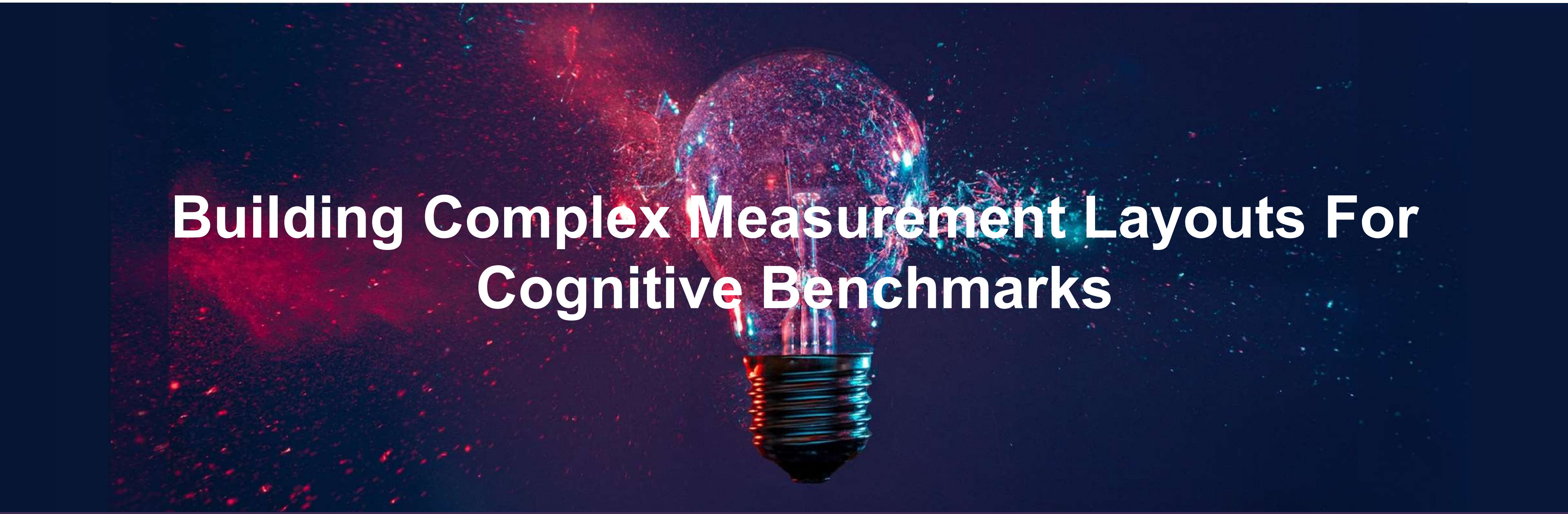
UNIVERSITY OF
CAMBRIDGE



Building Complex Measurement Layouts For Cognitive Benchmarks

Konstantinos Voudouris

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Vancouver, 20 February, 2024

Leverhulme Centre for the Future of Intelligence,
University of Cambridge,

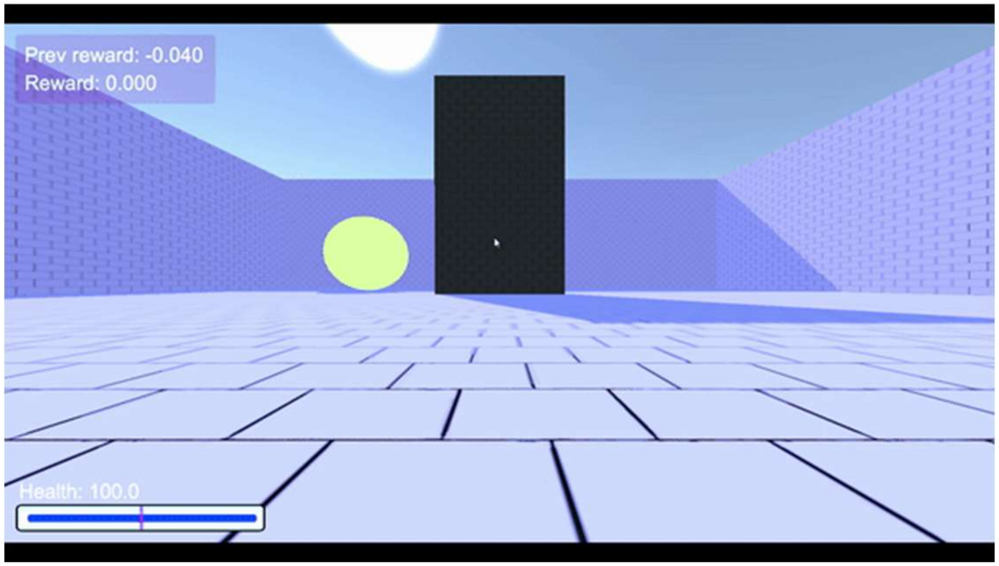
This Session

- Introduce key considerations for developing a useful benchmark for a measurement layout.
- Motivate the role of theoretical knowledge about capabilities in benchmark design and measurement layout development.
- Incrementally build a complex measurement layout for evaluating object permanence (and related capabilities).
- Extend the measurement layout to the multivariate case.
- Apply this measurement layout to real data from DRL agents and humans.

Choosing A (Primary) Capability

- Reinforcement Learning:
  - Long-term planning
  - Tool-use
  - Intuitive physics (object permanence, causality, solidity, inertia)
- Language Models:
  - Theory of Mind
  - Arithmetic
  - Detecting deception

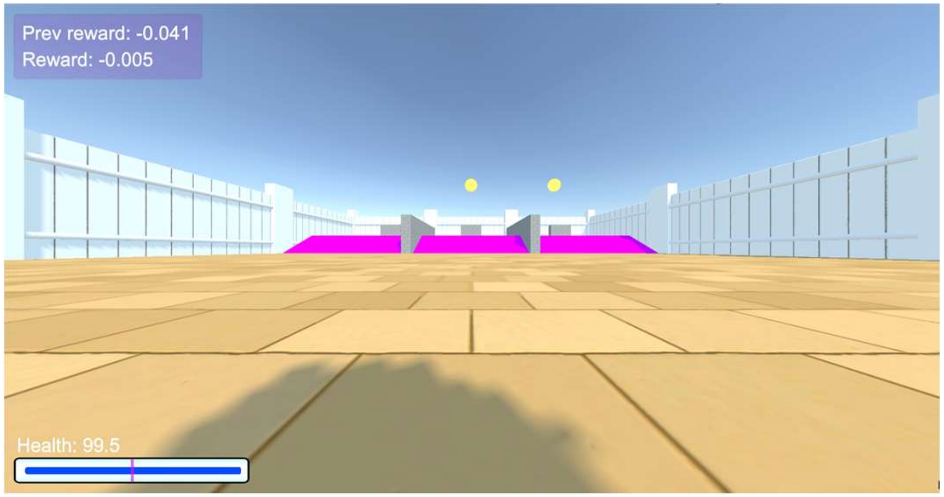
Today's Capability: Object Permanence



Construct Validity

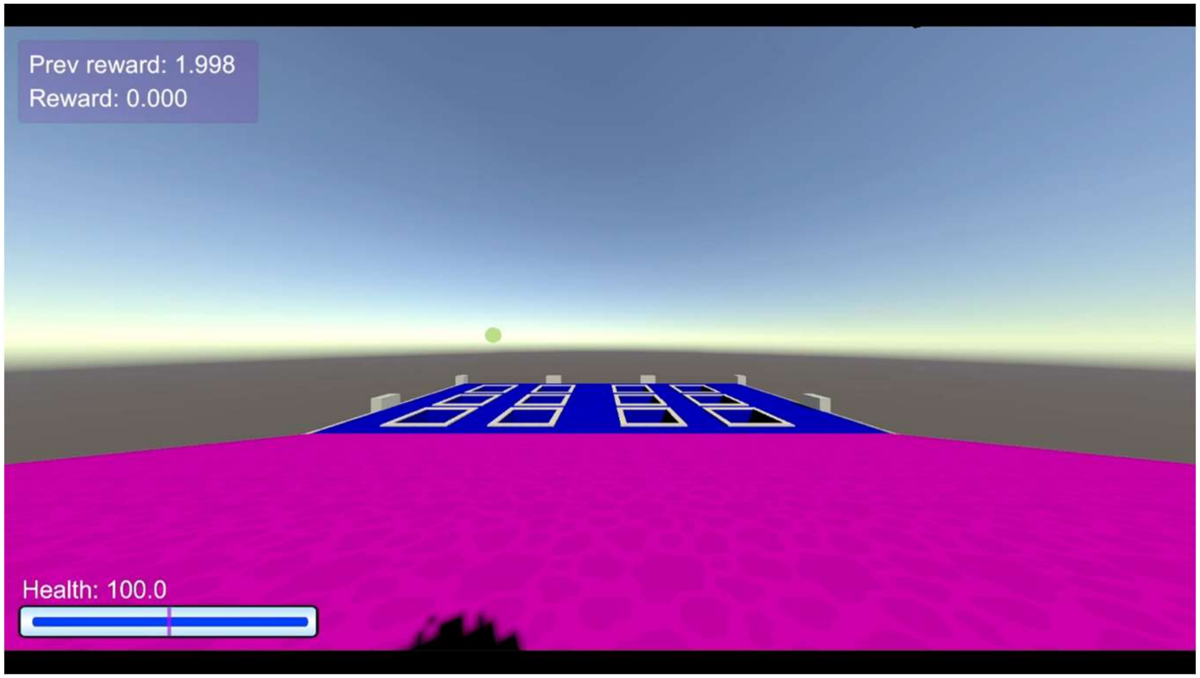
- To what degree does a test accurately measure what it is intended to measure?
- Difficult to guarantee:
  - Tests require validation against other measures.
  - Measures need to be reliable (test-retest).
  - May ultimately be circularly defined.
- In AI Evaluation, we can often draw on research evaluating capabilities in other systems: humans and other animals.

O-PIAAGETS: PCTB



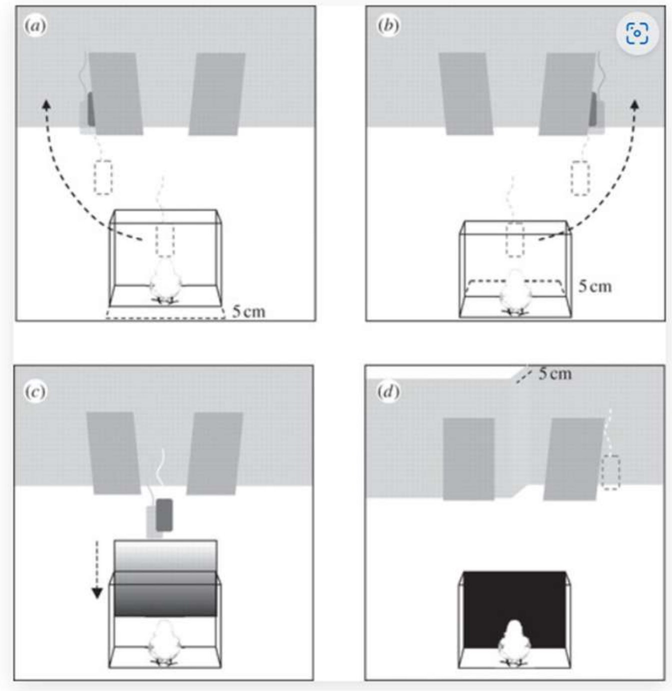
Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. science, 317(5843), 1360-1366.

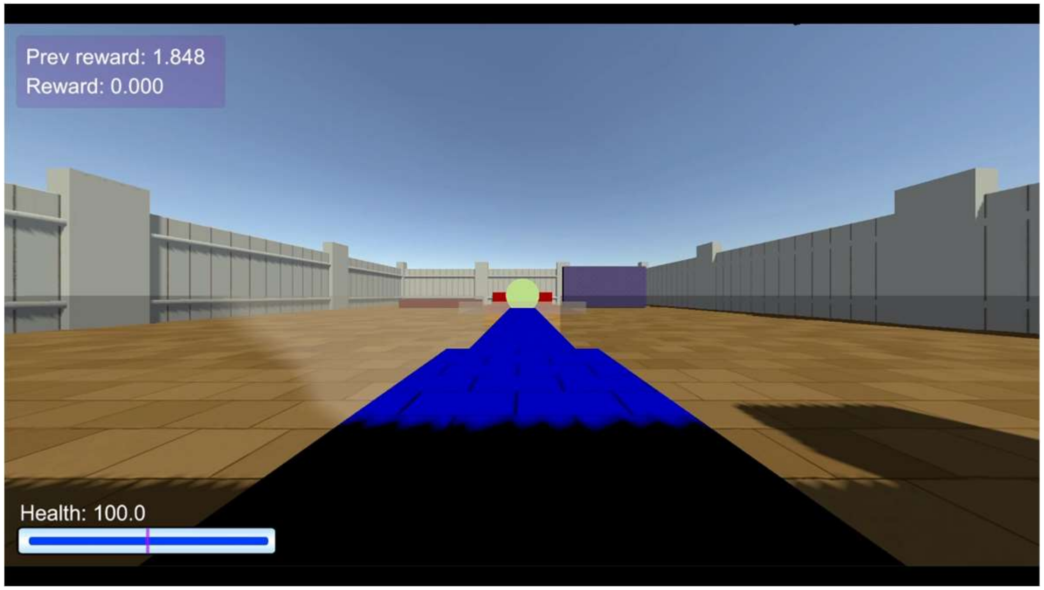
O-PIAAGETS: PCTB



Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. science, 317(5843), 1360-1366.

O-PIAAGETS: CV Chick Tasks

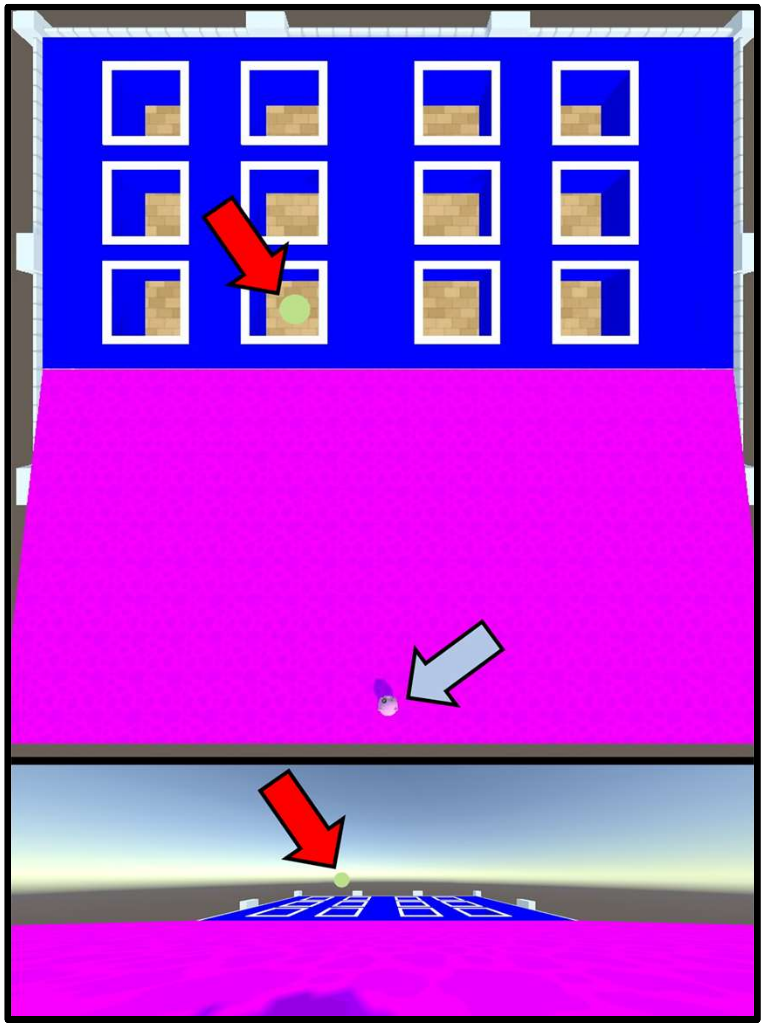



Prev reward: 1.848
Reward: 0.000
Health: 100.0

Chiandetti, C., & Vallortigara, G. (2011). Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2621-2627.

Internal Validity

- What could explain success/failure on this task?
- Object permanence
- Spatial Navigation
- Visual Acuity
- Idiosyncrasies of the test
- Vary as many features as possible



A Battery of Tasks

- Varying:
  - Whether the goal(s) are occluded
  - The shortest path to the goal/choice (a proxy for how long the goal is occluded for).
  - How circuitous that path is
  - The Euclidean distance to the goal(s)
  - The size of the goal(s)
  - The presence of lava
  - Where the goal(s) are placed (left, centre, right)
  - The type of task

A Battery of Agents

- 4202 tasks for 4 agents (Random, Heuristic, PPO, Dreamer-v3)
- 1608 tasks for human players
- Random Action Agent

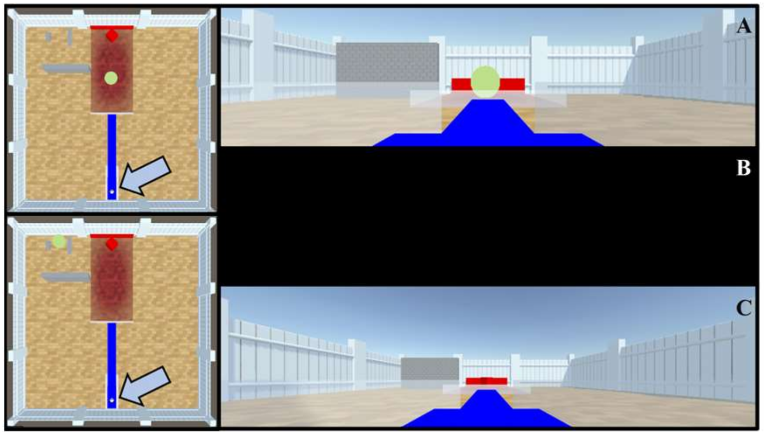
  - Randomly samples actions with equal weight and takes that action for a number of steps sampled from $U(1{,}20)$.
- Heuristic Agent

  - Navigates towards green and yellow goals and away from red lava, following a rigid rule. No Memory.
- Proximal Policy Optimisation (PPO) Agent

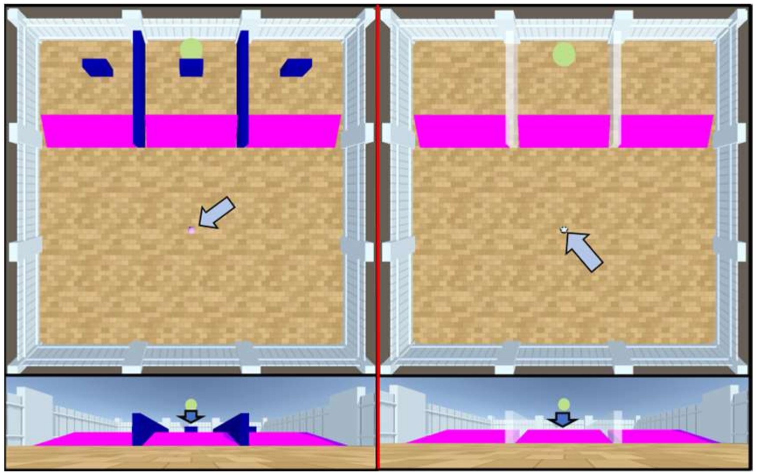
  - Two agents trained on different curricula.
- Dreamer-v3 Agent

  - Two agents trained on different curricula.
- Combined data from 30 humans.

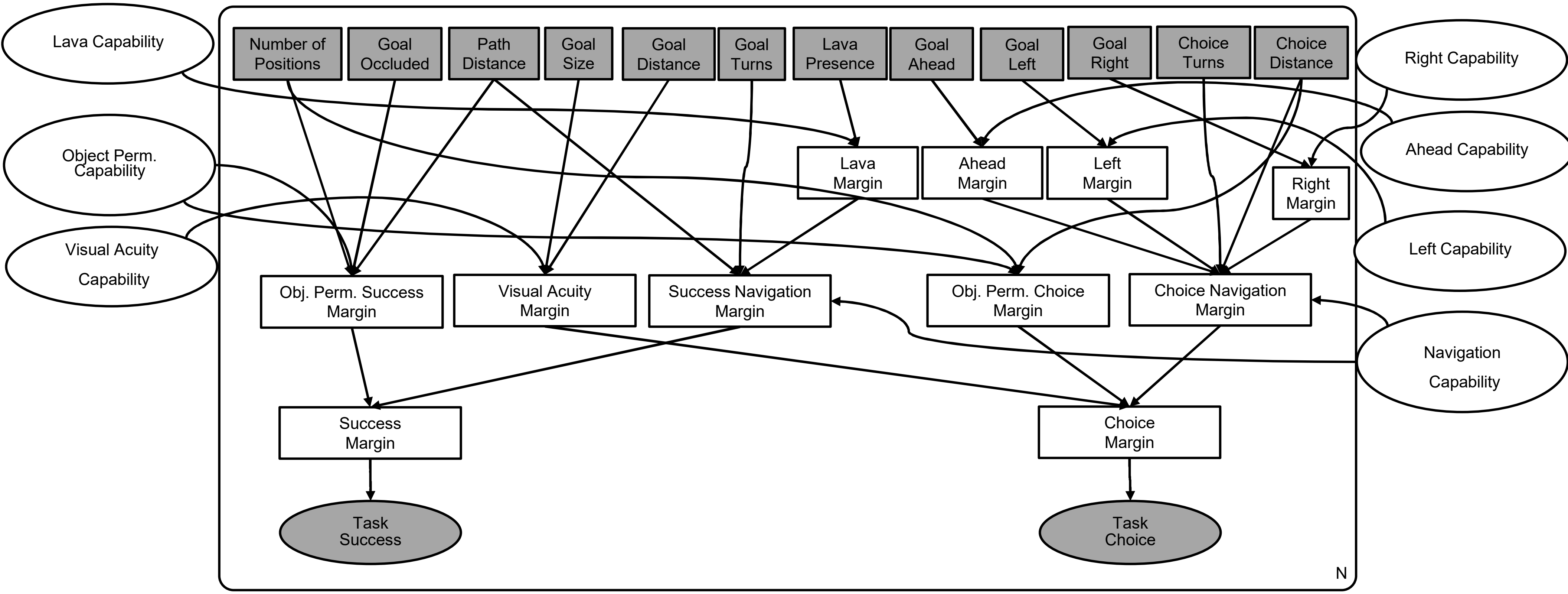
Response Variables



- Whether the agent obtained the goal (success)
- Whether the agent made the correct choice (choice)



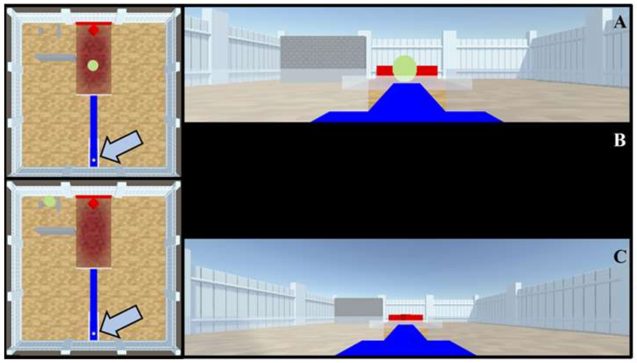
Let’s Start Building: End Goal



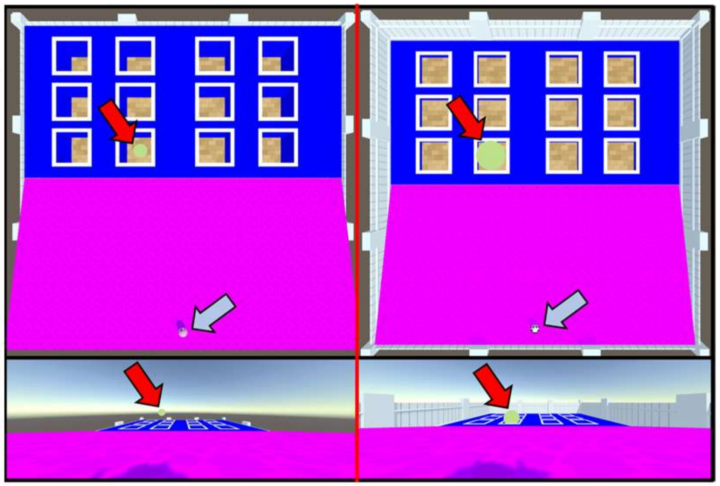
Let's Start Building

- https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial/blob/main/tutorial-notebooks/4_BuildingComplexMeasurementLayouts.ipynb

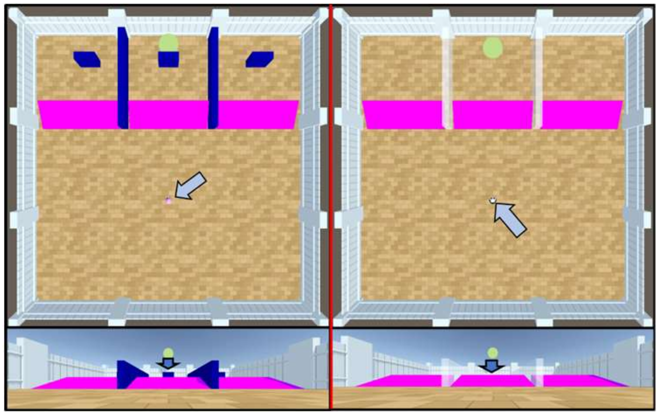
CV Chick Task



PCTB Grid Task



PCTB Cup Task



<u>FOLLOW THE NOTEBOOK UNTIL 16:50</u>

Wrapping Up

- Introduced key considerations for developing a useful benchmark for a measurement layout.
- Motivated the role of theoretical knowledge about capabilities in benchmark design and measurement layout development.
- Incrementally built a complex measurement layout for evaluating object permanence (and related capabilities).
- Extended the measurement layout to the multivariate case.
- Applied this measurement layout to real data from DRL agents and humans.