Dissecting dynamical components of complex decision-making using a computer game-based task

Gautam Agarwal[1], Mani Hamidi[2], Dongrui Deng[3], Tiago Quendera[4], Mattia Bergomi[5], Zachary Mainen[4]

[1]Keck Science Department at the Claremont Colleges; [2]Human and Machine Cognition Lab, University of Tubingen; [3]Xi'an Jiaotong University; [4]Systems Neuroscience Lab, Champalimaud Research; [5]Veos Digital

Introduction

In many everyday situations, we confront the "curse of dimensionality": we can pursue many possible paths of action, only some of which may be useful. How does the brain learn to make a proper decision in these situations? There are two competing explanations of this phenomenon (Fig. 1).

The first view (henceforth termed "bottom-up") is that the organism learns associations between its sensory inputs, motor outputs, and rewards, and adjusts its actions to maximize rewards. This process has been modeled using artificial neural networks, which are able to match or exceed human performance at complex tasks including Atari games and Go. However, these networks lack some crucial ingredients of human intelligence: humans can learn these tasks with less than 1/1000 of the training data and can maintain their ability when the goal or context is changed.

A second view (henceforth termed "top-down") is that organisms view the world as arising from a set of latent factors that follow certain rules (a "causal structure"). It has been proposed that humans learn these causal structures by sampling from a discrete space of policies, constrained by prior knowledge in the form of geometric, physical, and psychological intuition. While this view explains the rapid and discontinuous way in which organisms appear to learn, how the brain may implement it remains mysterious.

Since there are arguments for both "top-down" and "bottom-up" forms of learning in the brain (McClelland et al. 2010, Griffiths et al. 2010), our work seeks to compare the predictions of these models in a task that is both sufficiently controllable and sufficiently challenging to quantify complex skill learning. Based on previous work (Lake et al. 2017), we hypothesized that humans would tend toward "top-down" learning.

Aims

Aim 1 - Develop a system for quantifying complex skill learning

Deriving general principles of complex decision making requires a task that meets several competing design constraints (Allen et al. 2024). First, the "state space" of the task is sufficiently large, motivating subjects to develop efficient strategies to solve it. Second, the values of all states and actions should be pre-defined so that the value of a player's choice is unambiguous. Third, stimuli can be generated programmatically, allowing us to comprehensively sample subjects' changing strategy across stimuli throughout learning. Fourth, the task progresses in difficulty, so that it can be used to observe skill formation within individuals. Finally, the task generates enough data, within and across subjects, to sufficiently sample the distribution of strategies that subjects employ.

Aim 2 – Evaluate competing accounts of complex skill learning in humans

We will use the task to evaluate two major classes of models: the "bottom-up" view where an agent learns through distributed modifications of an associative network; and the "top-down" view where an agent intermittently proposes, evaluates, and updates causal theories of the world. Each view makes specific predictions: 1) the "bottom-up" view predicts that players gradually develop hierarchical representations that are shaped primarily by previously encountered states and rewards, while generalizing poorly for new puzzles. The distribution of action sequences is expected to reflect its associated reward (see Fig. 2B, C); 2) the "top-down" view predicts that players occasionally "leap" to certain strategies (new to them but common in the population), that exploit prior knowledge in a manner that generalizes well to new puzzles The distribution of action sequences is expected to reflect these priors (i.e. particular locations in Fig. 2B, C).

We will compare the predictions of these two models using online human data generated by Aim 1. In addition, we will compare human learning to that of an artificial neural network that implements "bottom up" learning (Mnih et al., 2015).

Aim 3 - Model the dynamics of plan formation preceding action selection

We will complement the online data collection of Aim 2 with pupillometry in the lab to model how subjects allocate cognitive resources in making a decision. We predict that trials with longer reaction times correspond to ones involving more planning. Specifically, we expect pupil size to grow larger for longer trials, and pupil saccades to increase linearly with trial length.

Methods

Task design

To satisfy the constraints outlined in Aim 1, we designed a parametric version of the traveling salesman problem, a classic "NP hard" problem (a general class of problems with no known efficient solution). In the task, subjects must plan a path around a hexagon to collect targets at appropriate locations and times. We implemented this task as a downloadable video game (available at hexxed.io) and publicized it via Portuguese and American media channels, enlisting ~10,000 players of all ages and genders worldwide to this date.

The game progresses through 6 levels of increasing difficulty. Within a level, subjects are presented with different stimuli, consisting of arrangements of 1 to 6 objects (the number and difficulty increasing with level). The subject maximizes their score by collecting the objects when they are as large as possible. At any moment, the subject can initiate 1 of 12 actions – they can either move to or attempt object collection from any of the 6 positions. Moving n spaces causes all objects to grow towards the edge by n steps, increasing their value. However, if an object reaches the edge before the subject reaches it, it disappears, and its value is lost.

By interviewing our human subjects during the testing phase, we learned that they learned to play this game in two related but distinct phases:

1) Learning the rules and objectives of the game – Since subjects were not given any verbal instructions, they must search for sufficiently rewarding action sequences. In the first level, both humans and AIs need to learn to collect a single object at its full size (Fig. 2A). This requires choosing from over 4,000 viable action sequences (Fig. 2B). Interviews indicated that rule learning process largely (but not completely) takes place in level 1.
2) Knowing the rules and objectives, identifying the best action – once subjects understand the rules, they must make an inference regarding how to apply these rules to maximize reward. Interviews indicated that this process starts in level two and onwards, when subjects must determine an optimal path to collect two or more objects (more directly comparable to the combinatoric optimization of the traveling salesman problem). Figure 2C shows several board configurations (henceforth termed "puzzles") of increasing difficulty. For each puzzle, we can visualize all the distinct ways of collecting objects as a compact "value landscape". By changing the number and arrangement of targets (n = 1 – 6), we can densely sample the full stimulus space (containing 164 unique puzzles), allowing us observe players' actions as they transition to expert-level performance.

We have limited the scope of our current analysis to the first of these two phases, because formally it corresponds to the problem with the smallest search space, and because understanding this phase should help constrain our models of the second phase.

Data collection

Every time a player installs the game, they are presented a screen indicating that the data they generate in the app is going to be collected by a secure server. If they choose to accept, a unique user ID is generated and stored locally on their device. When they start a level, the details of each interaction with the screen (screen spatial and temporal coordinates and the game action evoked), as well as the puzzle's information, are recorded. Upon completion of a level (whether the user quits, loses, or passes the level), the information is sent to a central, secure mySQL database. If the information

cannot be sent (e.g. if the user is offline), then it is saved locally within a data file and an attempt is made to send it the next time the user plays a level. In addition to the task-related data, subjects have access to a leaderboard which indicates their global ranking among all players; each leaderboard access event is also logged by the database. Finally, to correlate behavioral data with demographic information, users have the option to complete a short survey which includes questions about their age, gender, and big 5 personality traits.

Data analysis

The models of learning differ in their predictions of how players sample available action sequences. We quantify how restricted the attempted policies are, as well as how these policies change over consecutive trials, to compare model predictions. In addition, we evaluate whether demographic variables predict inter-subject differences.

Comparison to artificial intelligence (AI)-based agents

We implemented the task using OpenAI's Gym interface, allowing us to observe learning in artificial agents. Specifically, we represented the hexagonal game board as a 6x12 binary matrix, each entry indicating the presence of a target at the corresponding coordinate. The actions, state transitions, and criteria for level passing replicated those for the human version of the game. Our AI was instantiated using the Deep Q-learning Network (DQN), a "bottom-up" agent that learns using off-policy reinforcement learning (RL). The network takes as input the binary matrix indicating game state and outputs which action to choose. We compared 3 variants – a 1-layer linear network, a 3-layer ReLu network, and a 4-layer convolutional + ReLu network. All networks exhibited qualitatively similar behavior, but the convolutional network was consistently faster in its learning, so we used it for our comparisons of humans and AI.

Pupillometry

10 participants were recruited internally from Champalimaud Research and instructed to play the game using a touch-sensitive monitor, for up to 1 hour. At the same time, their pupil was measured at ~120 Hz using a pupil-labs eye tracking system. Bonsai software was used to register timestamps of eye tracking frames as well as touch-screen presses, which were used to align eye tracking time with game time. The pupil size and X-Y position were extracted from the recordings. These measurements were logged over a 15-second window (-5 to +10s) around the puzzle onset, along with the time of first action, which indicated the subject's reaction time for that puzzle. Data was combined across subjects to produce average plots (Fig. 7).

Results

Humans are selective

We first compared the distribution of paths taken by humans and "bottom-up" AIs. We find that even upon their first interaction with the game, humans try one of a small number of possible actions (Fig. 3A). Specifically, ~80% of subjects (n=1214) interact with the only lobe that has an object within it. Using a flattened version of the MDP from Fig. 2B, we find that the action sequences that people attempt are also very selective (see thick edge in Fig 3B). This pattern is not seen in the "bottom-up" AIs, which sample states of the MDP uniformly, as expected from randomly initialized agents. The highly restricted paths that people sample from the outset suggests that they can apply a "top-down" prior to the novel, unfamiliar environment of the game (ie., that occupying positions containing objects is more likely to be rewarding than occupying empty positions).

Humans are persistent

We then observed how subjects changed their policies across subsequent trials. To do so, we calculated the probability of observing a strategy conditioned on the strategy used on the previous trial (a "transition matrix"). We find that the transition matrix for humans shows a strong diagonal component, indicating that humans repeatedly sample the same policy for multiple trials before changing it. This is true regardless of whether the policy was rewarding (ie., is seen both in the top-left and bottom-right part of the diagonal) and is not observed in the AIs, which have a much more diffuse pattern of transitions (Fig. 4).

Humans show leaps of insight

We observed the probability that subjects would learn to solve level 1 as a function of the number of level attempts. We find that humans are on average faster than AI, and that the distribution is more heavy-tailed for humans than for AIs (Fig. 5). Observing individual learning curves, we find that AIs gradually improve their performance to reach the minimum criterion for passing the level; in contrast, humans show "leaps of insight", often transitioning suddenly from a completely unrewarding policy to the optimal one.

Learning efficiency decreases with age

To assess the influence of biological factors on human learning, we compared how quickly human players in different age groups were able to pass the first level. We found a strong age dependence, with the oldest group requiring roughly 3x more attempts to pass the level than the youngest group (Fig. 6).

Pupils show planning prior to action

We measured the pupils of 10 subjects while they completed the task in the lab. Because greater pupil size can indicate higher cognitive load, we hypothesized that pupil size would increase following puzzle onset. Surprisingly we found the opposite. Further investigation revealed that this decrease followed an increase in screen brightness (Fig. 7 left). Thus, pupil size could not be used as a reliable indicator of cognitive load for our task. Instead, we used eye velocity as a measure of cognitive processing, finding that it increased reliably following puzzle onset. Separating trials according to reaction time (i.e., time between puzzle onset and first action) revealed a biphasic speed profile: an initial eye movement that was relatively unaffected by reaction time, and more dispersed eye movements that were present in trials with longer reaction times (Fig. 7 center). Due to this correlation, reaction time could be used to predict the total eye movement on a given trial (Fig. 7 right).

Discussion, Conclusions and Recommendations

We report the successful implementation of an experimental system to observe the dynamics of complex skill learning across a large global human population, allowing us to compare the contributions of "bottom-up" and "top-down" forms of learning. This entailed 1) developing a complex cognitive task whose stimulus space could be parametrically sampled; 2) embedding the task within a mobile phone-based app that could collect user data remotely; 3) creating a secure online database to record player actions longitudinally over the course of skill development; 4) publicizing the task to reach an audience of ~10k players; 5) implementing a virtual environment for comparing learning in humans and AIs; 6) Comparing patterns of human and AI behavior using metrics designed to probe the predictions of the two models. This pipeline has led to the creation of a uniquely high-powered data set for evaluating models of human learning in complex environments.

Our results confirm the predictions of the "top-down" model of learning: 1) humans converge on the same, highly restricted subset of viable policies when first attempting the game. Because the interface and rules of the game do not bear any obvious relationship to their prior experience, this result suggests that subjects share a high-level theory that they use to select their actions within the novel environment of the game. Note that this pattern alone does not rule out the possibility that such a prior can be applied in new environments by a "bottom-up" learning agent, though how this would be implemented is less clear; 2) humans persist with the same, often unrewarding, policy across multiple trials. This pattern is consistent with the notion that humans are sampling from a discrete space of theories (Fig 1), collecting evidence of their current theory being a viable explanation for their goal before choosing to adopt or discard it. This contrasts with "bottom-up" learning, which searches a continuous space via gradient descent and thus changes its policy gradually (Fig. 1). 3) humans display leaps of insight, often suddenly transitioning during learning from completely unrewarding to optimal solutions, once again inconsistent with the graded change of gradient descent and in alignment with the sudden changes implied by the formation of a new theory (Figs 1, 5). We demonstrate the utility of our approach as a sensitive readout of demographic influences on complex skill learning, finding that human learning efficiency gradually and significantly decreases over the span of 6 decades. Future analysis will determine whether these differences result from specific changes in the indicators of "top-down" learning described above. Finally, we show a correlation between total

pupil movement and reaction time, supporting the use of reaction time as a noninvasive readout of the degree of planning used by a subject prior to a choice.

Outlook: We believe that the presented task offers a unique combination of experimental control and algorithmic complexity, allowing us to quantify how subjects apply efficient heuristics to discover solutions. Furthermore, this grant has allowed us to collect a data set observing a world-wide pool of individuals as they transition from novice to expert, skilled behavior. We have just begun to scratch the surface of the data collected. Currently, the first author is supervising 6 undergraduates to analyze and model different aspects of the data collected, and 2 psychology graduate students to collect data in a controlled laboratory setting. While "bottom-up" neural networks have become commonplace, with many groups employing them for optimization problems and developing tools for dissecting their behavior, there is less consensus (albeit a long historical tradition (Lake et al. 2017)) for implementing and studying "top-down" learning. In observing how our subjects apply and modify their policies as they confront new, more complex puzzles, we intend in time to gain a deeper appreciation of creative and intelligent behavior in humans.

References

Allen, K., Brändle, F., Botvinick, M. .. Shulz E. (2024). Using games to understand the mind. *Nat Hum Behav* **8**, 1035–1043.

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, *29*, 17-23.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

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.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348-356.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015, 02). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

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*.

Ullman, T. D., Goodman, ND., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. Cognitive Development, 27(4), 455-480.

Figures

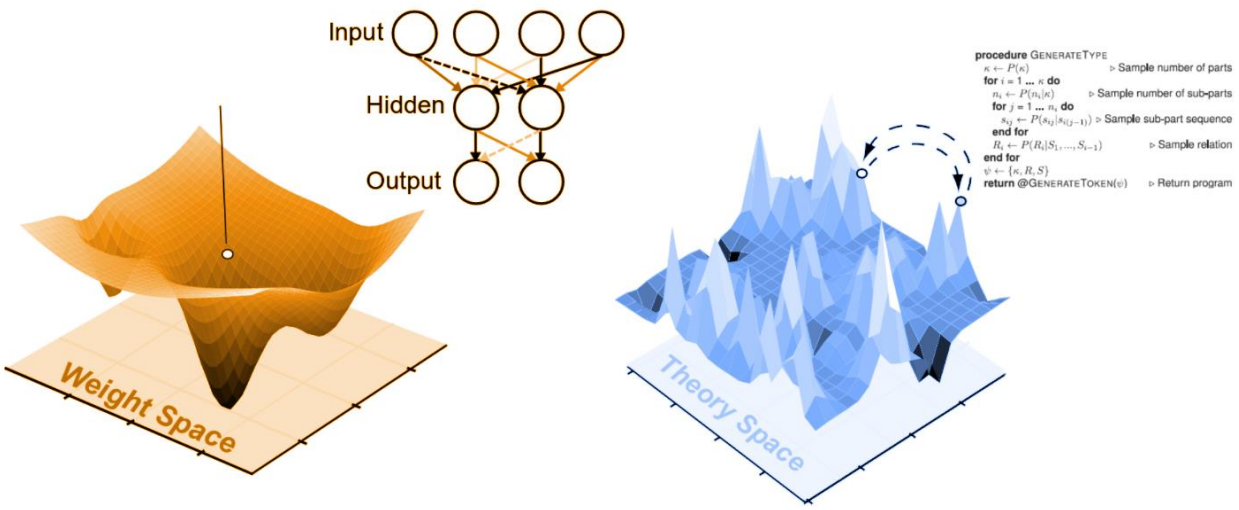

Figure 1. Two models of learning (adapted from Ullman et al. 2012). (*Left*, orange) In "bottom-up" learning, an agent uses gradient descent to adjust its weights in a manner that associates stimuli with rewarding actions. (*Right*, blue) In "top-down" learning, an agent organizes symbols into a theory that offers a candidate explanation of its environment; when the components of this theory are modified, the agent can 'leap' across the space of theories.

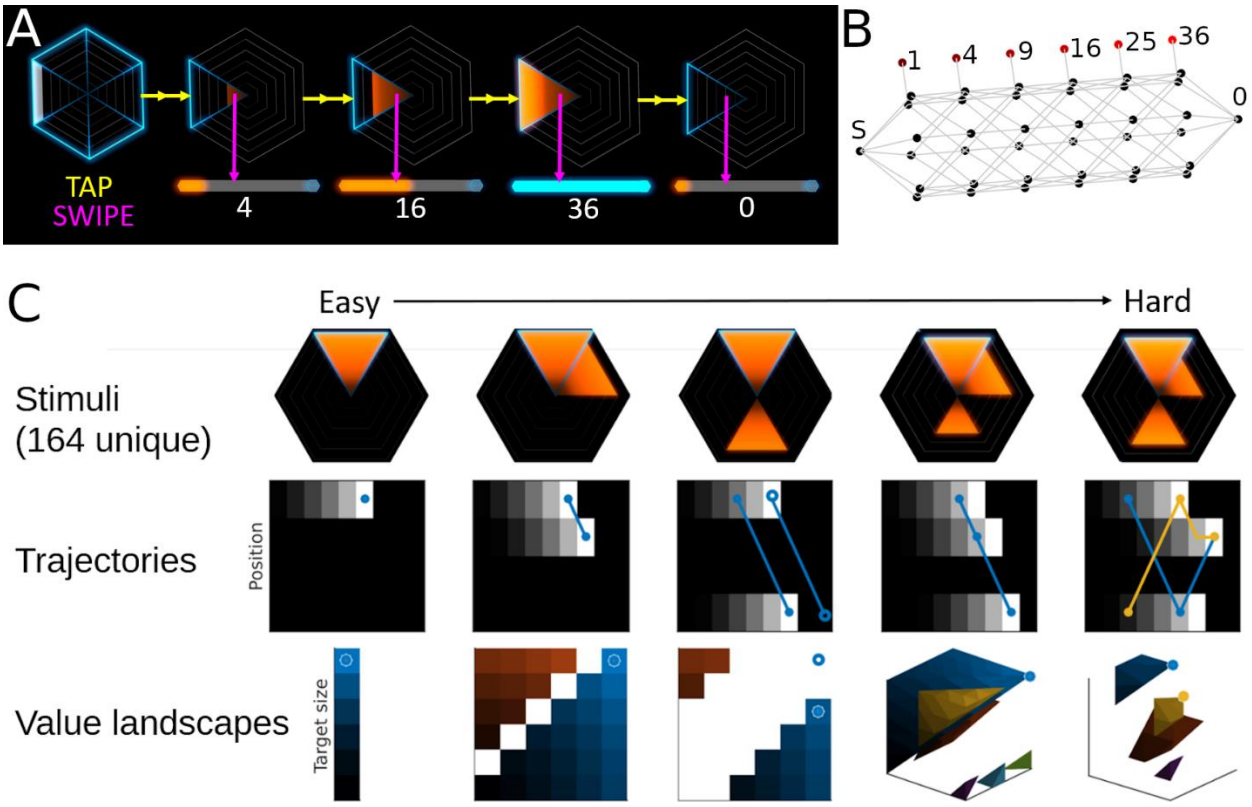

Figure 2: Task Design. A. Schematic of first level. Subject can either tap or swipe on the screen. Tapping causes the collector (blue outline) to move among the 6 lobes of the hexagon. The orange target grows by 1 unit each time the collector moves to an adjacent lobe of the hexagon. If the collector is on top of a target, the subject can swipe it to collect points (indicated by the number below the screen). However, the target disappears if it grows beyond the boundary of the hexagon. **B.** A Markov decision process (MDP) representing all possible states as nodes and actions as edges (some edges hidden for clarity). 'S' indicates start state, and numbers indicate rewards at all terminal states. The MDP resembles a hexagonal prism with terminal states aligning along one edge. **C.** As the number of targets grows, the state and solution

spaces grow exponentially. *Top row* shows example stimuli from levels 1 to 3. *Middle row* shows example trajectories a subject can take for a given stimulus (y axis: unwrapped position within hexagon; x axis: # of moves made. Gray ramps indicate the increasing value of objects, white being most valuable. Lines represent moves and dots represent collection attempts by agent. *Bottom row* shows a compressed value landscape for each puzzle. Each position within the landscape corresponds to a distinct solution. The position within the landscape corresponds to the size (between 1 and 6) at which each target is collected. An n-target puzzle's landscape is thus an n-dimensional hypercube. Color represents order in which targets are collected. Blue regions correspond to paths in which targets are collected from largest to smallest. The brightness of each position indicates the reward associated with it. White regions correspond to target size combinations that are impossible to achieve. Dots mark the locations of the corresponding trajectories depicted in the middle row.

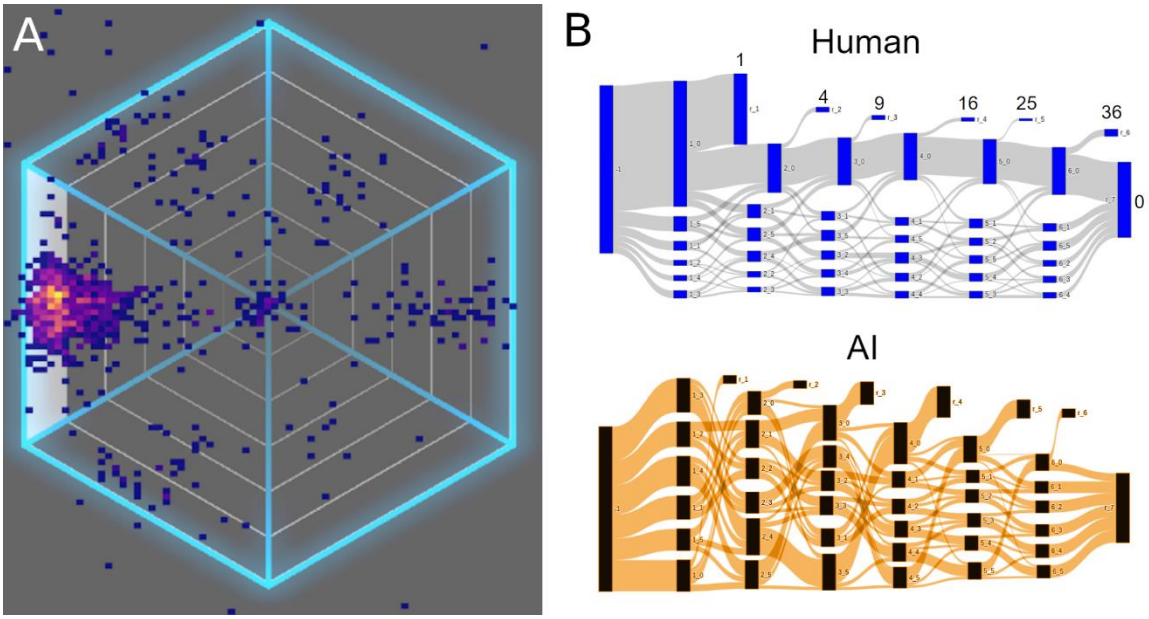

Figure 3: Humans are selective. A. Histogram of location of first tap across human subjects (n = 1214). Taps are highly concentrated in region where target will appear**. B.** Occupancy of paths within MDP (unrolled compared to the 3-d prism shown in Fig. 2B) pooled across all subjects and trial attempts. Subjects sample a subset of states and transitions within the MDP, indicated by the disproportionately thick path (signifying states in which the collector overlaps with the target to be collected). In contrast AI (bottom) sample paths more uniformly. Numbers indicate rewards at each terminal state, as in Fig 2B.

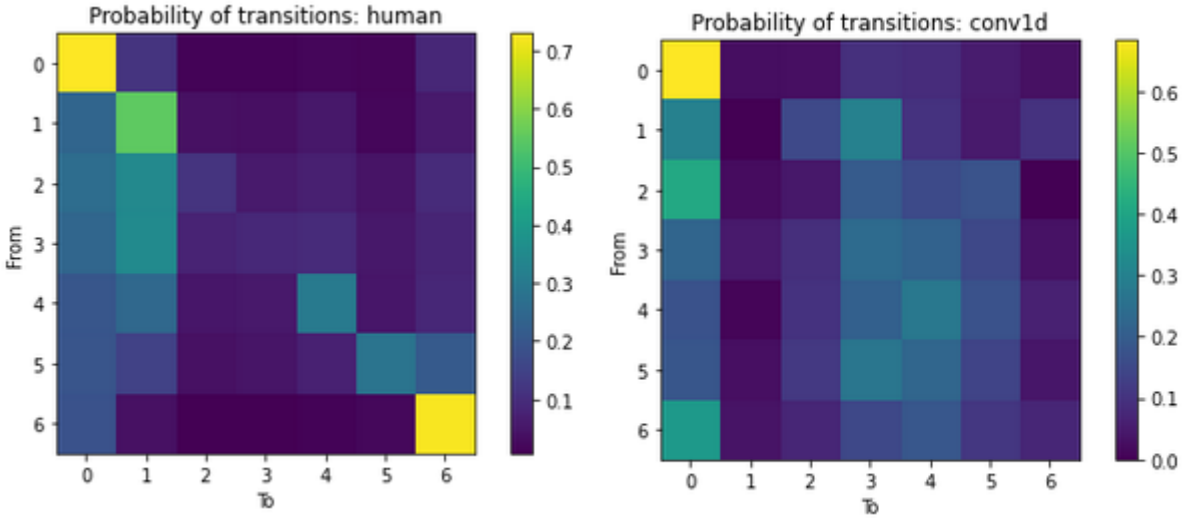

Figure 4. Humans are persistent. Transition matrices for humans (*left*) and AIs (*right*) showing the probability of collecting a specific reward within a trial conditioned on the amount of reward collected on the previous trial. Only humans show a strong overrepresentation of the diagonal, indicating that they are likely to keep trying the same policy across consecutive trials. While in principle the diagonal could represent transitions between different policies that offer the same reward, in practice we find that this overrepresentation is the result of the subject attempting the same policy multiple times.

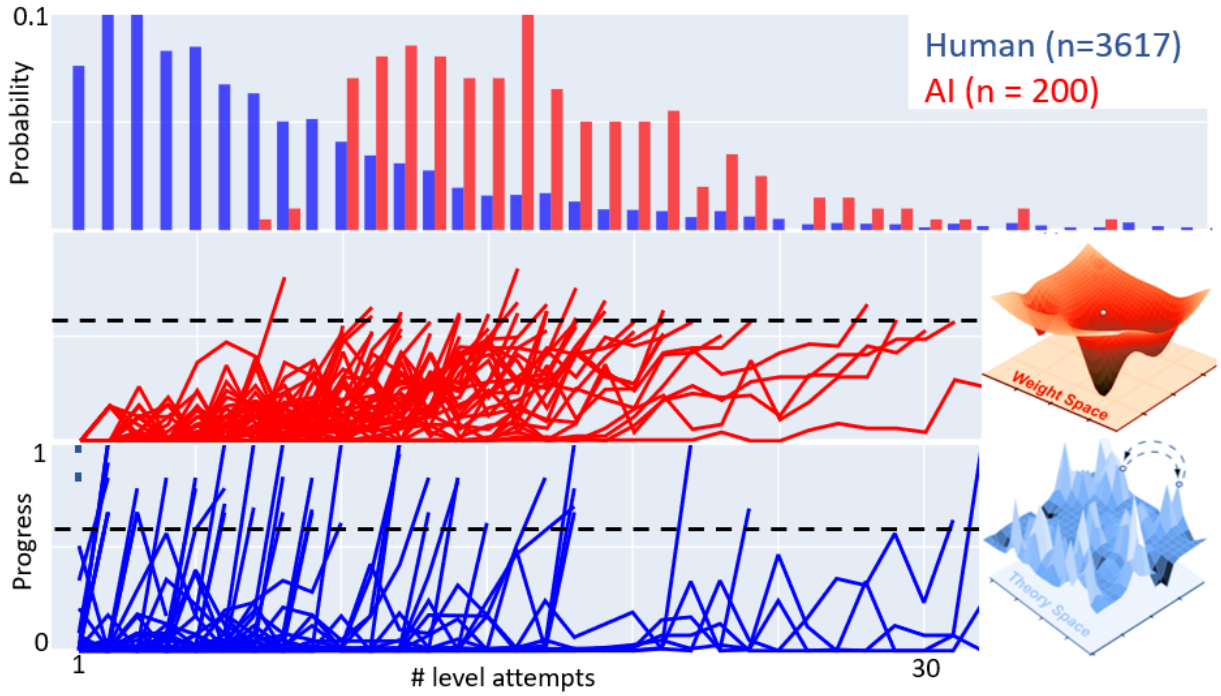

Figure 5. Humans show leaps of insight. *Top:* Probability of passing the first level as a function of number of level attempts for humans (blue) and AI. Human subjects who quit playing before passing the first level are excluded. *Lower left*: Learning curves for 50 randomly chosen AI (red) and humans (blue) as a function of level attempts. Progress is calculated as fraction of total possible points collected per level attempt. Note that unlike AIs, humans often collect very little reward before suddenly transitioning to a state of high performance. *Lower right*: Schematics of bottom-up (red) and top-down (blue) learning (from Ullman et al. 2012) are consistent with the learning curves we observe for AIs and humans respectively.

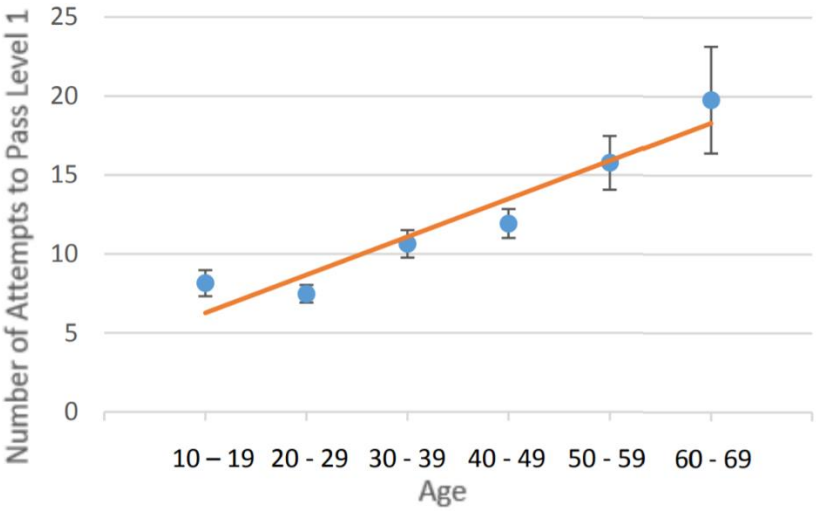

Figure 6. Learning efficiency decreases with age. When grouping players based on their age as reported in survey data (n=600), we find that young players require significantly fewer attempts on average to pass the first level. Blue dots and bars represent means and SEM; red line is best fit linear model.

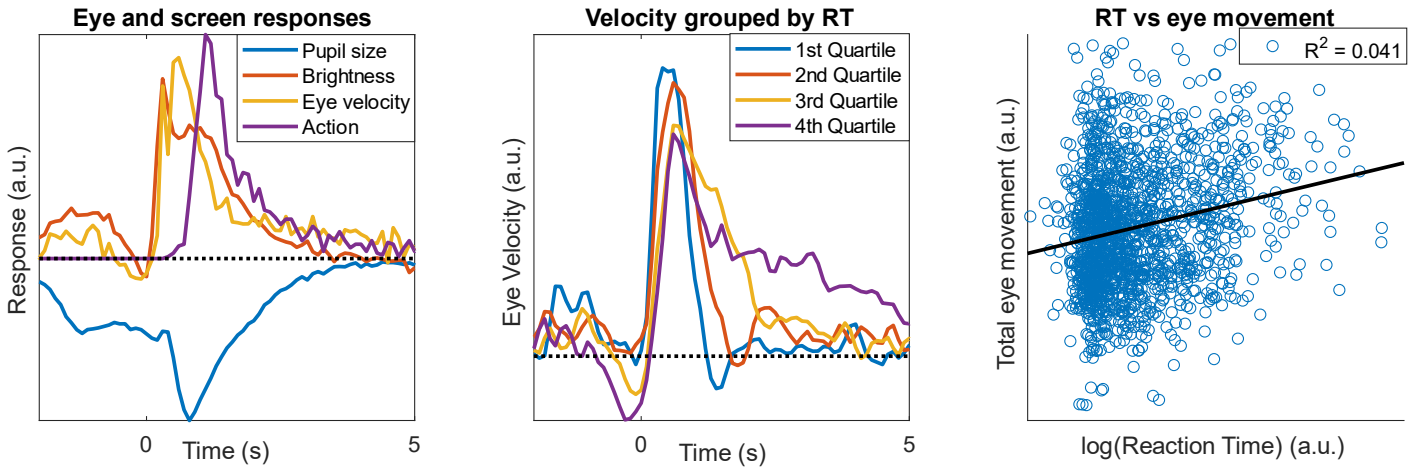

Figure 7. Eye movement increases with reaction time. *Left:* Following puzzle onset (t = 0s) and eye velocity (gold) increase while pupil size (blue) decreases. Changes in eye velocity precede subjects' first action (purple) (n = 10 subjects, 1581 trials). *Center:* separating eye velocity averaged into four groups of trials based on reaction times reveals two phases – an initial peak following puzzle onset (t=0) in which the pupil quickly reorients, followed by more temporally dispersed eye movements. The first phase is more consistent across reaction time groups, while the second phase increases with reaction time. *Right:* trials with longer reaction times contain greater total eye movement, suggesting longer delays signify more planning.