UNIVERSITY OF CAMBRIDGE

Economic and Social Research Council

kv301@cam.ac.uk

Direct Human-AI Comparison in the Animal-AI Environment

Konstantinos Voudouris[1 2], Matthew Crosby[1 3], Benjamin Beyret[1 4], José Hernández-Orallo[1 5], Murray Shanahan[1 3 4], Marta Halina[1 2], Lucy Cheke[1 2].

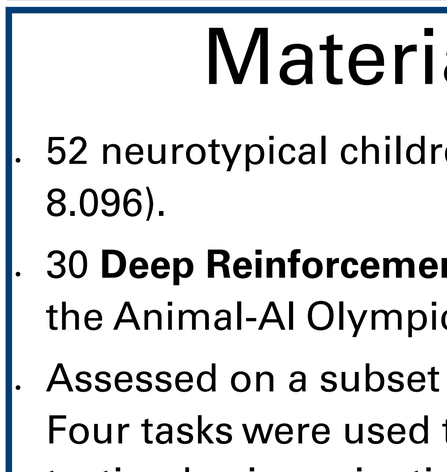
[1]Leverhulme Centre for the Future of Intelligence, UK, [2]University of Cambridge, UK, [3]Imperial College London, UK, [4]DeepMind, [5]Universitat Politècnica de València, Spain.

Introduction

- The field of **Artificial Intelligence (AI)** is making rapid progress, passing several benchmarks (e.g. Chess, Go, StarCraft II) previously thought to be too complex for current tech.
- Current AI benchmarks do not tell us much about the actual cognitive ability of AI systems.
- AI systems have also been shown to **lack robustness**, failing to generalise beyond their training distribution (e.g. Dong et al. 2018; Chollet 2019).
- Some suggest that AI systems are not **intelligently** solving these problems, but are merely taking 'shortcuts' (Geirhos et al. 2020).
- **Comparative psychology** has been developing experimental paradigms for telling between 'behavioural flexibility' and *Clever Hans Effects* for more than a century.
- The **Animal-AI Environment and Testbed** were developed to test AI systems on cognitive ability using these paradigms.
- Here we provide the first direct human-AI comparison in the Animal-AI Environment, facilitating a **mutually-beneficial dialogue** between AI and cognitive science.

Materials and Methods

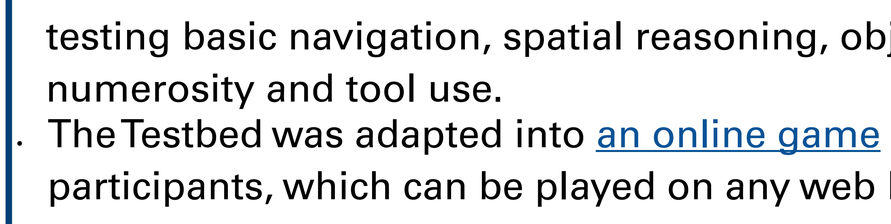
- 52 neurotypical children **aged 6-10** (median age=8, mean age = 8.096).
- 30 **Deep Reinforcement Learning (DRL) systems**, submitted to the Animal-AI Olympics Competition 2019.
- Assessed on a subset of 40 tasks from the Animal-AI Testbed. Four tasks were used to **test 10 different abilities**, including tasks testing basic navigation, spatial reasoning, object permanence, numerosity and tool use.
- The Testbed was adapted into an online game for human participants, which can be played on any web browser.



Top Left – A simple navigation task, where the goal is to move towards and touch the reward.

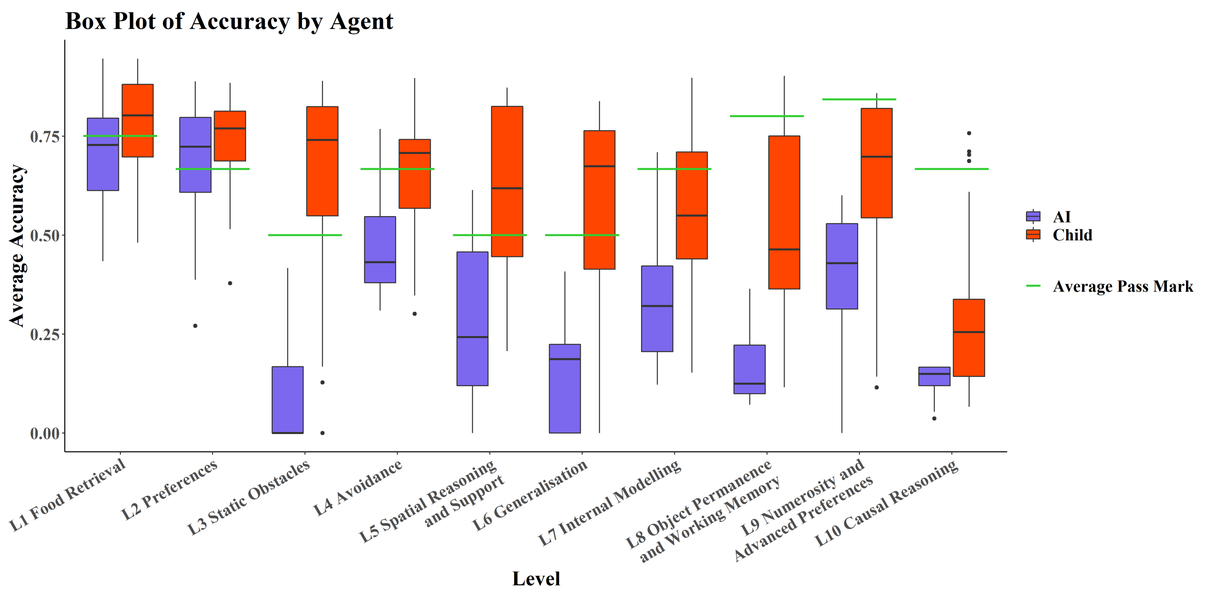
Bottom Left – A 6-arm Radial Arm Maze Task. Rewards are found at the end of each arm. Due to the time limit, players/agents need to be efficient by not visiting the same arm more than once.

Bottom Right – A forced choice horizontal 'hook' task. Players/agents start on a platform and must observe that the only way to succeed is to choose the right side and use the pushable block to move the reward out of the red zone.



Results

- Children and AIs performed similarly on basic tasks involving simple navigation ($W(30, 52)=560$, $p>.005$) and reward preferences ($W(30, 52)=670$, $p>.005$).
- They performed **significantly differently on all other abilities**.



- There was no significant difference between the age groups, and each age group was significantly different to the sample of AIs.
- There was a significant interaction effect for ability and agent level ($F(9, 738)=45.765$, $p<0.00001$) , suggesting that the degree to which children and AIs differ is different for each ability.
- AIs noticeably struggled with navigation around **static obstacles** and **generalising knowledge** about objects in the environment when causally irrelevant properties such as colour and shape were altered.
- **Tool use tasks appeared difficult** for both children and AIs, although some children were capable of solving these tasks.
- 'ironbar' and 'Trrrrr' were the top performing agents, however both performed significantly differently to children ($T^2(40, 14) =129.1$, $p<0.0001$; $T^2(40, 14)= 258.49$, $p<0.0001$, respectively).
- 'ironbar' and 'Trrrrr' outperformed children on several tasks in which **visual input was periodically withdrawn**.
- Exploratory cluster analysis suggested that **'ironbar' was the only agent to cluster with children**, despite differing from 'Trrrrr' by less than 0.2% overall (although this effect disappeared after dimensionality reduction was applied).

Conclusions

- AI research still has significant progress to make before it achieves human-level abilities across a range of common-sense problems. Cognitive science can help!
- The Animal-AI Environment offers a **unique paradigm for extending comparative psychological methods to AI**, permitting a mutually beneficial dialogue between the two domains.
- This is a proof-of-concept study demonstrating what is possible.

References

Chollet, F. 2019. "On the Measure of Intelligence." *arXiv.*

Dong, Y., F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. 2018. "Boosting adversarial attacks with momentum." *Proc. IEEE conf. comp. vision patt. Recog.*9185-9193.

Geirhos, R., J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. "Shortcut Learning in Deep Neural Networks." *arXiv preprint arXiv:2004.07780.*

Pre-print: Voudouris et al. (2021). Direct Human-AI Comparison in the Animal-AI Environment. https://doi.org/10.31234/osf.io/me3xy

Acknowledgements: Thanks to Prof. N. Clayton and Dr. R. Moore for helpful comments.