Humans learning a complex task are picky and sticky

T. Quendera (tiagoquendera@neuro.fchampalimaud.org)
Systems Neuroscience Lab, Champalimaud Research
Lisbon, PT

Dongrui Deng (ddr11758@stu.xjtu.edu.cn)
Xi'an Jiaotong University
Shaanxi, CN

Mani Hamidi (Mani.hamidi@uni-tuebingen.de)
Human and Machine Cognition Lab, University of Tubingen
Tubingen, DE

Mattia Bergomi (info@veos.digital)
Veos Digital
Rome, IT

Zachary F. Mainen (zmainen@neuro.fchampalimaud.org)
Systems Neuroscience Lab, Champalimaud Research
Lisbon, PT

Gautam Agarwal (gagarwal@kecksci.claremont.edu)
Keck Science Department, Claremont Colleges
Claremont, USA

Abstract

While neural networks approach human levels of performance in many complex tasks, they require much more training than humans. This may be because only humans can infer and apply generalizable principles from prior experience (Lake, Ullman, Tenenbaum, & Gershman, 2017). However, the statistics that underlie the human learning process are poorly understood and hard to investigate in the large state spaces found in most complex tasks (van Opheusden & Ma, 2019). We thus designed a cognitive task whose potential solutions are few enough for subjects to densely sample policy space, but complex enough to compel intelligent search. We launched the game as a smartphone-based app (hexxed.io) to collect data from 10k human participants. We find that unlike reinforcement learning agents (Deep-Q Networks ; DQNs), humans (1) search a highly restricted subset of the policy space; (2) attempt even poor solutions many times before discarding them; (3) arrive at the optimal policy suddenly and unpredictably with a "leap of insight". Our data suggest a "top-down" learning process by which humans propose explanatory solutions which they replace only upon collecting sufficient evidence to the contrary, in contrast to the "bottom-up" learning of DQNs that associates states with rewarding actions.

Keywords: Problem Solving; Search; Intelligence; Skill learning; Decision making; Epiphany;

Introduction

Although the human advantage in games like Space Invaders and Go has shrunken with the rise of neural networks, what remains is the unparalleled speed with which we become skilled (Lake et al., 2017). Skill learning can be cast as a search through the space of strategies, but the nature of this search, whether it arises in humans through reward-driven gradient descent (Mnih et al., 2015), rational inference (Pouncy, Tsividis, & Gershman, 2021)(Tsividis et al., 2021), or some other process (Martin-Maroto & de Polavieja, 2018) remains unknown. One difficulty is that while games provide excellent benchmarks for intelligent search, they are not designed to reverse-engineer skill learning. We developed a game for this purpose and demonstrate how it can shed light on human learning processes.

Methods

Upon starting the game with minimal instructions ("Controls: tap and swipe"), subjects must learn to grow a target within a single lobe of a hexagon to its full size (but no further) and subsequently collect it. This task can be represented as an MDP in the form of a hexagonal prism with over 1,000 candidate action sequences, only a small fraction of which achieve criterion performance (i.e. at least 21 of 36 possible points) (Figure 1). Subjects must complete 6 trials (ie. one "level attempt") at criterion performance in order to progress. Our goal was to observe enough players reaching criterion to characterise search in terms of two probability distributions: the probability of selecting any particular action sequence (i.e., inductive bias) and the probability of selecting an action sequence in the next trial given the current selection (i.e., learning). To this end we developed, launched and marketed a videogame-based task (`http://hexxed.io`), which attracted 10k players of which roughly one third reached criterion. We also implemented our task using OpenAI's gym environment, allowing us to compare human learning to that of model-free reinforcement learning agents known as Deep-Q Networks (DQNs). Here we present a 3-layer convolutional network but results were similar for 3-layer fully-connected and a linear network. We use these DQNs as a "null model", a known point of comparison for highlighting unique aspects of human learning that any accurate model should reproduce.

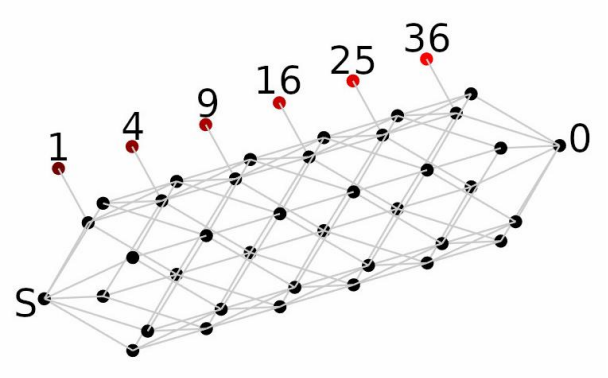

Figure 1: The possible search space of actions in level 1 can be mapped as deterministic a Markov Decision Process (MDP) and visualised as a hexagonal prism of states (dots) and actions (edges). 'S' indicates initial state and numbers indicate terminal states.

Results

Humans are picky

From their first move, humans appear to sample only a relatively small subset of the possible actions: 80% of subjects (n=1214) choose the lobe that contains the rewarded target (Figure 2A). We find that subjects traverse a "highway" bordering terminal states (i.e., interacting with the target-containing lobe), whereas the more agnostic DQNs more evenly sample states of the MDP (Figure 2B).

Humans are sticky

We next looked at how policies are modified by visualising the probability of using an action sequence conditioned on the previous trial's action sequence. Because the full transition matrix (determined by the number of unique action sequences) is enormous, here we group sequences by the 7 possible rewards they produce (Figure 3A). Whether grouped or un-grouped, humans' matrices are uniquely overrepresented along the diagonal because they stick to their

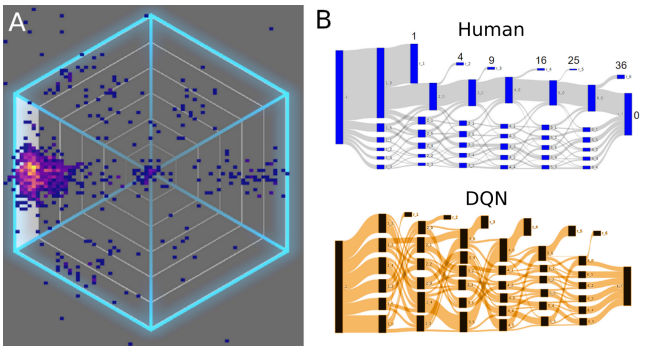

Figure 2: (A) A heatmap of humans' initial interaction with the screen overlaid on a screenshot of the task (n=1214). Despite the absence of instructions, humans tend to select the section of the screen that has available rewards (signified by the grey bar on the screen). (B) Flattened versions of the MDP where node/edge size indicates density of visits by humans (top; n = 3617) and DQNs (bottom; n = 200).

chosen action sequence across trials. When observed across level attempts, we find that humans show "leaps of insight". Subjects often transitioned suddenly after a long bout of no rewards to a highly (often optimally) rewarding policy. In contrast, we find that DQNs gradually improve their performance until they reach the minimum criterion. (Figure 3B)

Learning efficiency decreases with age

To examine whether the task can distinguish biological factors influencing learning, we included a survey of age, gender, and personality. The most striking differences were found across age, with the oldest group requiring roughly three times more attempts to pass the level than the youngest group (Figure 4).

Discussion

We designed a task in which subjects repeatedly attempt the same deterministic puzzle, allowing us to track precisely how individuals search a large solution space. Humans are "picky": they restrict their search to those policies that "make sense" (Dubey, Agrawal, Pathak, Griffiths, & Efros, 2018). Humans are "sticky": they persistently employ even unrewarding policies across trials, seemingly maladaptive for a deterministic task. Nonetheless, the "leaps of insight" we see (Figure 3B) are consistent with widely observed jumps in performance during complex skill learning (Gray & Lindstedt, 2017) that may reflect a process of accumulating evidence for or against a potential solution (Courville & Daw, 2007). Our results support the proposal that humans learn complex skills by some form of top-down "program induction" rather than a bottom-up gradient descent characteristic of neural networks (Lake et al., 2017).

Acknowledgments

This work was supported by grant #360/18 from Bial Foundation.

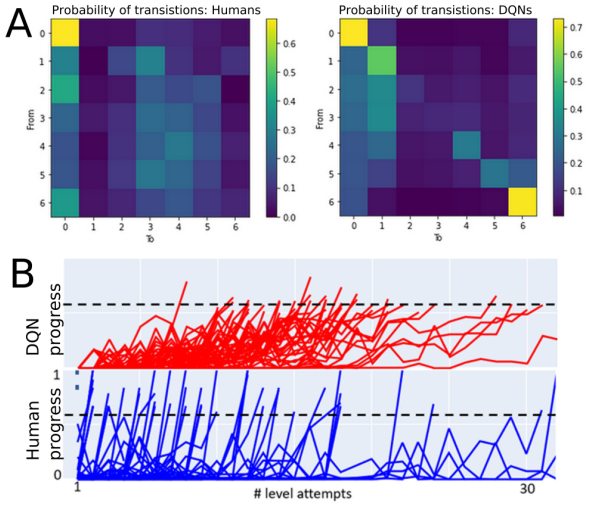

Figure 3: (A) Transition matrices for humans (left; n = 3617) and DQNs (right; n = 200) showing the probability of collecting a specific reward within a trial conditioned on the amount of reward collected on the previous trial. (B) Lower left: Learning curves for 50 randomly chosen DQNs (red) and humans (blue) as a function of level attempts. Progress is defined as a fraction of total possible points collected per attempt.

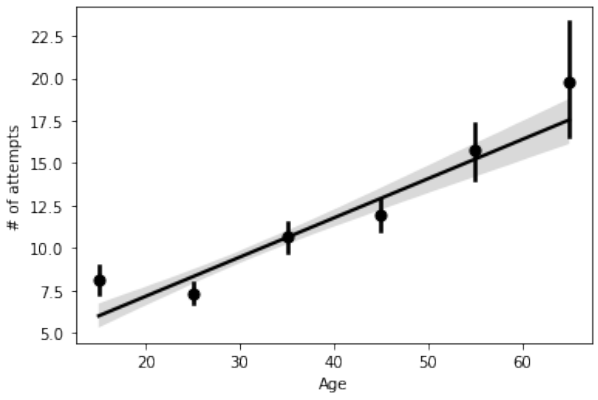

Figure 4: Figure 4: When grouping players based on their reported age (n=594), we find that young players require significantly fewer attempts on average to clear the first level. Dots and bars represent means and SEM; line is best fit linear model (r2=0.081, Wald Test p=1.509e-12).

References

Courville, A. C., & Daw, N. (2007). The rat as particle filter. *Advances in neural information processing systems*, *20*.

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018). Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*.

Gray, W. D., & Lindstedt, J. K. (2017). Plateaus, dips, and leaps: Where to look for inventions and discoveries during skilled performance. *Cognitive science*, *41*(7), 1838–1870.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Martin-Maroto, F., & de Polavieja, G. G. (2018). Algebraic machine learning. *arXiv preprint arXiv:1803.05252*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Pouncy, T., Tsividis, P., & Gershman, S. J. (2021). What is the model in model-based planning? *Cognitive Science*, *45*(1), e12928.

Tsividis, P. A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., ... Tenenbaum, J. B. (2021). Human-level reinforcement learning through theory-based modeling, exploration, and planning. *arXiv preprint arXiv:2107.12544*.

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133.