



Building Good Benchmarks



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE

Konstantinos Voudouris
Department of Psychology,
University of Cambridge

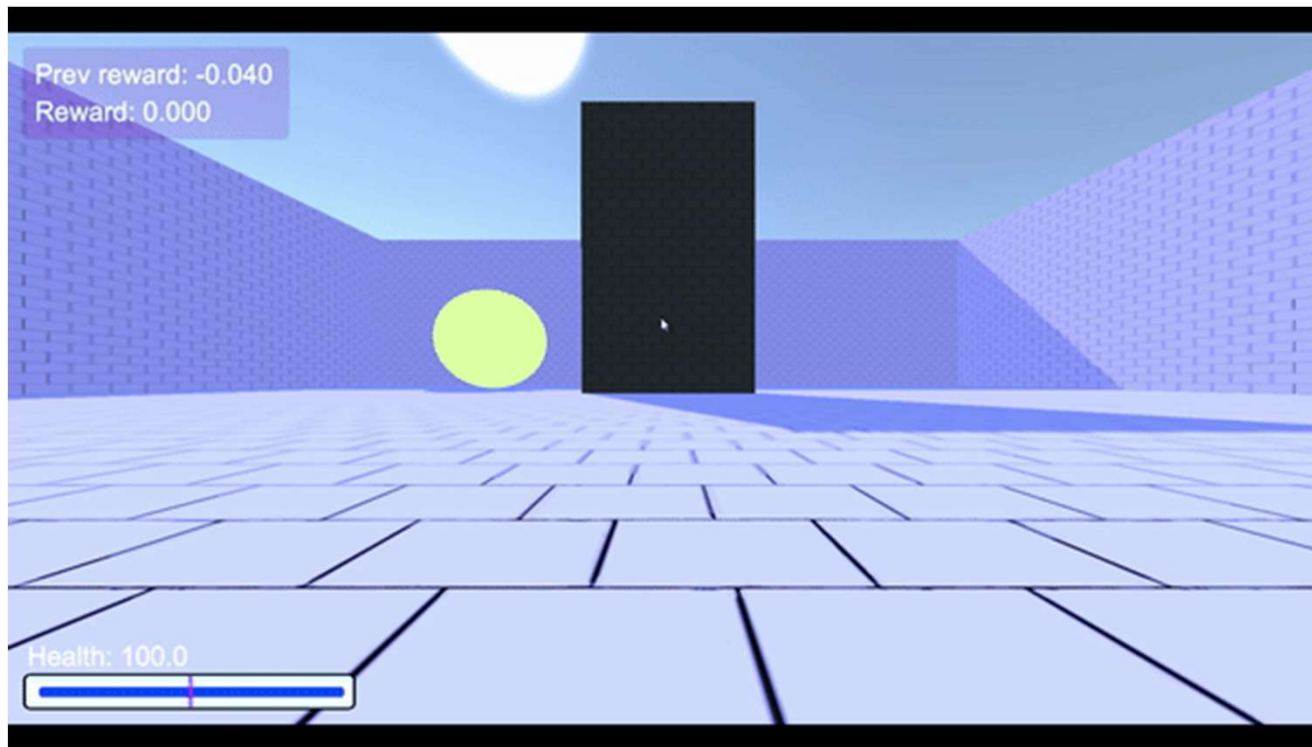
Session Plan

- 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).
- Apply this measurement layout to **real data** from DRL agents.
- Introduce several **model diagnostics** for Hierarchical Bayesian Networks.
- BONUS (if time): Extend the measurement layout to the **multivariate case**.

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.

Tests of Object Permanence



Chimpanzee - Success

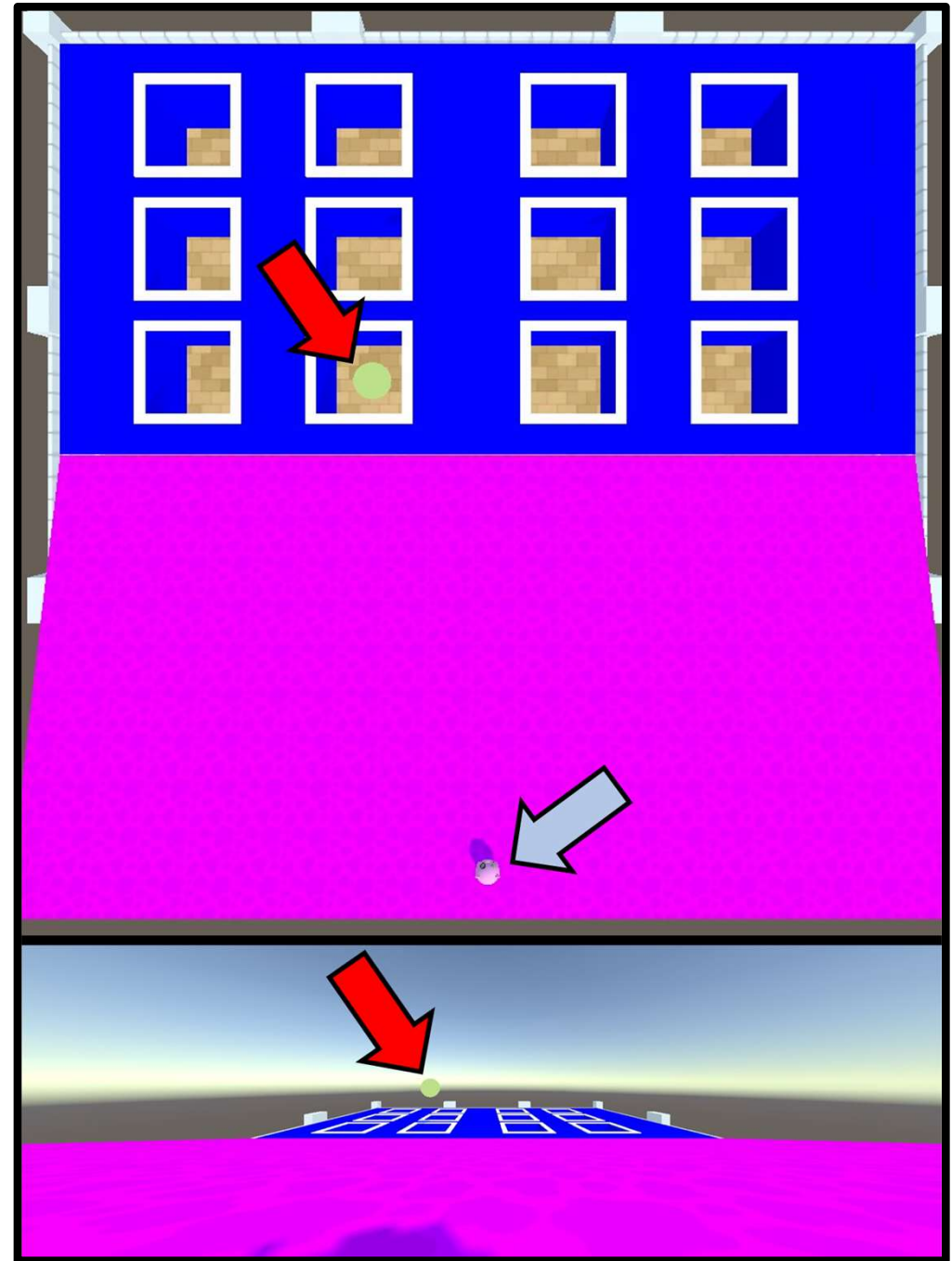


6-Year-Old Human - Fail

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.
Voudouris, K., Donnelly, N., Rutar, D., Burnell, R., Burden, J., Hernández-Orallo, J., & Cheke, L. G. (2022). Evaluating object permanence in embodied agents using the animal-AI environment. *Proceedings of the Evaluation Beyond Metrics Workshop, Vienna, 2022*.

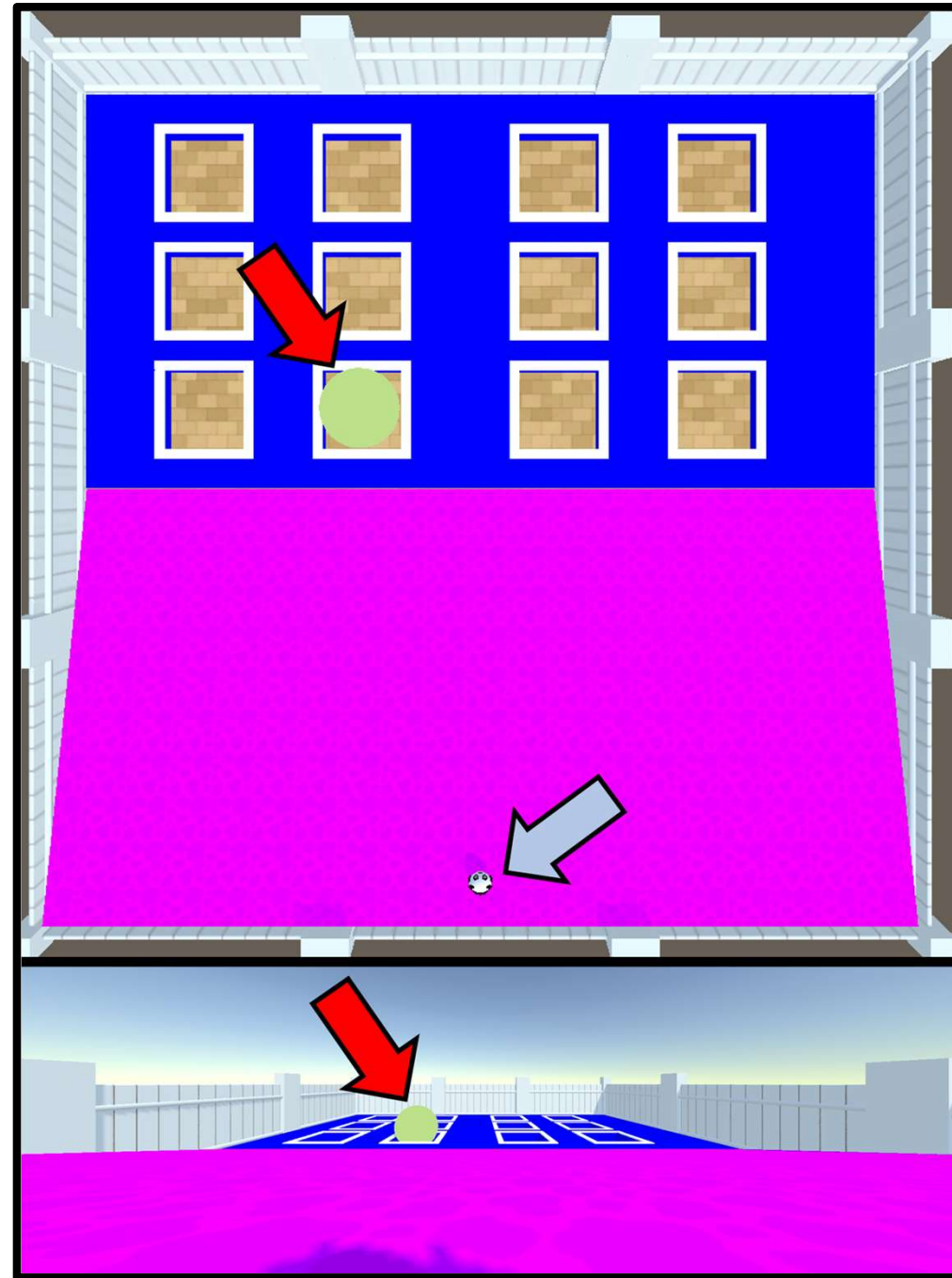
Internal Validity

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



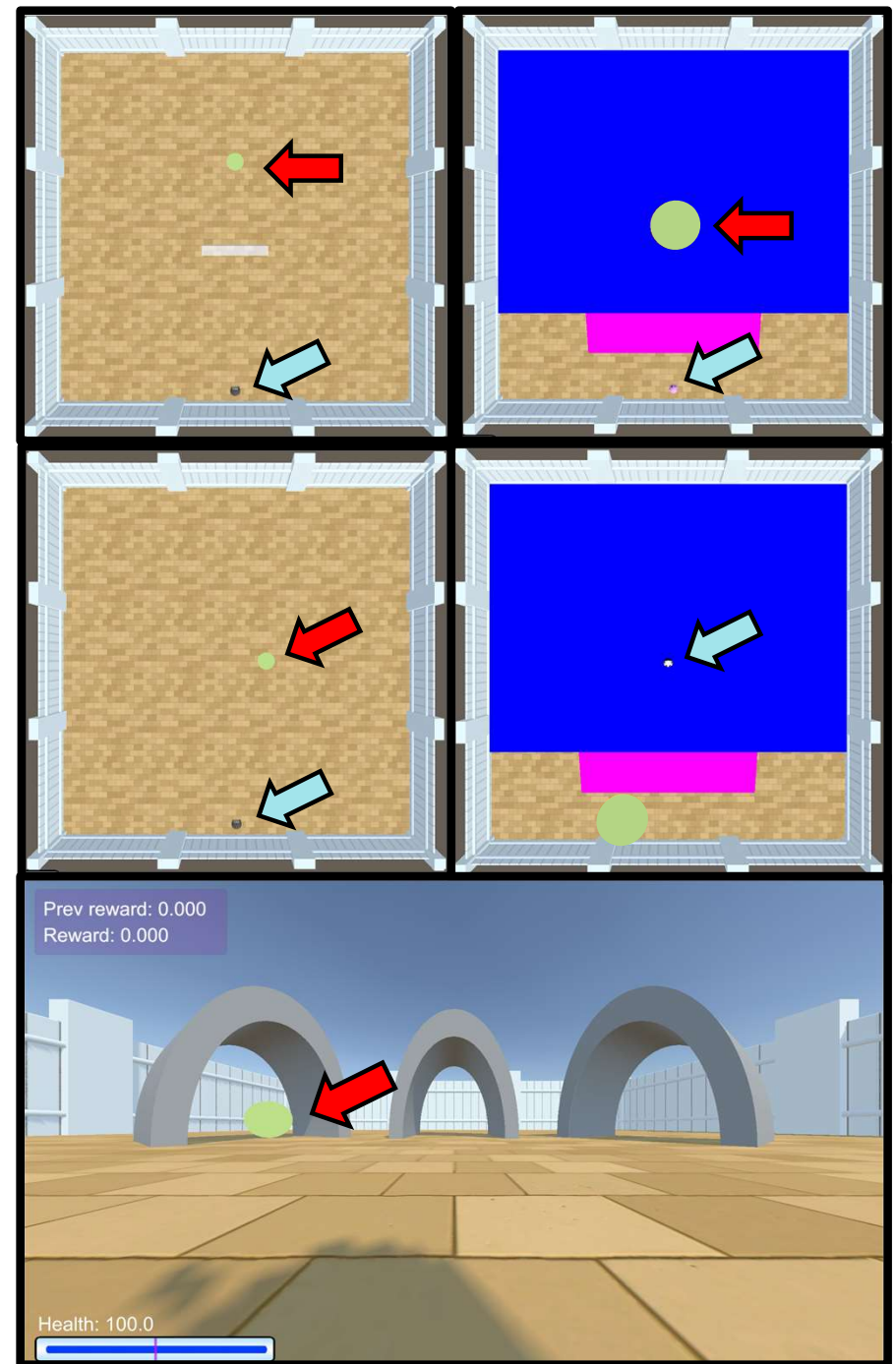
Internal Validity

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



Internal Validity

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



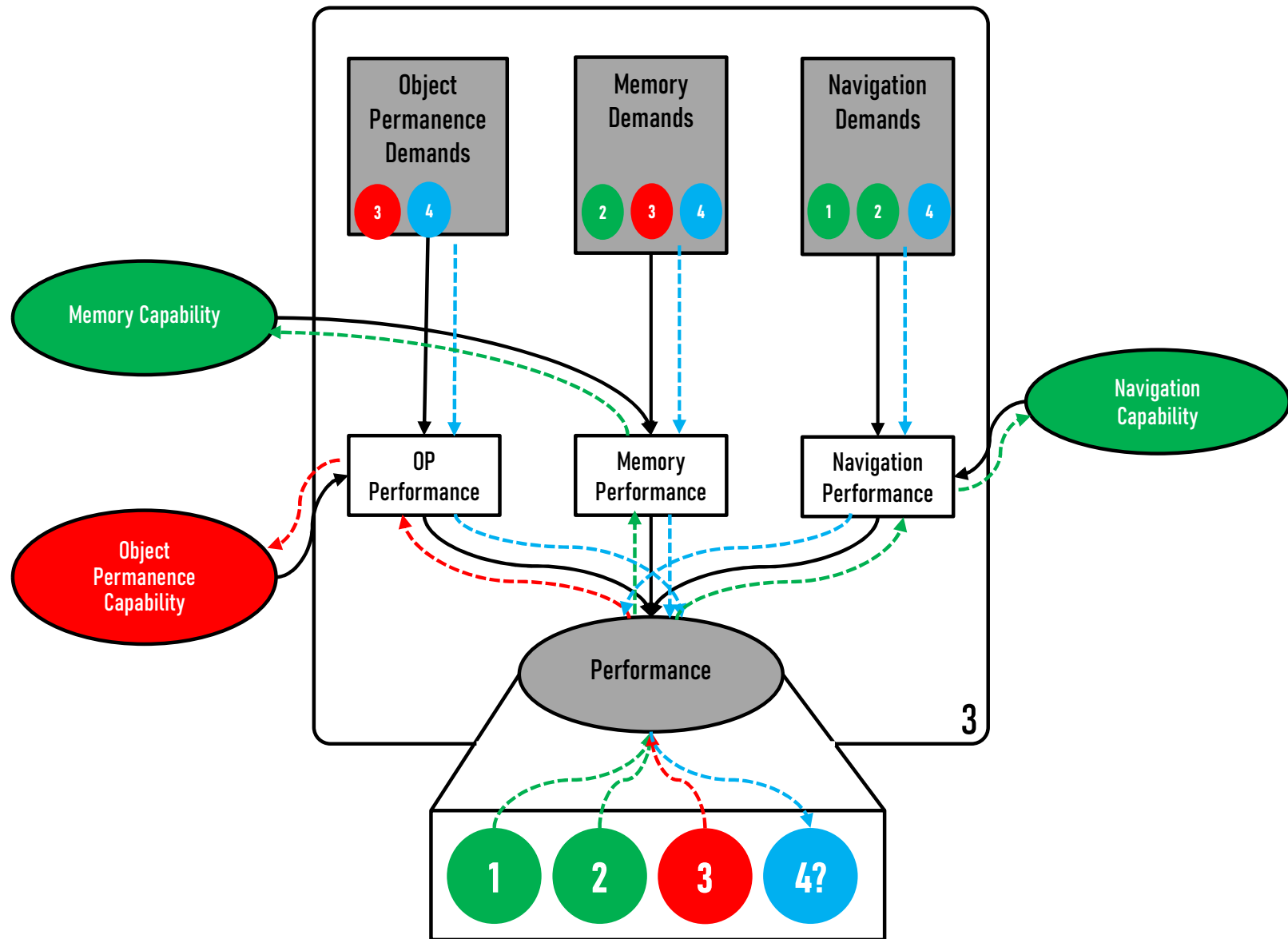
A Battery of Tasks

- 119 Basic Tasks
- 240 Grid Practice Tasks
- 192 Grid Object Permanence Tasks
- Varying:
 - Goal size (`mainGoalSize`)
 - Goal position (`goalPosition`) – centred at zero, left negative
 - Whether goal is occluded (`goalOccluded`)
 - How many holes there are in the grid (4, 8, 12)

A Battery of Agents

- 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 goals, following a rigid rule.
- Proximal Policy Optimisation (PPO) Agent
 - Two agents trained on different curricula.
- Dreamer-v3 Agent
 - Two agents trained on different curricula.

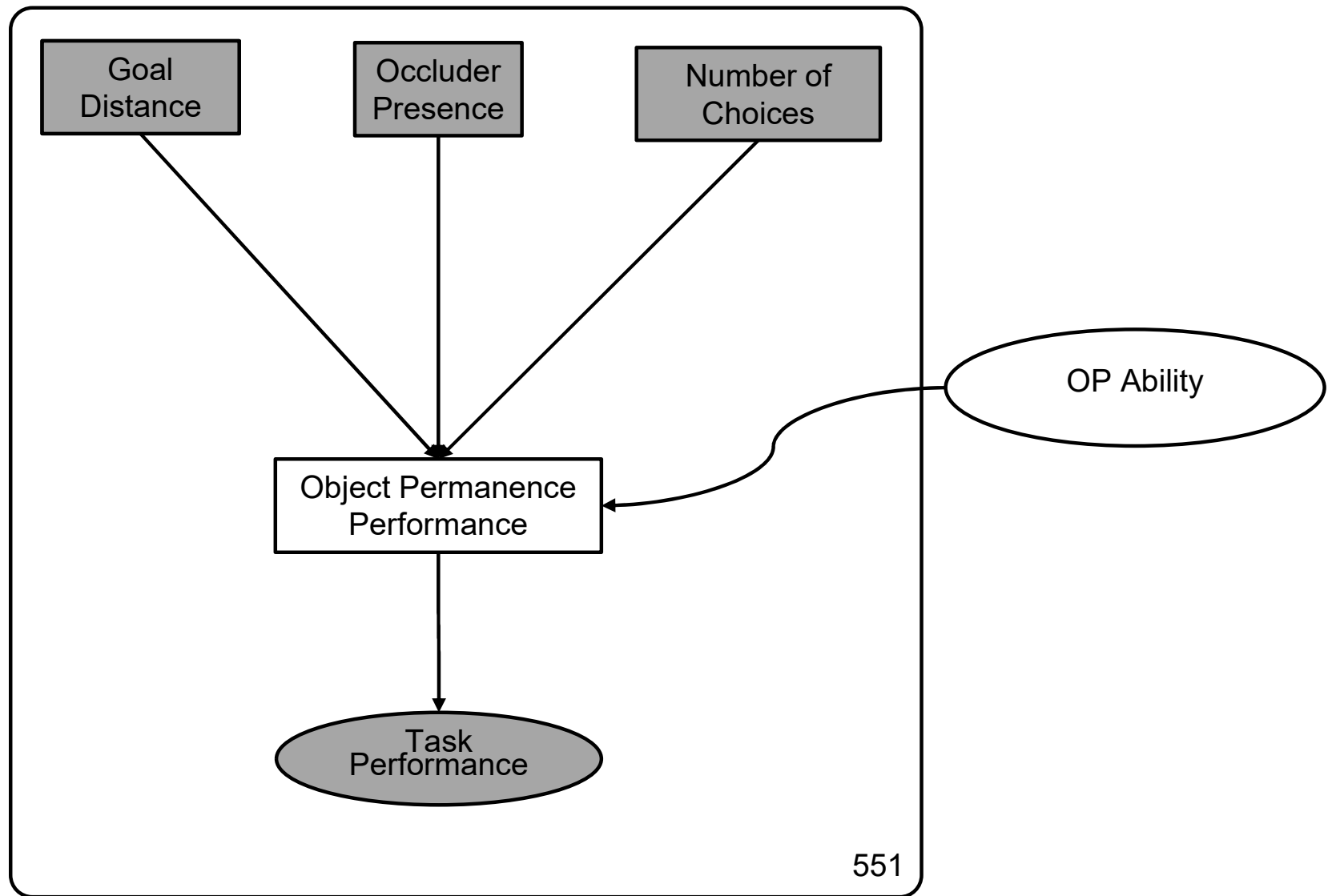
Let's Start Building



How to Define Object Permanence?

- Cognitive science tells us that it is (in part) a working memory task.
- The longer an object is occluded, the harder the task (decaying memory traces)
- The more places an object *could* be occluded, the harder the task (memory substitutability)
- The more objects to be tracked under occlusion, the harder the task (working memory load)

Object Permanence



Introducing Navigation

