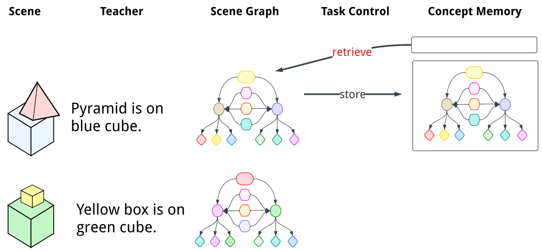
Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

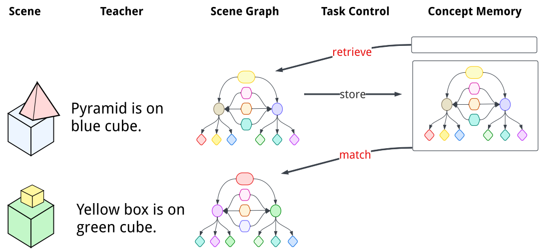
Grounded definition of $\text{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

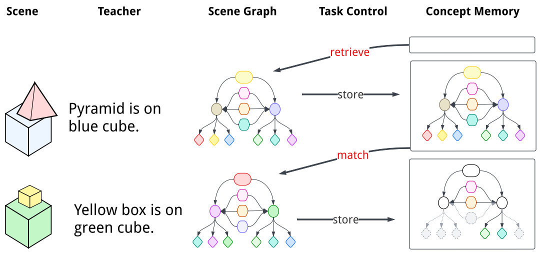
Cognitive theory of structure mapping (Gentner AP 1987)

$\text{sim}(G_S, G_C) = \sum_e w(e) \times \text{corr}(e, G_C)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)



Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$

Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

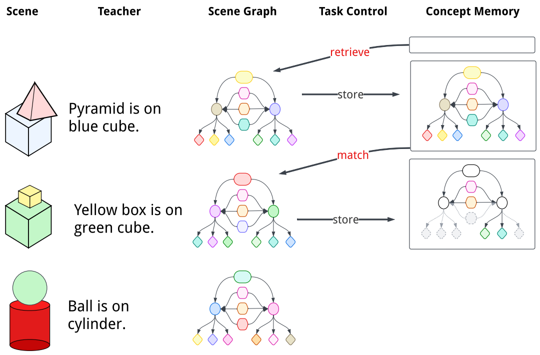
Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

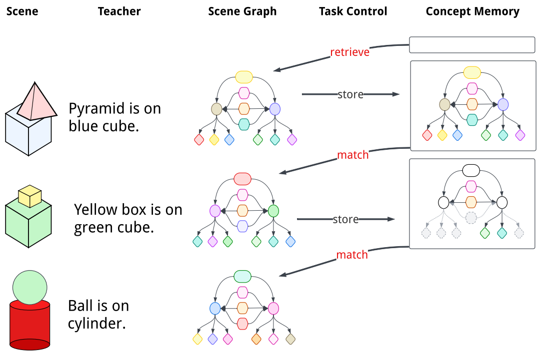
Grounded definition of $\text{on}(\text{o}_1, \text{o}_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\text{sim}(G_s, G_c) = \sum_e w(e) \times \text{corr}(e, G_c)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

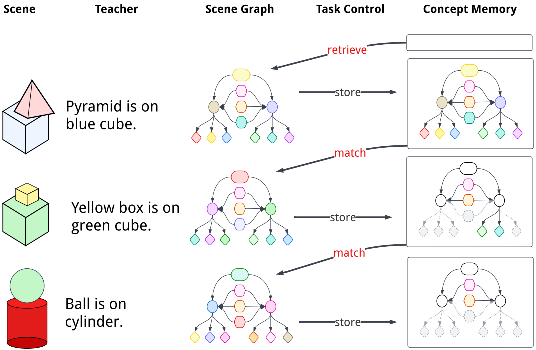
Grounded definition of $\text{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\text{sim}(G_s, G_c) = \sum_e w(e) \times \text{corr}(e, G_c)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

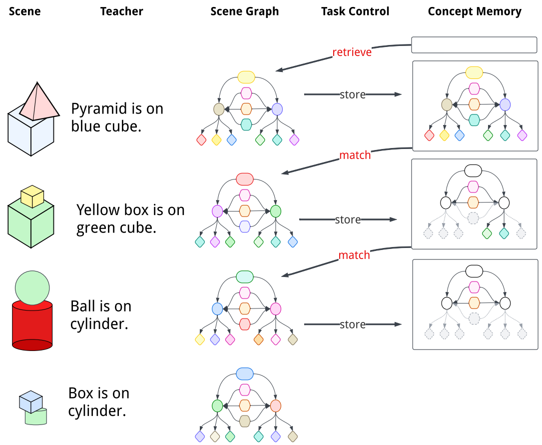
Grounded definition of $\text{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\text{sim}(G_S, G_C) = \sum_e w(e) \times \text{corr}(e, G_C)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

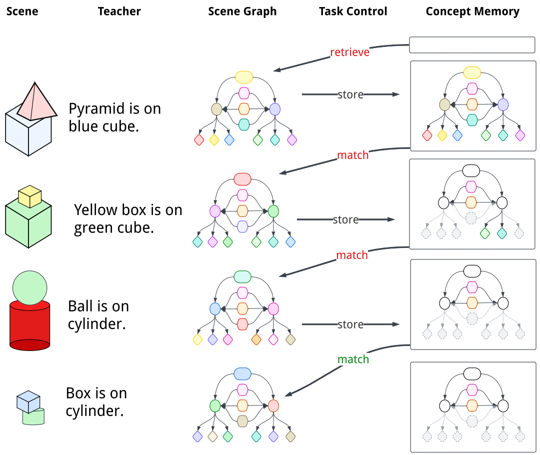
Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)

$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$



Interactive Concept Learning

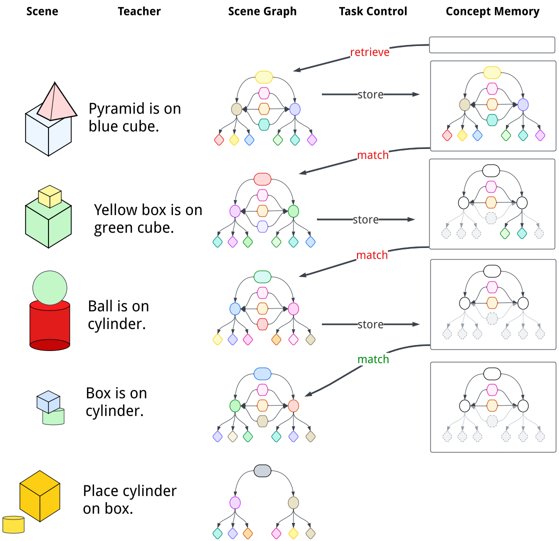
Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

Grounded definition of
$\text{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

Cognitive theory of structure mapping (Gentner AP 1987)
$\text{sim}(G_S, G_C) = \sum_e w(e) \times \text{corr}(e, G_C)$



Interactive Concept Learning

Mohan et al. ACS 2020; Hancock, Mohan, & Forbus JAIR 2025 (in review)

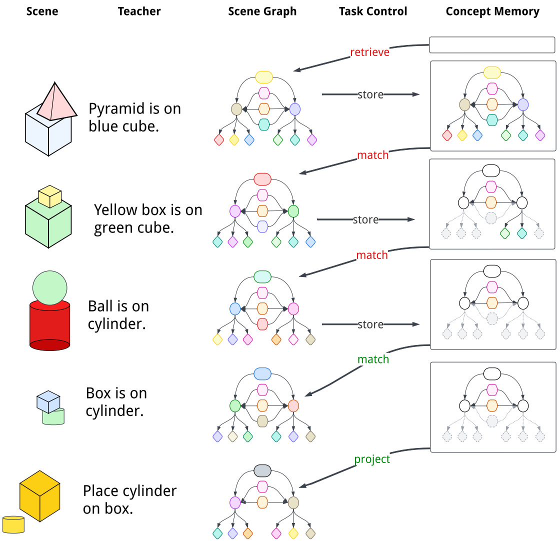
Grounded definition of $\mathrm{on}(o_1, o_2)$

Scene Graph Example

CV: shapeA(o1), colorA(o1), sizeA(o1), shapeB(o2), colorB(o2), sizeB(o2)
CDC: nZ(o1,o2), oZ(o1,o2), oX(o1,o2)
RCC: touchX(o1,o2), intersectZ(o1,o2)

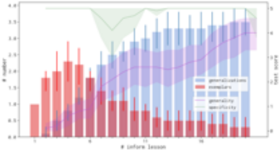
Cognitive theory of structure mapping (Gentner AP 1987)

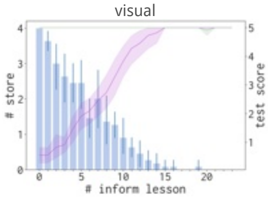
$$\mathrm{sim}(G_s, G_c) = \sum_e w(e) \times \mathrm{corr}(e, G_c)$$

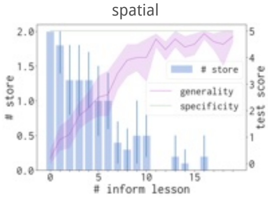


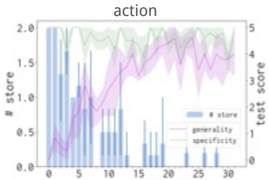
Empirical Observations

- Experimental scheme
  - A trial of N lessons
  - Lesson: an instantiation, generality measurement (true positive rate), specificity measurement (true negative rate)
- Findings:
  - General: visual, spatial, action & events, composite objects
  - Bidirectional: recognition and creation
  - Fast: learns from few examples, rapid generalization, small leakage
  - Active: learns only when needed
- A demonstration

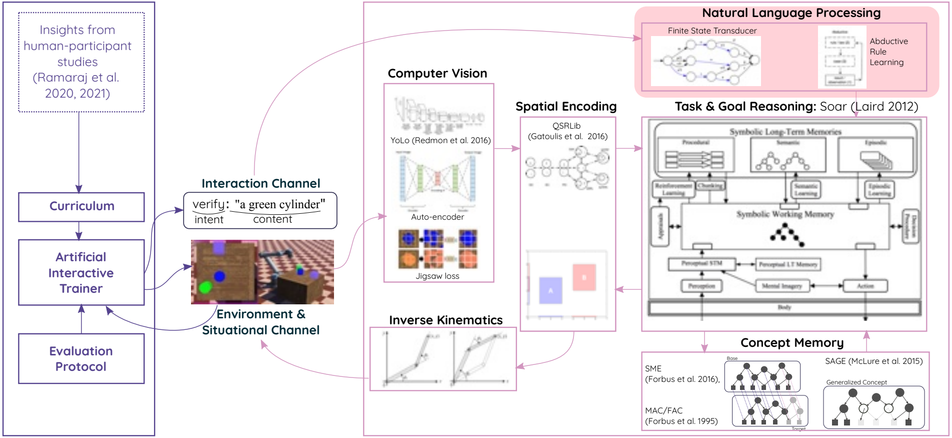






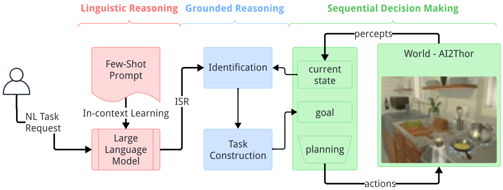


Architecting LLMs with Task Reasoning



Architecting LLMs with Task Reasoning

Grover and Mohan, ICAPS demonstration 2024



- SoA frames language-to-action problem as a sequence to sequence problems
- Our approach frames it as a task model acquisition problem
  - In-context learning for language to meaning representation
  - Grounded reasoning to instantiate a goal
  - Planning to generate a sequence of actions
- Ongoing work: concept memory with LLM-based pattern matching

Few-shot Prompt

language: Please clean that mug.
ISR: {INTENT: request, action:0: {TYPE: clean-task, obj:0: {TYPE: mug}}}
language: Can you please wash this plate?
ISR: ....
language: ....
ISR: ...
....
language: Please place the bread in the fridge.

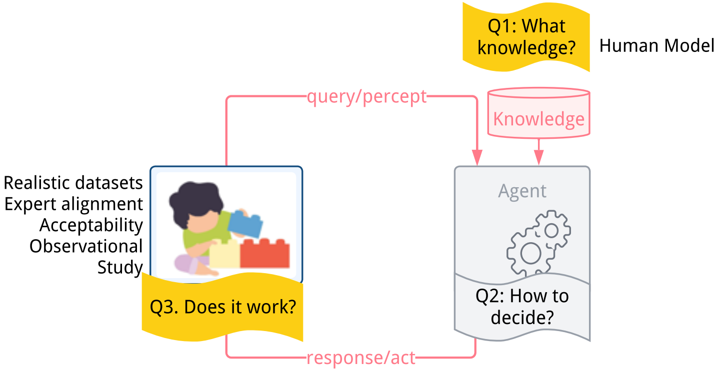
Response

ISR: {INTENT: request, action:0: {TYPE: find-and-put-task, obj:0: {SPEC: any, TYPE: apple}, obj:1: {SPEC: unique, TYPE: bread}, rel:0: {TYPE: in, ARG1: obj:0, ARG2: obj:1}}}

isa(o1, door), isa(o2, stove), isa (o3, fridge), isa (o4, knife), isa (o5, bread), isa (o6, countertop)....,
at(o1, <loc1>), at(o2, <loc2>), at(o3, <loc3>)....
iscontainer(o3), issurface(o6),....
....
Current State

Generated Goal in(o5, o3)

Health Behavior Coaching



Agents for Human Learning

Agents are designed for static human needs & preferences

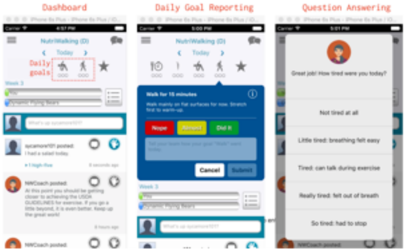
- Human are continual learners and evolve throughout our lives
- Adoption depends on responsiveness

NSF/NIH SCH with Kaiser Permanente

- How can agents help people develop healthy behaviors?
- Publications
  - AI: IAAI/AAAI2017
  - HCI: TiiS2020, TiiS2021
  - Medicine: JMIR2019, JMIR2017
  - Engineering: EMBS2016
- First ecological, long-term evaluation of adaptive AI behavior
- Collaboration with psychologists, user-experience researchers, clinicians, patients

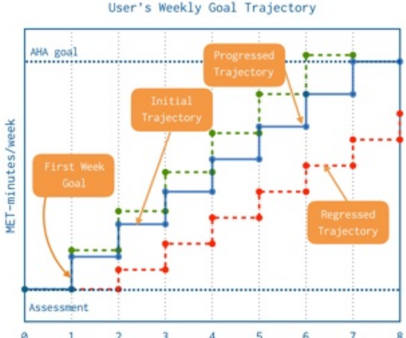
Coaching Agent in mHealth

- Support sedentary individuals in regular exercise
- AHA recommendation: 30 minutes, 5 times a week
- Designed in collaboration with a physical therapist



Collaborative Adaptive Goal Setting

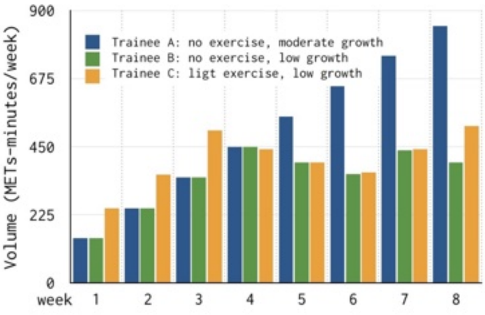
1. Determine current exercise volume
2. Propose different extents, evaluate with user
3. Assume a uniform step growth model until AHA goal
4. Schedule exercise for the week, maximize opportunity
5. Measure behavior, self-efficacy, & difficulty
6. Revise growth model, replan next week



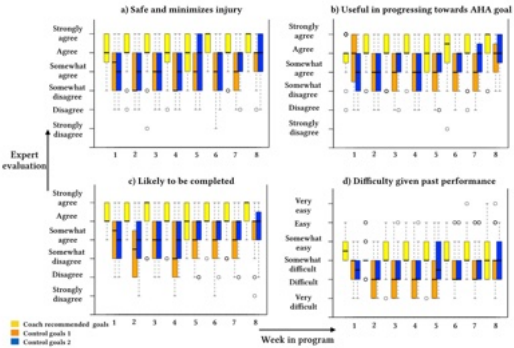
Evaluation Paradigm

Mohan et al. TiiS 2020; Mohan et al. AAAI 2017

1. Realistic Datasets: simulated patient profiles



2. Alignment: choice studies with expert panel

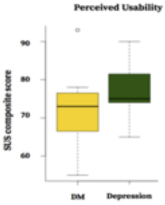


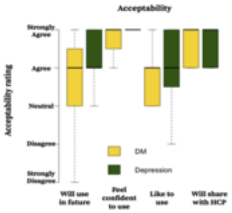
Evaluation Paradigm

Mohan et al. TiiS, 2020; Hartzler et al. EMBC 2016

3. Acceptability cognitive walkthroughs with patients
N=15, diabetes and depression

- Could provide users with control (P9)
- Helps you take responsibility (P1), with more choice (P7)
- Allows you to set goals that you can strive for (P8).





4. Impact ecological observational study
N=21, 6 weeks

1. Increased exercise volume by 20%
2. Over-optimistic with self-assessment
3. Personalized goals + collaborative selection led to more successful completion

	(1)	(2)	(3)
Independent Variables↓	Goal Volume	Performed Exercise	Performed Exercise
Week	9.608*	12.392*	-0.487*
	(5.166)	(12.202)	(12.007)
Goal Volume			0.618***
			(0.119)
Mean Dependent Variable	601.098	392.250	392.250
	(23.138)	(24.830)	(24.830)
Random effect	✓	✓	✓
Marginal R^2	0.004	0.005	0.378
Conditional R^2	0.868	0.662	0.639

Table 2. Mixed-effect linear regression models for goal volume (column 1) and performed exercise volume (column 2). Volume is measured in MET-mins/week. The numbers in parentheses are standard errors. *** $p < 0.001$, ** $p < 0.05$, * $p < 0.1$

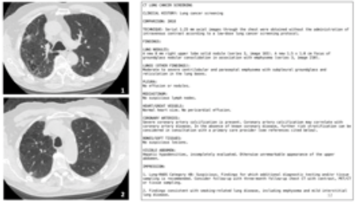
Evaluating GenAI Systems on Realistic Datasets

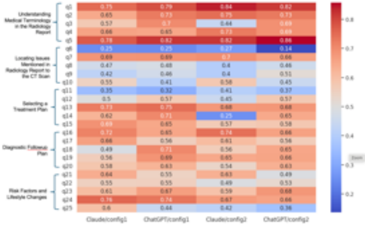
Rajagopal et al. HealthIUI at ACM IUI 2025

Real problem: can GenAI support people's informational needs?

- Studied patient-radiologist dyadic sensemaking interaction
- Identified 10 different themes and curated a realistic QA dataset
- Evaluated ChatGPT and Claude wrt expert responses.
    1. High error rate (ChatGPT: 20%, Claude: 40%)
    2. Inability to ground interaction in images
    3. GenAI responses were long; had irrelevant, superfluous, banal elaborations
    4. Radiologist response geared towards helping decisions, while GenAI responses towards extensive enumeration and definitions

Question	Are these nodules, then are they present in in healthy individuals as well? Or is that specifically the cancer that's manifesting itself. Or what's the definition of that?
Radiologist	So nodules are very widely present in most adult humans, I would say about one in 2 adults have pulmonary nodules, and many of them, I would say a vast majority of them do not turn out to be cancer. However, some of these nodules occasionally turn out to be cancer. And we, as radiologists, review various different factors, including how the nodule is looking and shaped in these ct scans. How big are the nodules. How fast is it growing over time? And where is the nodule located? These are some of the imaging features that we look-into in addition to patient age, gender, and any other clinical history of other types of cancer. These holistically go into our understanding of how suspicious are these nodules for potential lung cancer.
Claude	To answer your question, let me clarify a few points about lung nodules: Definition: Lung nodules are small, round or oval-shaped growths in the lung. They appear as white spots on chest CT scans. Occurrence in healthy individuals: Lung nodules can occur in both healthy individuals and those with lung disease. Many lung nodules are benign (non-cancerous), especially in younger people without a history of smoking. New nodules in this case: The report mentions "A new 8 mm right upper lobe solid nodule". This is significant because: It's new (wasn't present in the 2018 comparison scan) It's solid (completely opaque on CT, which can be more concerning than non-solid nodules) It's 8 mm in size (nodules larger than 6 mm are generally considered more suspicious) Cancer risk: While not all nodules are cancerous, new solid nodules in the context of a lung cancer screening program (especially in individuals with risk factors like smoking history) are considered suspicious until proven otherwise. Classification: This finding, along with the "new 1.5 x 1.0 cm focus of groundglass nodular consolidation", has led to the Lung-RADS Category 4B classification, which indicates a high suspicion for malignancy. Next steps: The report recommends additional diagnostic testing, which may include follow-up CT, PET/CT, or tissue sampling to determine if these nodules are indeed cancerous. In summary, while lung nodules can be present in healthy individuals, new solid nodules of this size in the context of a lung cancer screening are considered suspicious and require further evaluation. They're not definitively cancer at this point, but they represent findings that need close follow-up and potentially more invasive testing to rule out malignancy.
ChatGPT	In the context of your report, a "nodule" refers to a small, abnormal growth in the lung tissue. Nodules can have various causes, and not all nodules are cancerous. They can appear in both healthy individuals and those with other conditions, such as infections, inflammation, or benign growths. However, nodules are often monitored closely in individuals undergoing lung cancer screening, as they can sometimes be an early sign of lung cancer. In your case, the report mentions a new 8 mm solid nodule in the right upper lobe, which is being closely evaluated. The fact that it is labeled under "Lung-RADS Category 4B: Suspicious" suggests that this particular nodule is considered concerning enough to warrant further diagnostic testing or a biopsy to rule out or confirm malignancy. The key point is that while lung nodules can occur in healthy individuals, some (like this one) may raise concern for cancer and require additional follow-up.

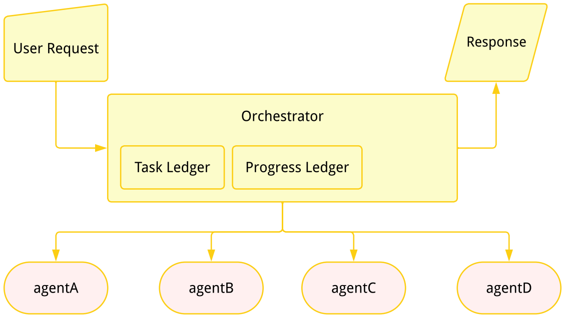




Outlook

Agent Frameworks

Complex behavior emerges from an interplay of diverse, modular reasoning and execution

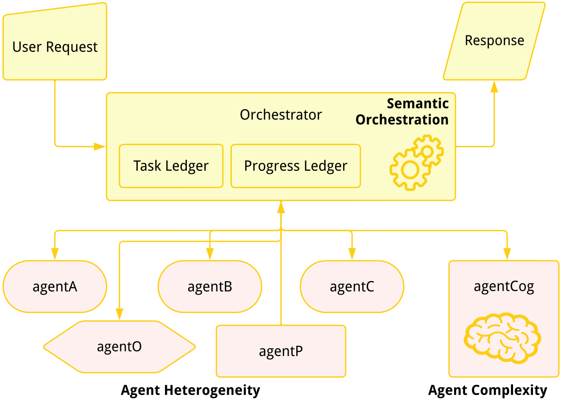


Bansal, G., Vaughan, J.W., Amershi, S., Horvitz, E., Fourney, A., Mozannar, H., Dibia, V. and Weld, D.S., 2024. Challenges in Human-Agent Communication. arXiv:2412.10380

Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J., Alber, J. and Chang, P., 2024. Magentic-one: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468.

Agent Frameworks

Complex behavior emerges from an interplay of diverse, modular reasoning and execution

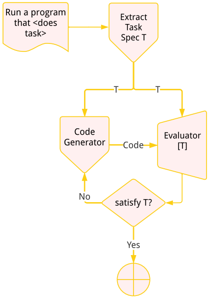


Bansal, G., Vaughan, J.W., Amershi, S., Horvitz, E., Fourney, A., Mozannar, H., Dibia, V. and Weld, D.S., 2024. Challenges in Human-Agent Communication. arXiv:2412.10380

Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J., Alber, J. and Chang, P., 2024. Magentic-one: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468.

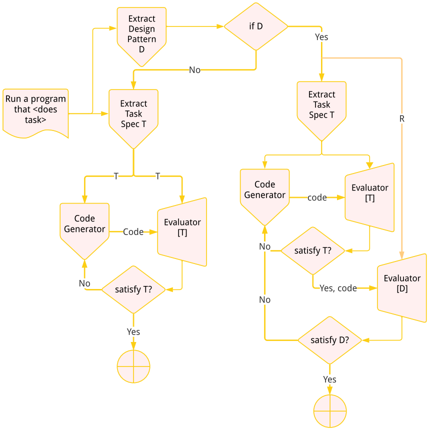
Semantic Orchestration

Augment GenAI inference with compositional SDM (Q2/reasoning)
Modular organization of decision control; help user set appropriate expectations



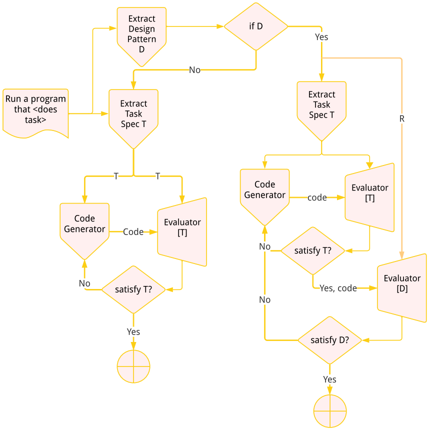
Semantic Orchestration

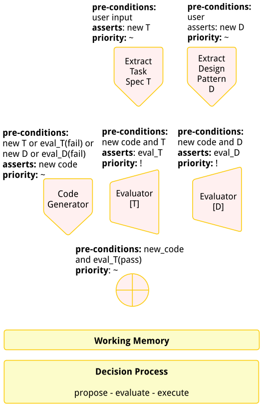
Augment GenAI inference with compositional SDM (Q2/reasoning)
Modular organization of decision control; help user set appropriate expectations



Semantic Orchestration

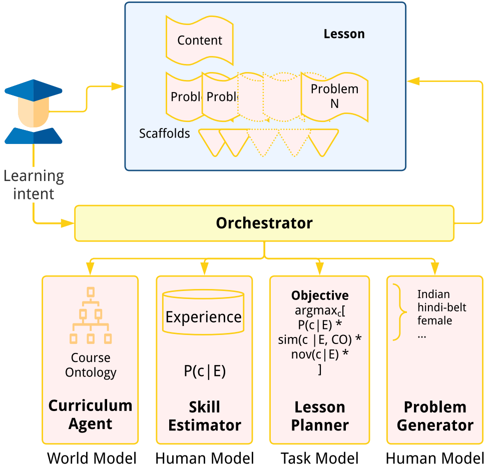
Augment GenAI inference with compositional SDM (Q2/reasoning)
Modular organization of decision control; help user set appropriate expectations





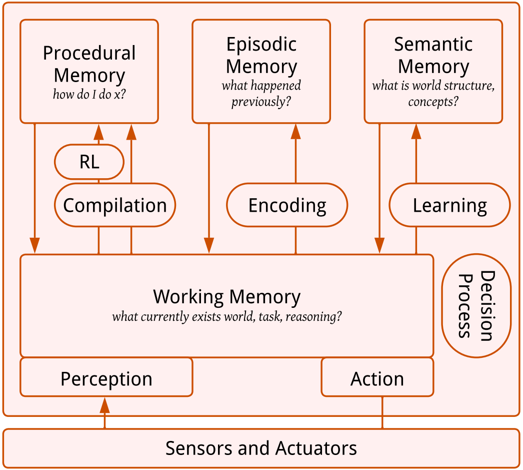
Agent Heterogenity

Agents vary in function, purpose, and inference



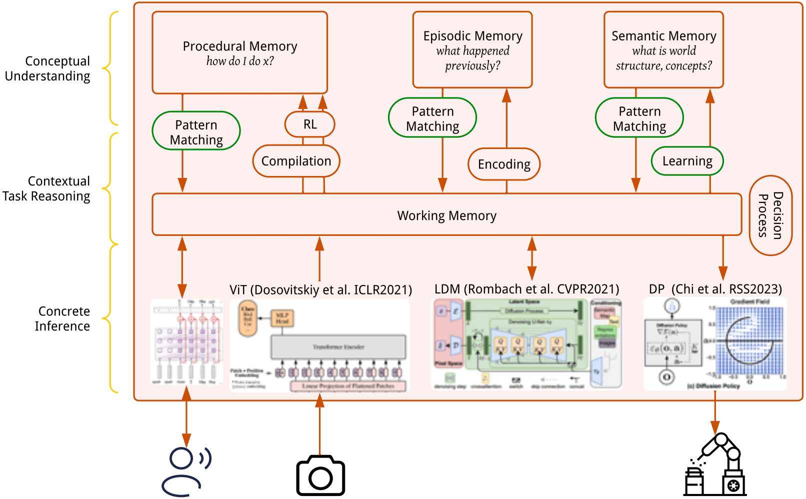
Agent Complexity

Agent with multiple cognitive capabilities



Laird, J.E., Lebiere, C. and Rosenbloom, P.S., 2017. A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. AI Magazine, 38(4), pp.13-26.

Agent Complexity: Cognitive Architectures for the Real World



Also, Sumers, T.R., Yao, S., Narasimhan, K. and Griffiths, T.L., 2023. Cognitive Architectures for Language Agents. Transactions in Machine Learning Research.

Thank You!

World models: Piotrowski et al. ICAPS2023, Piotrowski et al. ICAPS2021
Task models: Grover and Mohan, ICAPS-D2024, Mohan et al. IUI-W2019, Mohan and Laird AAAI2011
Human models: Ramaraj et al. ROMAN2021, Mohan TiiS2021, Mohan et al. TiiS2020, Mohan et al. JAIR2019, Mohan et al. 2017

Q1: What knowledge?

query/percept

Realistic benchmarks
Rajagopal et al. HealthIUI2025

Observational/Choice Studies
Mohan TiiS2020
Mohan et al. JAIR2019
Mohan et al. AIES2019

RCT partial-factorial design
Mohan TiiS2021
Springer et al. JMIR2017
Pirolli et al. JMIR2017

Q3. Does it work?

Agent

Q2: How to decide?

response/act

Learning Fast & Slow
Laird and Mohan AAAI2018, Blue Sky Award

Open-world Learning
Mohan et al. AIJ2024
Piotrowski et al. ICAPS-D2024
Piotrowski et al. AAMAS2023

Interactive Task Learning
Mohan et al. ACS2020
Mohan and Laird AAAI2014

Analogical Generalization
Hancock et al. JAIR2025 (in-review)

My Amazing Colleagues

Shreya Rajagopal, Poorvesh Dongre, William Hancock, Preeti Ramaraj, Sachin Grover, Wiktor Piotrowski, Jacob Le, Kalai Ramea, Matthew Klenk, Charles Ortiz, Roni Stern, Johan de Kleer, Matthew Shreve, Victoria Bellotti, Bob Price, Anusha Venkatakrishnan, Andrea Hartzler, Peter Pirolli, James Kirk, Aaron Mininger, Ken Forbus, John Laird