PRECIS

From Animals to Algorithms: Comparative Psychology for the Study of Artificial Intelligence

Author: Konstantinos Voudouris
Institution: Department of Psychology, University of Cambridge
Supervisors: Dr Lucy G. Cheke (Department of Psychology), Dr Marta Halina (Department of History & Philosophy of Science)
Examiners: Prof. Nicola Clayton FRS (University of Cambridge), Prof. Cameron Buckner (University of Florida)
Grade: Passed with no corrections.
Date Awarded: 25th October 2024

Konstantinos Voudouris's dissertation *From Animals to Algorithms: Comparative Psychology for the Study of Artificial Intelligence* constitutes an innovative, interdisciplinary approach to the evaluation of artificial intelligence (AI). AI systems are beoming increasingly sophisticated, capable of completing complex tasks such as playing games, recognising images, and producing human-like language. The problem is, we often cannot explain nor predict their behaviour. Comparative psychologists have created methodological tools for studying the behaviour of another class of complex system, non-human animals. While many have advocated for conducting behavioural experiments on AI systems, treating them as though they are participants in the laboratory, the value of a comparative psychological approach has been underappreciated. This dissertation remedies this by addressing the limitations of traditional AI evaluation methods using tools from comparative psychology, laying the foundation for a fruitful research programme in an age of increasingly capable AI. While the focus of the dissertation is the intersection between comparative psychology and AI evaluation, Voudouris also draws on psychometrics, Bayesian statistics, and the philosophy of science to advance the way we empirically investigate what contemporary AI systems can and cannot do.

Thesis Summary

Chapter 1 of the dissertation establishes the foundation for the argument that a comparative psychological approach can significantly enhance the study and understanding of AI systems. The chapter begins with an overview of the rapid advancements in AI, highlighting their impressive capabilities in tasks such as image recognition, language processing, and reinforcement learning, while highlighting how modern AI systems remain largely opaque, with unpredictable and unexplainable behaviour. There have been several calls since the inception of AI as a field to remedy this situation by studying this behaviour experimentally, broadly using the empirical methodologies of the natural sciences. While the past decade has seen considerable work applying tools from human cognitive psychology to study behaviour in computer vision models (Ritter et al., 2017), reinforcement learning agents (Leibo et al., 2018), and large language models (Binz & Schulz, 2023), little emphasis has been placed on applying tools from non-human cognitive psychology (although see Buckner, 2023; Crosby, 2020; Hernández-Orallo, 2017b; Shanahan et al., 2020). The central thesis of this chapter and the dissertation is that comparative psychology can provide valuable tools and methodologies for the study of AI behavior. By leveraging experimental designs and concepts from comparative psychology, researchers can gain deeper insights into the cognitive capabilities of AI systems. The unique contribution of comparative psychology, compared to human psychology, is that its tools and methodologies have been developed with the explicit appreciation that the subjects of study are not like us, and that many of our assumptions are anthropomorphic, anthropocentric, and should be challenged empirically. This approach has several potential benefits, including more meaningful evaluations, improved interpretability of AI behavior, and the development of more sophisticated and robust AI systems.

Chapter 2 critiques the standard practice of AI evaluation, which is to produce larger and more general datasets and benchmarks which purport to test a number of capabilities simultaneously. Voudouris defines a capability following a common definition in the philosophy of psychology (where it usually termed a *capacity*), namely, the disposition to complete tasks with certain features (Cummins, 2000, pp. 122–123). This account permits a large range in the generality of a capability, depending on the features that are considered. For instance, a system can possess the disposition to navigate a maze with either a fixed size or any finite size, two distinct but related capabilities. These capabilities vary in terms of their generality. Voudouris argues that the standard practice of evaluation provides poor tools for testing the capabilities of a system. First, these datasets and benchmarks have low validity – they often do not measure the capabilities that they are intended to measure (see also Hernández-Orallo, 2017a; Raji et al., 2021). Second, they lack adequate control conditions, meaning that there are multiple competing and equally plausible explanations for behaviour that invoke distinct and incompatible capabilities. Third, researchers often aggregate performance across benchmarks into a handful of summary statistics. At best, these measures are ordinal, meaning that we can make judgments about whether a system is better or worse at a task, but not a judgment about how *much* better they are. This would demand interval or ratio measurements, in which magnitudes are meaningful, which are not provided by the standard practice of AI evaluation. These arguments serve as the basis for the contributions of the remaining chapters of the dissertation.

Chapter 3 introduces one of the major contributions of the dissertation, the Animal-AI Environment, which was co-designed by the author. This is a platform designed for applying comparative psychological methods to the evaluation of AI systems, constituting a *virtal laboratory* for direct comparison between humans, animals, and embodied AI systems. Animal-AI provides a controlled setting where AI agents can be tested on cognitive tasks similar to those used in animal behavior studies. Chapter 3 presents a comprehensive overview of the environment and how it can be used to advance interdisciplinary research at the intersection of cognitive science and AI, demonstrating its utility in three reinforcement learning experiments. An example of a task built in Animal-AI is presented in Figure 1. This task conceptually replicates operant chamber tasks common in associative learning and comparative psychology (Skinner, 1938). The agent is a unit sphere spawned in a 40x40x40 arena. The agent is able to move forwards, backwards, and rotate, and it can be controlled by both a computational model (e.g., reinforcement learning) and by human players through a controller. The agent receives observations of the environment, including pixel images of what is in front of them. The arena can be populated with a number of objects explicitly designed for replicating comparative psychology experiments, including rewarding objects and buttons, but also dispensers, walls, ramps, and tunnels. The author directly contributed to the design of the objects in the environment and the interfaces for humans and a number of reinforcement learning libraries, including *stable-baselines* and *Dreamer*. This means that Animal-AI has broad appeal and is easy to use for psychologists and machine learning researchers alike.

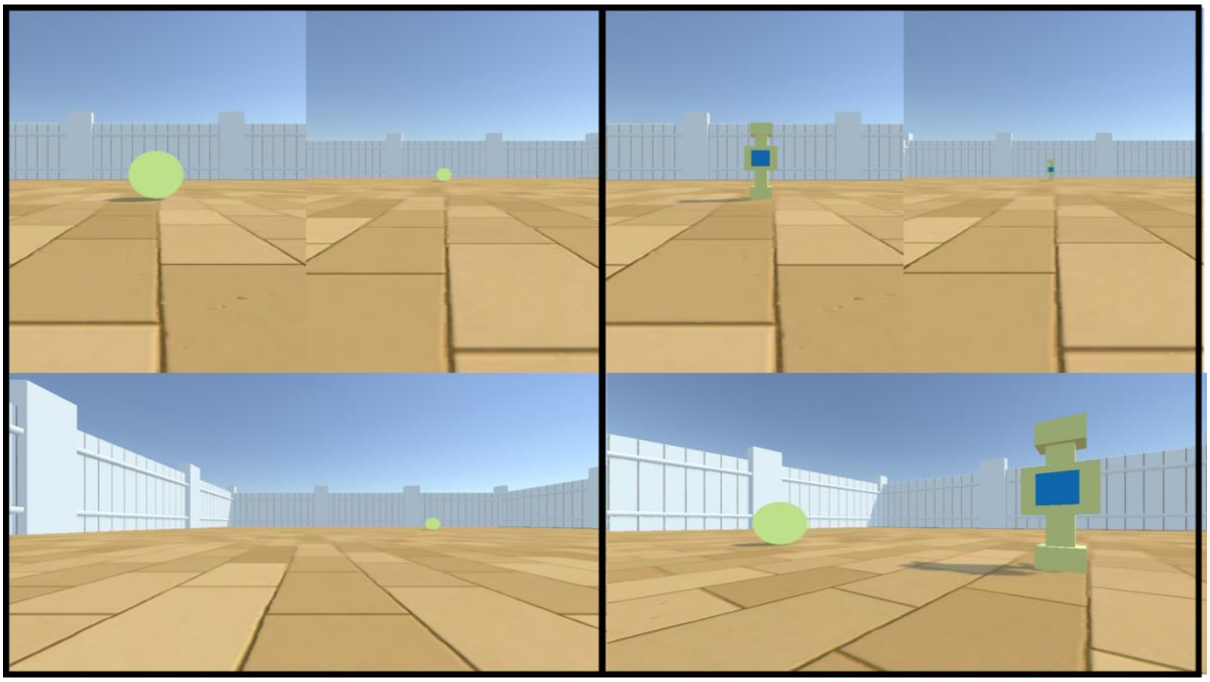


Figure 1 – A button-press task in Animal-AI, replicating operant chamber tasks in comparative psychology. ***Left****: Rewarding green spheres placed at different distances from the agent.* ***Right****: The button that can be pressed to produce a rewarding green sphere. The agent must learn to press the button to produce the reward.*

Chapter 4 puts Animal-AI to use for studying a key aspect of physical cognition in embodied agents, namely, object permanence, the capability to track objects while they are occluded. *Object Permanence In Animal-AI: Generalisable Suites* (O-PIAAGETS) is a battery of experiments for studying object permanence in embodied agents, conceptually replicating hundreds of experimental designs from comparative and developmental psychology. It contains over 12,500 distinct testing instances along with over 9,000 control conditions, serving as a significant improvement over other benchmarks for assessing object permanence in terms of both size and validity. It can be used to study object permanence in humans, reinforcement learning agents, and generally any system capable of producing actions in the environment.

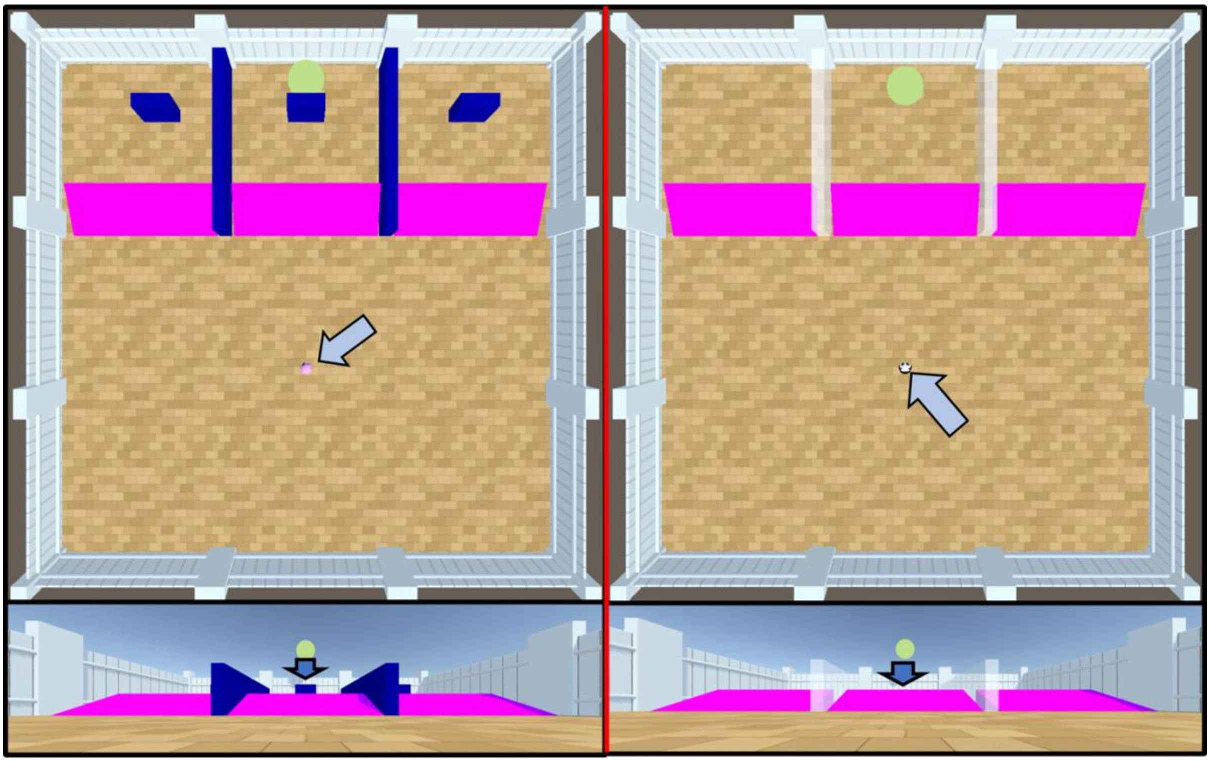


Figure 2 - A three-cup task from O-PIAAGETS, conceptually replicating the cup tasks in the Primate Cognition Test Battery (Herrmann et al., 2007). The agent is indicated by the grey arrows. The movement of green spheres is indicated by the blue arrows. ***Left****: An object permanence version of this task. The rewarding green sphere drops from a height behind the opaque blue wall.* ***Right****: A control version of this task, where the sphere is not occluded. In both cases, navigating over the incorrect ramp means that the agent cannot escape and loses all its points.*

Figure 2 shows a conceptual replication of a task from the Primate Cognition Test Battery (Herrmann et al., 2007). In the original task, a rewarding item is hidden in one of three cups. The participant receives the item, usually a piece of food or a toy, if they correctly identify which cup it is hidden under. If they choose the incorrect cup, they receive no reward. In Animal-AI, this is conceptually replicated with a rewarding green sphere dropping behind an opaque wall behind a ramp. If the agent navigates over the incorrect ramp, they cannot exit and they cannot obtain the rewarding item. The right panel of Figure 2 is an example of a control condition, where the reward is not occluded

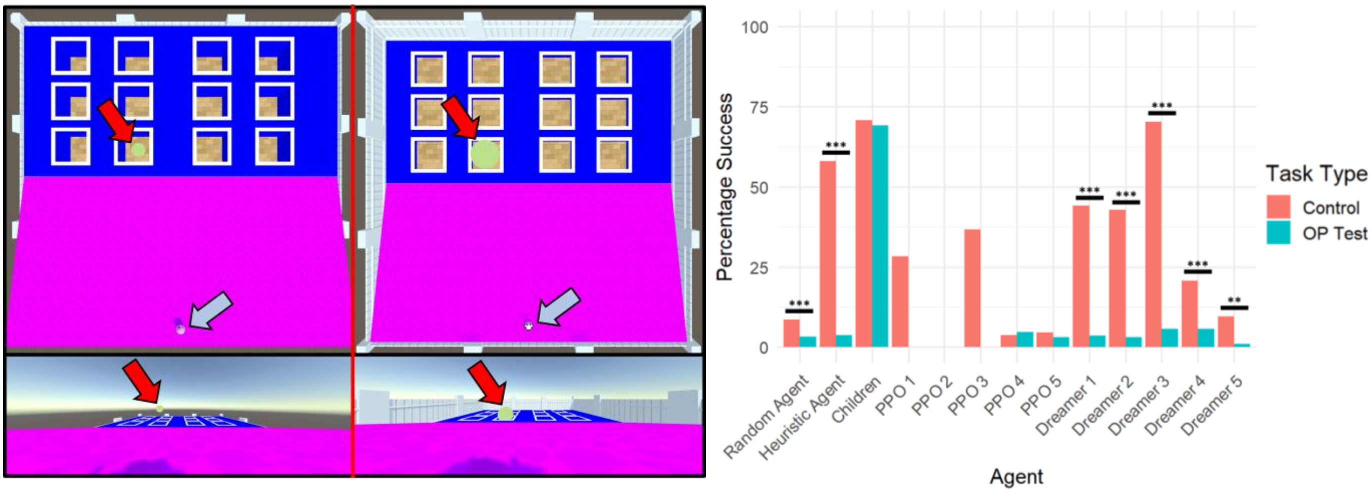


*Figure 3 – **Left**: The grid task, which goals indicated by the red arrows and the agent's position indicated by the grey arrows. The right panel shows a control condition, where the goal is visible from the agent's starting position. The left panel shows an object permanence test condition, where the goal drops from a height and is occluded. **Right**: Results on these tasks for 13 agents, including Children and two deep reinforcement learning agents (PPO and Dreamer) trained on five different curricula each.*

while the agent is making the choice (it can be seen from the top of the ramp). Good performance on the control conditions and poor performance on the object permanence conditions suggests that the agent lacks object permanence.

Figure 3 presents results on a similar task from O-PIAAGETS, reported in Chapter 6. This task, called the Grid Task, involves a single rewarding green sphere dropping from a height into one of a number of holes in the ground. In the control condition, the goal is (partially) visible once it has dropped, whereas in the object permanence condition, it is occluded. Thirteen types of agenta were evaluated on over 400 variants of this task. Baselines were established by a random agent, which takes randomised actions in the environment, a heuristic agent, which navigates towards visible rewards, and children aged 4-7 years old. Two deep reinforcement learning algorithms were also evaluated, Proximal Policy Optimisation (PPO; Schulman et al., 2017) and Dreamer-v3 (Hafner et al., 2023), trained on five different curricula. It is clear from the performance of the Dreamer agents that they are capable of completing many control tasks, but as soon as objects are occluded, their performance is significantly worse, suggesting that they lack the object permanence capability. This contrasts with children, who perform similarly on both conditions.

While O-PIAAGETS constitutes an advancement in object permanence testing in the context of the standard practice of AI evaluation, it still invites researchers to aggregate performance across the whole benchmark, leading to ordinal, rather than interval or ratio, measurements. Chapters 5 and 6 improve this situation by introducing the *measurement layout approach*, a statistical paradigm inspired by (Multivariate) Item Response Theory and psychometrics (Reckase, 2009). Measurement layouts are

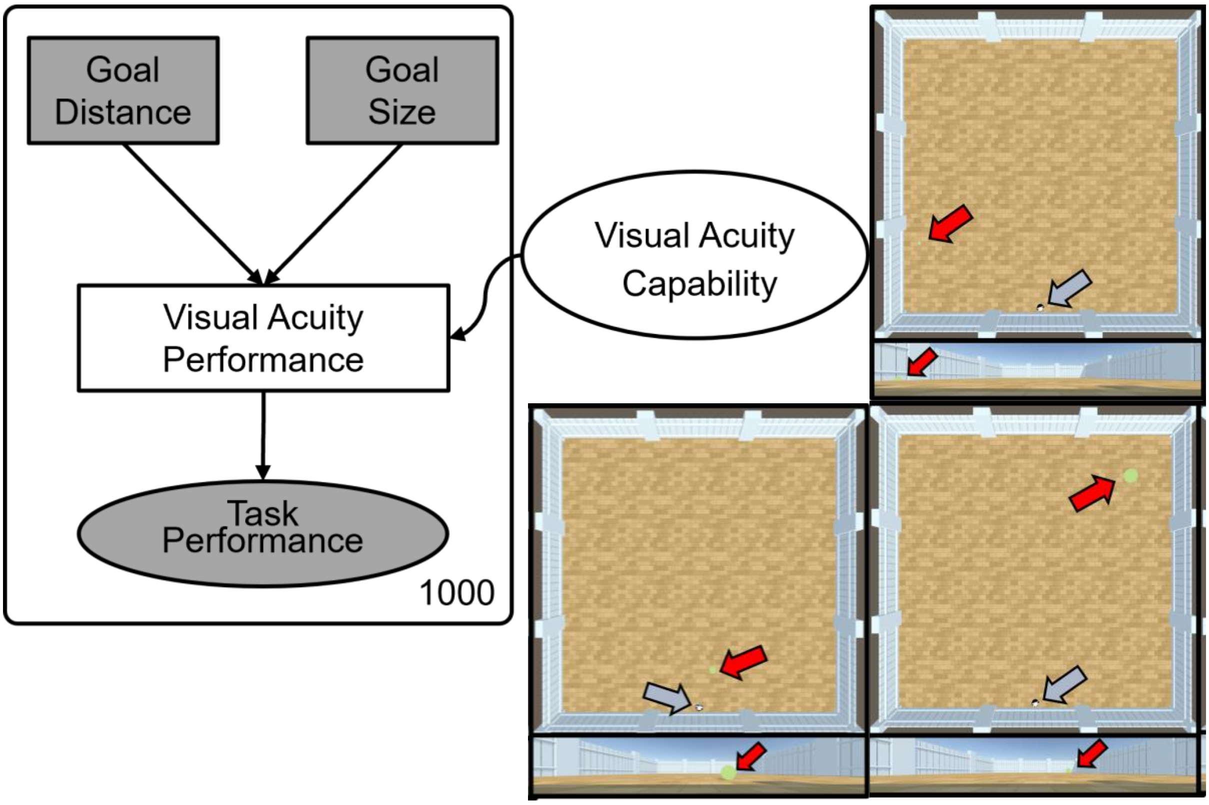
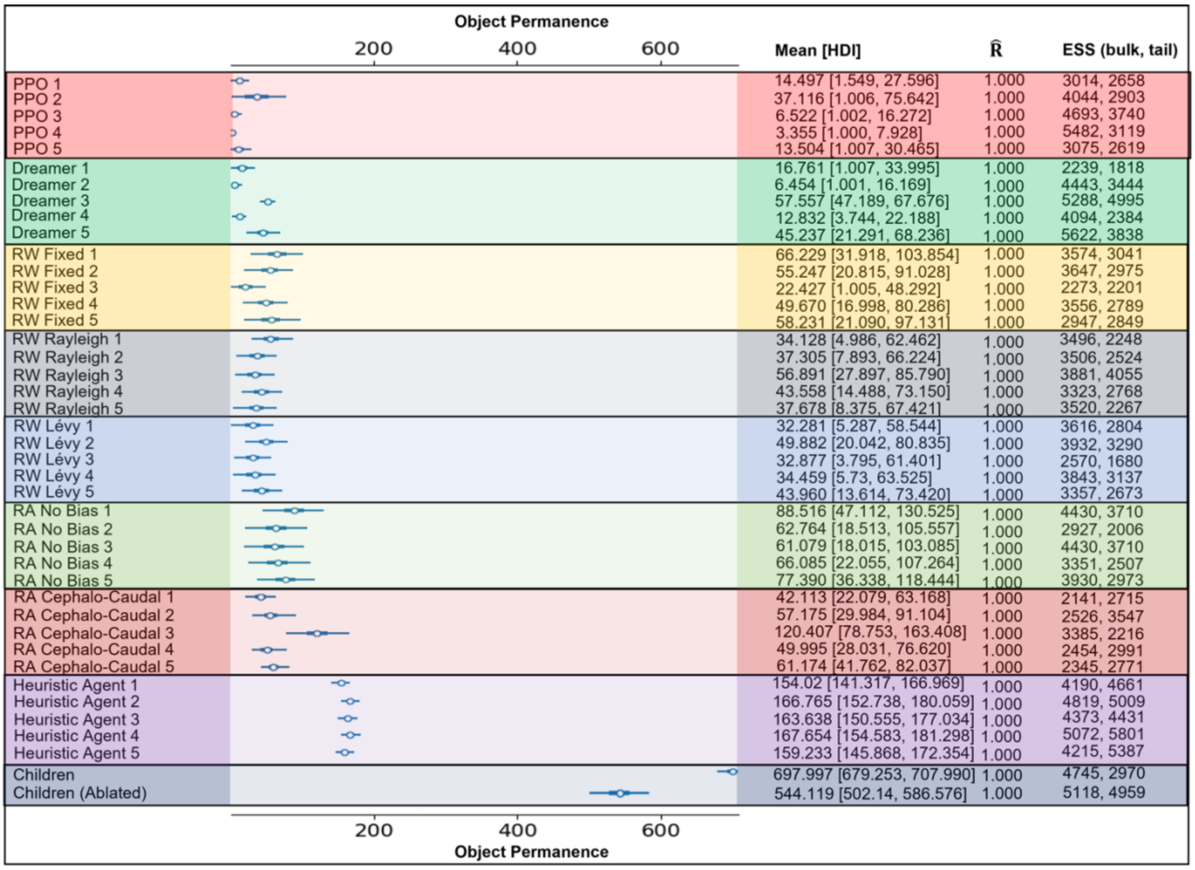


*Figure 4 – **Left**: A measurement layout relating two task features to task performance, via a latent visual acuity capability. **Right**: Three tasks in Animal-AI varying the size and distance of the rewarding green sphere (the goal).*

hierarchical Bayesian networks that relate the properties of single testing instances with agent performance via intermediary latent capabilities (Burden et al., 2023).

Chapter 5 introduces these models through a simple example. An agent in Animal-AI is tasked with obtained rewarding green spheres of varying sizes and at varying distances from its starting position (see Figure 4). The agent's disposition to solve these tasks can be characterised by its *visual acuity*: how well it can see small and more distant objects. By taking an agent's performance across a battery of tasks with different goal sizes and goal distances, measurement layouts can be used to infer a latent visual acuity capability. Since measurement layouts are Bayesian, capabilities are estimated as probability distributions over measures defined in terms of observable features of the task, in this case, the combination of goal size and goal distance. This allows us to compute the probability that an agent will pass a particular task given a goal's size and its distance from the agent. Measurement layouts can be scaled up to disentangle multiple latent capabilities that are sensitive to different features of a task and all differentially contribute to success. Furthermore, they enable us to integrate theoretical knowledge from the cognitive sciences into the measurement procedure. Finally, the resulting capability estimates, operationalised as posterior probability distributions, are given in interpretable units as defined by the task demands. This means that the magnitudes of the capability estimates are meaningful and can be used in comparisons between systems and species, thus serving as interval and/or ratio measurements.

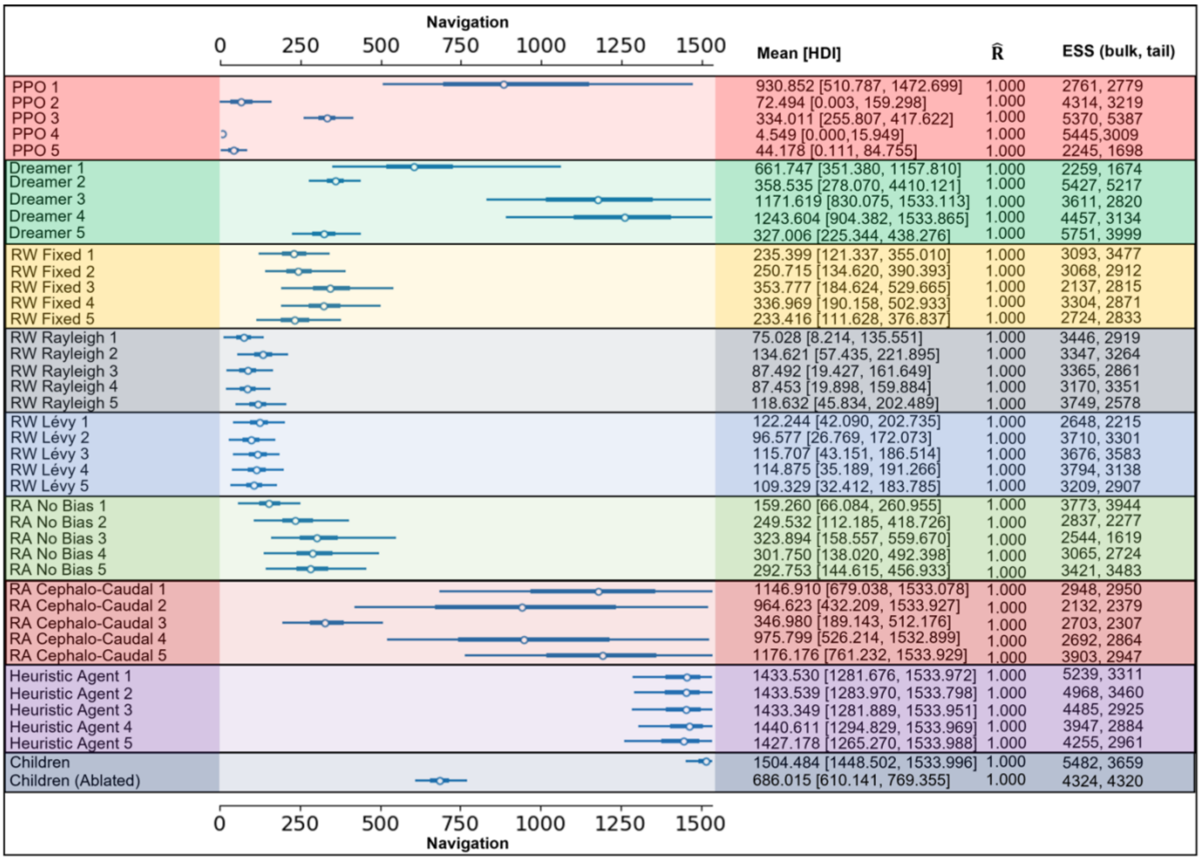
Chapter 6 presents an extensive analysis of the capabilities of agents on a subset of O-PIAAGETS, drawing both on classical methodologies in experimental psychology as well as the new measurement layout approach. Forty computational agents, including PPO and Dreamer-v3, were evaluated on over 4,200 O-PIAAGETS tasks and compared to children aged 4-7 years old. Measurement layouts were used to disentangle thirteen latent capabilities, including object permanence and navigation, producing probability densities over measures with interpretable units.



	Mean [HDI]	$\hat{R}$	ESS (bulk, tail)
PPO 1	14.497 [1.549, 27.596]	1.000	3014, 2658
PPO 2	37.116 [1.006, 75.642]	1.000	4044, 2903
PPO 3	6.522 [1.002, 16.272]	1.000	4693, 3740
PPO 4	3.355 [1.000, 7.928]	1.000	5482, 3119
PPO 5	13.504 [1.007, 30.465]	1.000	3075, 2619
Dreamer 1	16.761 [1.007, 33.995]	1.000	2239, 1818
Dreamer 2	6.454 [1.001, 16.169]	1.000	4443, 3444
Dreamer 3	57.557 [47.189, 67.676]	1.000	5288, 4995
Dreamer 4	12.832 [3.744, 22.188]	1.000	4094, 2384
Dreamer 5	45.237 [21.291, 68.236]	1.000	5622, 3838
RW Fixed 1	66.229 [31.918, 103.854]	1.000	3574, 3041
RW Fixed 2	55.247 [20.815, 91.028]	1.000	3647, 2975
RW Fixed 3	22.427 [1.005, 48.292]	1.000	2273, 2201
RW Fixed 4	49.670 [16.998, 80.286]	1.000	3556, 2789
RW Fixed 5	58.231 [21.090, 97.131]	1.000	2947, 2849
RW Rayleigh 1	34.128 [4.986, 62.462]	1.000	3496, 2248
RW Rayleigh 2	37.305 [7.893, 66.224]	1.000	3506, 2524
RW Rayleigh 3	56.891 [27.897, 85.790]	1.000	3881, 4055
RW Rayleigh 4	43.558 [14.488, 73.150]	1.000	3323, 2768
RW Rayleigh 5	37.678 [8.375, 67.421]	1.000	3520, 2267
RW Lévy 1	32.281 [5.287, 58.544]	1.000	3616, 2804
RW Lévy 2	49.882 [20.042, 80.835]	1.000	3932, 3290
RW Lévy 3	32.877 [3.795, 61.401]	1.000	2570, 1680
RW Lévy 4	34.459 [5.73, 63.525]	1.000	3843, 3137
RW Lévy 5	43.960 [13.614, 73.420]	1.000	3357, 2673
RA No Bias 1	88.516 [47.112, 130.525]	1.000	4430, 3710
RA No Bias 2	62.764 [18.513, 105.557]	1.000	2927, 2006
RA No Bias 3	61.079 [18.015, 103.085]	1.000	4430, 3710
RA No Bias 4	66.085 [22.055, 107.264]	1.000	3351, 2507
RA No Bias 5	77.390 [36.338, 118.444]	1.000	3930, 2973
RA Cephalo-Caudal 1	42.113 [22.079, 63.168]	1.000	2141, 2715
RA Cephalo-Caudal 2	57.175 [29.984, 91.104]	1.000	2526, 3547
RA Cephalo-Caudal 3	120.407 [78.753, 163.408]	1.000	3385, 2216
RA Cephalo-Caudal 4	49.995 [28.031, 76.620]	1.000	2454, 2991
RA Cephalo-Caudal 5	61.174 [41.762, 82.037]	1.000	2345, 2771
Heuristic Agent 1	154.02 [141.317, 166.969]	1.000	4190, 4661
Heuristic Agent 2	166.765 [152.738, 180.059]	1.000	4819, 5009
Heuristic Agent 3	163.638 [150.555, 177.034]	1.000	4373, 4431
Heuristic Agent 4	167.654 [154.583, 181.298]	1.000	5072, 5801
Heuristic Agent 5	159.233 [145.868, 172.354]	1.000	4215, 5387
Children	697.997 [679.253, 707.990]	1.000	4745, 2970
Children (Ablated)	544.119 [502.14, 586.576]	1.000	5118, 4959

Figure 5 - Forest plots showing the mean and highest-density intervals (HDIs) of the object permanence capability distributions for 41 agents, disentangled from 6 other latent capabilities. RW and RA refer to different types of random agents (random walker vs. random action).

Figure 5 presents summaries of the probability densities of the latent object permanence capabilities extracted from 42 fitted measurement layouts. The measurement is in units of the manhattan distance the agent must travel to obtain the goal multiplied by the number of places the goal could be hidden in. Goal distance is used as a proxy for the amount of time the goal is occluded for. Clearly, all computational agents have noticeably lower object permanence capabilities than children. In contrast, Figure 6 presents the same summaries for the latent navigation capability. A number of agents, including the rule-following heuristic agents and two Dreamer agents have high navigation capabilities, measured in terms of the distance and tortuosity of the route to the goal from the agent's starting point. These capability profiles cannot be inferred from aggregating performance across different testing conditions, demonstrating a key contribution of measurement layouts. A nuanced capability profile can be inferred from testing an agent on a well-annotated and designed benchmark of tasks, and redundancy is minimised because every task can be used to measure multiple capabilities at once.



	Mean [HDI]	$\hat{R}$	ESS (bulk, tail)
PPO 1	930.852 [510.787, 1472.699]	1.000	2761, 2779
PPO 2	72.494 [0.003, 159.298]	1.000	4314, 3219
PPO 3	334.011 [255.807, 417.622]	1.000	5370, 5387
PPO 4	4.549 [0.000,15.949]	1.000	5445,3009
PPO 5	44.178 [0.111, 84.755]	1.000	2245, 1698
Dreamer 1	661.747 [351.380, 1157.810]	1.000	2259, 1674
Dreamer 2	358.535 [278.070, 4410.121]	1.000	5427, 5217
Dreamer 3	1171.619 [830.075, 1533.113]	1.000	3611, 2820
Dreamer 4	1243.604 [904.382, 1533.865]	1.000	4457, 3134
Dreamer 5	327.006 [225.344, 438.276]	1.000	5751, 3999
RW Fixed 1	235.399 [121.337, 355.010]	1.000	3093, 3477
RW Fixed 2	250.715 [134.620, 390.393]	1.000	3068, 2912
RW Fixed 3	353.777 [184.624, 529.665]	1.000	2137, 2815
RW Fixed 4	336.969 [190.158, 502.933]	1.000	3304, 2871
RW Fixed 5	233.416 [111.628, 376.837]	1.000	2724, 2833
RW Rayleigh 1	75.028 [8.214, 135.551]	1.000	3446, 2919
RW Rayleigh 2	134.621 [57.435, 221.895]	1.000	3347, 3264
RW Rayleigh 3	87.492 [19.427, 161.649]	1.000	3365, 2861
RW Rayleigh 4	87.453 [19.898, 159.884]	1.000	3170, 3351
RW Rayleigh 5	118.632 [45.834, 202.489]	1.000	3749, 2578
RW Lévy 1	122.244 [42.090, 202.735]	1.000	2648, 2215
RW Lévy 2	96.577 [26.769, 172.073]	1.000	3710, 3301
RW Lévy 3	115.707 [43.151, 186.514]	1.000	3676, 3583
RW Lévy 4	114.875 [35.189, 191.266]	1.000	3794, 3138
RW Lévy 5	109.329 [32.412, 183.785]	1.000	3209, 2907
RA No Bias 1	159.260 [66.084, 260.955]	1.000	3773, 3944
RA No Bias 2	249.532 [112.185, 418.726]	1.000	2837, 2277
RA No Bias 3	323.894 [158.557, 559.670]	1.000	2544, 1619
RA No Bias 4	301.750 [138.020, 492.398]	1.000	3065, 2724
RA No Bias 5	292.753 [144.615, 456.933]	1.000	3421, 3483
RA Cephalo-Caudal 1	1146.910 [679.038, 1533.078]	1.000	2948, 2950
RA Cephalo-Caudal 2	964.623 [432.209, 1533.927]	1.000	2132, 2379
RA Cephalo-Caudal 3	346.980 [189.143, 512.176]	1.000	2703, 2307
RA Cephalo-Caudal 4	975.799 [526.214, 1532.899]	1.000	2692, 2864
RA Cephalo-Caudal 5	1176.176 [761.232, 1533.929]	1.000	3903, 2947
Heuristic Agent 1	1433.530 [1281.676, 1533.972]	1.000	5239, 3311
Heuristic Agent 2	1433.539 [1283.970, 1533.798]	1.000	4968, 3460
Heuristic Agent 3	1433.349 [1281.889, 1533.951]	1.000	4485, 2925
Heuristic Agent 4	1440.611 [1294.829, 1533.969]	1.000	3947, 2884
Heuristic Agent 5	1427.178 [1265.270, 1533.988]	1.000	4255, 2961
Children	1504.484 [1448.502, 1533.996]	1.000	5482, 3659
Children (Ablated)	686.015 [610.141, 769.355]	1.000	4324, 4320

Figure 6 - Forest plots showing the mean and highest-density intervals (HDIs) of the object permanence capability distributions for 41 agents, disentangled from 6 other latent capabilities. RW and RA refer to different types of random agents (random walker vs. random action).

Part II of the dissertation turns to methodological and philosophical questions in the study of non-human behaviour, in an effort to more completely characterise the challenges of behaviourally evaluating the capabilities of AI systems. In its over 100 year history, scientific comparative psychology has grappled with a number of issues in turning the tools of human psychology over to the study of non-human animals. To what extent are scientists biased by their own intuitions and preconceptions about behaviour when interpreting what animals can and cannot do? To what extent are measures used on humans valid when used with animals? What sorts of hypotheses should be on the table when trying to explain how animals behave? Chapters 7, 8, and 9, present extensive case studies of methodological challenges in comparative psychology, before drawing lessons for a comparative psychological approach to AI evaluation.

Chapter 7 for how hypotheses are generated in comparative psychology, and in particular, the role of analogical reasoning in comparative psychology. Analogies drawn from human psychology and associative learning are conduits for transferring knowledge from well-studied phenomena to less understood ones. In particular, Voudouris argues that analogies justify that hypotheses are worthy of pursuit because of their established explanatory potential. The use of analogies also sheds light on the longstanding debate about the existence of a distinction between associative learning and cognition. A novel account of the ontogeny of this distinction is offered, namely, that the apparent distinction in the scientific literature does not necessarily track a difference in behavioural processes, but is rather an artefact of the analogical reasoning processes which gave rise to hypotheses in the literature. By framing the associative-cognitive distinction as a product of analogical reasoning, this chapter provides a more charitable interpretation of comparative psychology than is usually offered by philosophers of science. It suggests that the field's reliance on analogies from human psychology and associative learning does not necessarily imply an endorsement of a problematic dichotomy between behavioral processes. Furthermore, analogical reasoning is presented as a justifiable strategy for hypothesis generation in comparative psychology, as well as in AI evaluation.

Chapter 8 turns to the practice of preferring simpler hypotheses in comparative psychology, often referred to as Morgan's Canon (1894; 1903): *In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development.* Due to the ubiquity of this line of thinking, hypotheses that appeal to simpler processes, such as associative learning, tend to be taken as simple default explanations of an animal's behaviour. However, philosophers of science have long pointed out the lack of evidence for any measure of the simplicity of behavioural processes that would justify such inferences (Fitzpatrick, 2008, 2017; Meketa, 2014). This chapter accepts these arguments, but claims that the simplicity of behavioural processes is a useful idealisation for generating alternative hypotheses. While there may be no evidence that associative learning is simpler than, say, episodic memory or tool use, thinking in these terms assists comparative psychologists in generating plausible alternative hypotheses to test empirically, thus idealising away from the complexities of animal behaviour (see Potochnik, 2017; Weisberg, 2007). This is particularly useful in cases where evidence is limited and there is a vast potential hypothesis space, as in the case of non-human animal research and, incidentally, AI evaluation. As pointed out in the philosophy of science literature scientists often need heuristics and cognitive aids for generating hypotheses in such situations, and idealisations about the simplicity of behavioural processes are one example.

Chapter 9 collates the methodological challenges discussed in the preceding two chapters and puts them to 220 practicing comparative psychologists in a survey. The questions probed their preferences for simpler hypotheses and their views on the apparent distinction between associative learning and cognition. The results are synthesised with the analyses across the dissertation, presenting key methodological considerations for studying the behaviour of AI systems. First, AI Evaluation must pay close attention to the problem of contrastive underdetermination, which refers to the situation where several incompatible explanations are consonant with the available evidence. Comparative psychologists have innovated several methods for generating novel alternative hypotheses to explain non-human animal behaviour that challenge anthropocentric and anthropomorphic assumptions. Second, AI Evaluation must blend theory and experiment to produce valid measures of the capabilities being targeted, and engage in critical debate about those measures in open fora, as has been practiced by comparative psychologists for several decades. With an eye on the methods of AI Evaluation, we can advance our understanding of AI capabilities without succumbing to the pitfalls concomitant with studying behavioural systems very different from humans.

Interdisciplinary Contribution

From Animals to Algorithms presents an interdisciplinary approach to evaluating artificial intelligence by drawing on the methodologies of comparative psychology. The increasing behavioural sophistication of AI systems, and the concurrent difficulty of explaining their behaviour, may well become one of the fundamental challenges of cognitive science in the future. Recognising this, Voudouris argues that the tools developed to study animal cognition offer valuable insights. This dissertation makes several contributions. First, it presents a comprehensive critique of the prevailing reliance on large datasets and benchmarks in AI evaluation, highlighting their limitations in validity and measurement rigor. Second, it presents the most comprehensive and up-to-date overview of the Animal-AI Environment, a platform unique in its mission to unify comparative psychology and AI research which facilitates direct comparisons between humans, animals, and AI agents on shared cognitive tasks. Animal-AI, coupled with the development of specific test suites like O-PIAAGETS for assessing object permanence, enables the application of comparative psychological methods in a controlled digital setting. Third, this dissertation presents a novel statistical paradigm inspired by psychometrics and Bayesian statistics, to move beyond simple rankings of AI performance and achieve more nuanced, quantifiable measurements of capabilities. Finally, this dissertation makes contributions to the philosophy of science in its discussion of hypothesis generation strategies in comparative psychology. This synthesis of comparative psychology, computer science, statistics, and philosophy offers a novel and much-needed framework for evaluating AI in a way that integrates into the cognitive sciences.

The research outputs of this dissertation have been, and continue to be, published at venues across comparative psychology, machine learning, artificial intelligence, and philosophy of science. A selected subset of these works is given immediately below:

- 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. *EBeM'22: Workshop on AI Evaluation Beyond Metrics, IJCAI, July 25, 2022, Vienna, Austria*.
- Voudouris, K., Farrar, B. G., Cheke, L. G., & Halina, M. (*forthcoming*). Morgan's Canon and the Associative-Cognitive Distinction Today: A Survey of Practitioners. *Journal of Comparative Psychology*.
- Voudouris, K., Liu, J. D., Siwinska, N., Schellaert, W., & Cheke, L. G. (2024). Investigating Object Permanence in Deep Reinforcement Learning Agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).
- Burden, J., Voudouris, K., Burnell, R., Rutar, D., Cheke, L., & Hernández-Orallo, J. (*under review*). Inferring Capabilities from Task Performance with Bayesian

Triangulation. *Journal of Artificial Intelligence Research*. (*arXiv preprint arXiv:2309.11975*).

- Voudouris, K. et al. (*under review*) The Animal-AI Environment: A Virtual Laboratory For Comparative Cognition and Artificial Intelligence Research. *Behavior Research Methods*. (*arXiv preprint arXiv:2312.11414*).
- Voudouris, K. (*under review*). Analogies and the Associative-Cognitive Distinction in Comparative Psychology. *Biology & Philosophy*.
- Voudouris, K. (*under review*). Cognitive Simplicity as an Idealisation. *Erkenntnis*.

Future Directions

The dissertation opens up several future research opportunities. The Animal-AI Environment continues to be developed as a research tool for cognitive scientists and AI researchers to work together on mutual research problems. Measurement layouts constitute a new paradigm for measuring cognitive capabilities in machines and humans, facilitating direct comparison between them and a more nuanced appraisal of our progress towards intelligent machines. They are now being extended and applied to other classes of system, including large language models. Novel philosophical analyses of the science of comparative psychology advance the debates on the associative-cognitive distinction, Morgan's Canon & principles of parsimony, and hypothesis generation strategies. Combined, this dissertation represents a significant step forward in the interdisciplinary study of Artificial Intelligence, laying the foundation for a robust, comparative psychological approach to AI Evaluation.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Buckner, C. (2023). Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour. *The British Journal for the Philosophy of Science*, *74*(3), 681–712.

Burden, J., Voudouris, K., Burnell, R., Rutar, D., Cheke, L., & Hernández-Orallo, J. (2023). *Inferring Capabilities from Task Performance with Bayesian Triangulation* (arXiv:2309.11975). arXiv. http://arxiv.org/abs/2309.11975

Crosby, M. (2020). Building Thinking Machines by Solving Animal Cognition Tasks. *Minds and Machines*, *30*(4), 589–615. https://doi.org/10.1007/s11023-020-09535-6

Cummins, R. (2000). *How does it work? Versus What are the laws? Two conceptions of psychological explanation*.

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. *Mind & Language*, *23*(2), 224–246.

Fitzpatrick, S. (2017). Against Morgan's Canon. In K. Andrews & J. Beck (Eds.), *The Routledge Handbook of Philosophy of Animal Minds*. Routledge.

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). *Mastering Diverse Domains through World Models*. arXiv preprint arXiv:2301.04104.

Hernández-Orallo, J. (2017a). Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, *48*(3), 397–447.

Hernández-Orallo, J. (2017b). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.

Herrmann, E., Call, J., Hernàndez-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.

Leibo, J. Z., d'Autume, C. de M., Zoran, D., Amos, D., Beattie, C., Anderson, K., Castañeda, A. G., Sanchez, M., Green, S., Gruslys, A., Legg, S., Hassabis, D., & Botvinick, M. M. (2018). *Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents*. arXiv preprint arXiv:1801.08116.

Meketa, I. (2014). A critique of the principle of cognitive simplicity in comparative cognition. *Biology & Philosophy*, *29*(5), 731–745.

Potochnik, A. (2017). Idealization and the Aims of Science. In *Idealization and the Aims of Science*. University of Chicago Press.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). *AI and the Everything in the Whole Wide World Benchmark*. arXiv preprint arXiv:2111.15366).

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. *Proceedings of the 34th International Conference on Machine Learning*, 2940–2949.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shanahan, M., Crosby, M., Beyret, B., & Cheke, L. (2020). Artificial Intelligence and the Common Sense of Animals. *Trends in Cognitive Sciences*, *24*(11), 862–872.

Skinner, B. F. (1938). *The Behavior Of Organisms: An Experimental Analysis* (R. M. Elliott, Ed.). Appleton-Century-Crofts, Inc.

Weisberg, M. (2007). Three Kinds of Idealization. *The Journal of Philosophy*, *104*(12), 639–659.