Reinforcement Learning and Insight in the Artificial Pigeon

Thomas R. Colin (thomas.colin@plymouth.ac.uk)
School of Mathematics and Computing, University of Plymouth
Plymouth, U.K.

Tony Belpaeme (tony.belpaeme@plymouth.ac.uk)
School of Mathematics and Computing, University of Plymouth
Plymouth, U.K.

Abstract

The phenomenon of insight (also called "Aha!" or "Eureka!" moments) is considered a core component of creative cognition. It is also a puzzle and a challenge for statistics-based approaches to behavior such as associative learning and reinforcement learning. We simulate a classic experiment on insight in pigeons using deep Reinforcement Learning. We show that prior experience may produce large and rapid performance improvements reminiscent of insights, and we suggest theoretical connections between concepts from machine learning (such as the value function or overfitting) and concepts from psychology (such as feelings-of-warmth and the einstellung effect). However, the simulated pigeons were slower than the real pigeons at solving the test problem, requiring a greater amount of trial and error: their "insightful" behavior was sudden by comparison with learning from scratch, but slow by comparison with real pigeons. This leaves open the question of whether incremental improvements to reinforcement learning algorithms will be sufficient to produce insightful behavior.

Keywords: reinforcement learning; insight; creativity

Introduction

Insight moments are one of the most spectacular manifestations of human creativity. Revolutionary insights are paradigmatic examples of creativity, whether historically suspicious (Aristotles' "Eureka!", Newton's apple), or better documented such as those described by the mathematician Poincaré (1909) or the chemist Kekulé (Rothenberg, 1995). In this article, however, we focus on the insights which occur in everyday human and animal problem-solving.

Over a century of research in psychology underlies our knowledge of insightful problem-solving. In contrast, to our knowledge there has been relatively little work considering insight from an artificial intelligence perspective, especially since the momentous advent of deep learning techniques in AI. We seek to remedy this omission. The objective is not to build a precise model of biological neural processes, but to uncover analogies between the two domains of deep Reinforcement Learning (RL) and biological insight. We do this by simulating a classic experiment on insight (Epstein, Kirshnit, Lanza, & Rubins, 1984), dealing with insight in the pigeon.

We will first discuss established results from insight research in psychology on humans and animals, and the difficulties associated with modeling insight problems from a machine learning perspective. We will then describe the original experiment and its simulation, and the results obtained using a simple deep RL approach (a deep actor-critic). Finally, we discuss the analogies between insight and various sub-disciplines within reinforcement learning, suggesting directions for future research.

Background: insight

Psychological research on insight begins with studies on chimpanzees by Köhler (1921). These studies sought to demonstrate that animals, far from being Cartesian automatas as suggested in the work of Thorndike (1898), are capable of human-like intelligence. One of Köhler's experiments involved attaching a banana to the ceiling of the chimpanzee enclosure, and placing a box within the enclosure. The chimpanzees had to carry the box underneath the banana and climb onto it in order to reach the fruit. When solving the problem, the chimpanzees displayed behavior that more closely resembled Aristotle's "Eureka!" than the trial-and-error learning of cats locked in puzzle-boxes by Thorndike (1898). In Köhler's "gestalt" perspective, it was understood that chimpanzees had to interpret the situation from scratch in order to discover the "roundabout" way of reaching for the objective.

Later work by Birch (1945) showed that chimpanzee insight was not achieved from scratch, but was instead made possible by relevant prior experiences. Epstein et al. (1984) showed that with adequate training, "even" pigeons could display the kind of insight observed in chimpanzees. Epstein's findings are robust: several variations of this experiment were performed by Epstein and colleagues, and the original was recently replicated by Cook and Fowler (2014). For Epstein, who was a student of Skinner, this made the argument that seemingly complex mental processes could be explained from behaviorist principles.

There has been continued interest in insight since the cognitive turn in psychology. This body of work has established several key behavioral, cognitive, and metacognitive characteristics of insight:

1. The insight sequence: search – (impasse) – restructuring – verification (Ohlsson, 2011; Weisberg, 2015).
2. Insights are sudden and surprising to the problem-solver, as evidenced by "feeling-of-warmth" ratings measuring subjective closeness to the solution (Metcalfe & Wiebe, 1987).
3. The "restructuring" which accompany insight involves

changes in problem representation (Knoblich, Ohlsson, & Raney, 2001), in the heuristics used (Kaplan & Simon, 1990), and in the constraints on operators (MacGregor, Ormerod, & Chronicle, 2001).

4. Insight depends on previous experience (Wiley, 1998) and is facilitated by sleep (Wagner, Gais, Haider, Verleger, & Born, 2004).

Recent research on insight has used imaging techniques such as fMRI[1]. Much of this work has focused on associative cortices (notably middle and temporal gyri) and on hemispheric differences (Kounios & Beeman, 2015); however the involvement of structures associated with executive control is a robust finding (prefrontal cortex, especially anterior cingulate cortex), and recent ultra high-field work (Tik et al., 2018) suggests the involvement of deeper brain structures during insight, including those underlying biological reinforcement learning (subcortical dopaminergic structures including the striatum, thalamus, nucleus accumbens and ventral tegmental area).

Summarizing: a rich body of research has investigated insight according to different psychological research paradigms, establishing the key characteristics of insight enumerated above. However, the precise nature of the cognitive mechanisms that enable insight remains unclear.

This is not to say that there have not been models, or theories, of the cognitive basis of insight (computational, mathematical, or otherwise); those of Hélie and Sun (2010), Friston et al. (2017), Schilling (2005), and Stephen, Boncoddo, Magnuson, and Dixon (2009) are among the most influential. A review of and comparison with these variegated models is beyond the scope of this paper, if only due to their great diversity, which ranges from bayesian inference (Friston et al., 2017) to dynamical systems (Stephen et al., 2009) and graph theory (Schilling, 2005). We note in passing that the model presented later in this article may be compatible with several of these other models: for instance phase transitions such as those described by Stephen et al. (2009) are conjectured to occur in neural networks.

None of the four models mentioned above aim to give rise to artificial agents capable of solving problems through insight[2]. In contrast, we seek to produce a model of insight problem-solving which, when implemented, not only predicts the behavior of a biological insightful problem-solver, but also solves the problem.

AI: which insight problems to model?

Most of the contemporary insight literature focuses on humans, using a wide array of experimental designs (for instance, the nine-dots problem (MacGregor et al., 2001), the mutilated checkerboard problem (Kaplan & Simon, 1990), or the Compound Remote Associates (Bowden & Jung-Beeman, 2003)). Despite their apparent variety, virtually all insight studies involving humans make use of verbal instructions which define the objective for the problem-solver in their language.

Consider the nine-dot problem: the instructions specify the number of segments, with constraints over their properties (four segments, drawn in a sequence "without lifting the pen"; every dot should end up on one of the segments). Language thus allows for a description of the desired "goal-state" which is abstract enough to specify the solution without giving it away. Simulating such a problem using AI would require either very task-specific algorithms (which seems to defeat the point of replicating human insight), or the algorithmic mastery of language as a prerequisite for understanding instructions.

A "roundabout" solution is to focus instead on insight experiments which feature animals solving problems that are not specified by instructions, but instead by some intrinsic need, typically for food, and by the situation in which the experimenter puts the animal[3]. This is the approach taken in this article.

Insightful (real) pigeons

The experiment by Epstein et al. (1984) is a reproduction of Köhler's banana-and-box experiment, adapted for pigeons. Chimpanzees would naturally want to acquire a banana; but pigeons might not be interested in that fruit. Therefore Epstein et al. first reinforced pecking a facsimile banana (hereafter just "the banana") by providing a suitable food reward upon pecks. In the "test" situation, the banana is suspended from the ceiling of the room, such that pigeons cannot reach it by stretching towards it (they do not attempt to fly towards it (Cook & Fowler, 2014)). However, a small cardboard cube ("the box") has been placed in the pigeon's Skinner box. The problem is solved when the animal pushes/pecks the box underneath the banana and, standing on the box, reaches for/pecks at the banana; see figure 1.

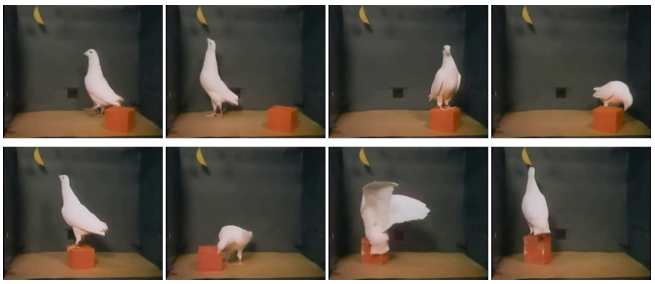

Figure 1: Left-to-right, then top-to-bottom: a pigeon solves the banana-and-box test (snapshots from https://youtu.be/mDntbGRPeEU, with permission from Dr. Epstein).

[1] See Sprugnoli et al. (2017) for a review of brain imaging studies.

[2] A notable exception is the model of MacLellan (2011), who investigates insight as a change of heuristics in a search process, and tests this on the nine-dot problem.

[3] See Shettleworth (2012) for a judicious review of insight research on animals.

Prior to this apparent display of ingenuity, the behavior of Epstein's pigeons was carefully *shaped*. Shaping is a technique used in animal training (with closely related applications in certain behavioural therapies for humans), consisting of reinforcing successive approximations of a desired behavior. Two skills ("behavioural repertoires") are taught to the pigeons by reinforcing the corresponding behaviors:

- In the absence of the banana: push a box to a green spot.
- With the box nailed underneath the banana, and in the absence of a spot: jump on the box and peck the banana.

Teaching pigeons to push a box towards an objective is considerably more difficult than getting them to hop onto the pre-placed box. To achieve this, Epstein et al. proceeded gradually, the shaping sequence including teaching the pigeons to move the box, then progressively placing the box at an increased distance from the spot. Additionally, the pigeons were sometimes put in the presence of the box and in the absence of both banana and spot, in order to extinguish aimless pushing behavior (which eventually would result, via a random walk within the Skinner box, in reaching the correct position and thereby triggering the food reward).

It is of special importance that the two behaviors are not exactly applicable to the final test: the pigeons are trained to push the box towards a green spot, but in the test situation they must spontaneously generalize this behavior to a slightly different problem: pushing towards the yellow banana. It is by combining two behaviors, and generalizing one behavior to a novel situation, that the pigeons solve the test task.

Epstein's pigeons proved remarkably adept in the test - all of them succeeding in minutes, save for one, and presenting behavior that seemed insightful: after a period of hesitation and some trial and error, the pigeons began acting in a seemingly directed, intentional manner, moving the box towards the banana and jumping on top of it. The lone laggard failed in a manner reminiscent of AI failures: during the test, a projector had been used to illuminate the (filmed) performance. When the additional lighting was turned off, the pigeon succeeded quickly.

Simulation

Admittedly, the displays of insight by Epstein's pigeons are less impressive than those of Köhler's chimpanzees: they received substantial training in the form of shaping. However, just as pigeons could not solve the test without having first acquired relevant skills, so chimpanzees were not able to solve insight problems without having first engaged in spontaneous play with the relevant objects (Birch, 1945). This suggests that similar cognitive mechanisms may be at play, and that it may be wise to begin by modeling the version of the task completed by pigeons.

In addition to requiring no instructions or verbal skill, the task used by Epstein et al. (1984) allows for a simulation which preserves much of what makes the task difficult: the pigeons had to combine pre-existing skills (pushing the box, and jumping on top of it to peck at the banana) while also generalizing to a new stimulus (pushing is shaped using a green dot, but in the test situation the pigeons must aim instead for a banana).

Thus, in simulating this task, we seek to preserve the difficulty inasmuch as it is relevant to problem solving, as opposed to the complete difficulty of the task including subjective perception and full physical coordination.

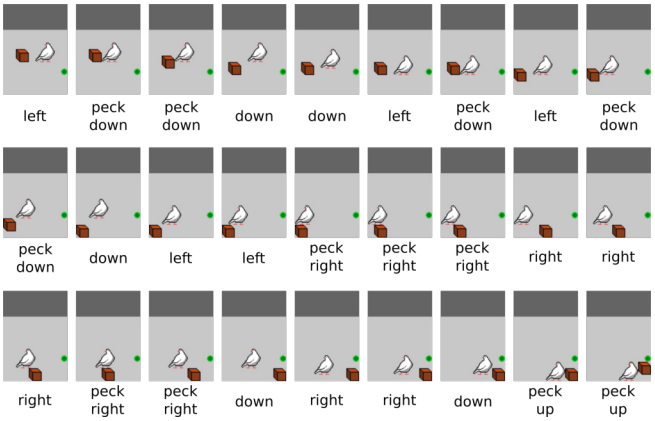

Figure 2: Successive frames of an artificial pigeon solving the "push box to spot" shaping task. The pigeon succeeded despite a sub-optimal policy (first pushing the box in the wrong direction, then pecking it out of corners). Also note the stochasticity of the "peck" actions: pecking actions have up to 4 different outcomes.

Task and shaping model

We model the task as an RGB image, such that the complete situation is perceived at each time-step. The pigeon, box, banana and spot consist of squares identifiable by size and color. For visualization, an interpretable representation is also provided (see figure 2). The dimensions of the various elements, and the dynamics of the actions are chosen to match those observed in the experiment. In particular, the size of the various elements (Skinner box, pigeon, box, banana, spot), the effects of the actions (walking, directional pecking, and jumping) and the consequences of interactions (box movement) closely match those of the initial experiment.

Specifically, the pigeon has 9 actions: walking in either cardinal direction, pecking towards either cardinal direction, and jumping on/off the box. Walking is deterministic and moves the pigeon by 1 square in the corresponding cardinal direction unless an obstacle is present. Pecking the box will result in its stochastic displacement in the general direction opposite to that from which it was pecked: assuming there are no obstacles and the box is not fixed in place, the box moves with equal probability (0.25) by 1 or 2 squares forward, or by 1 square forward and 1 in either perpendicular direction. With respect to direction, the pigeon can push the box south if the northern edge of the pigeon is at least as far north as the northern edge of the box, and if the pigeon is adjacent to the box;

likewise (mutadis mutandis) for the other directions (refer to figure 2 for some examples of stochastic box movement and pigeon positioning). The white pigeon is 3×3, the orange box 2×2, the green spot and yellow banana are 1 square each, and the background environment 10×10. Assuming squares approximately 4cm across, this roughly matches the size of the real objects (10x10cm box, 7x2cm facsimile banana, 4x4cm spot, approx. 25x8cm pigeons), Skinner box (45x45cm for the square box), and the effects of recognizable discrete actions in the original. In Skinner boxes, pigeons are rewarded by receiving food through a little window; in this simulation, a reward of 10 is provided instantaneously upon success.

The artificial pigeons undergo shaping similar to that used by Epstein et al.: artificial pigeons perform the **push-to-spot**, **jump-and-peck**, or **push-extinction** tasks. In the push-to-spot task, the box is initially placed immediately next to the spot. The distance between the box and the spot is sampled uniformly between 0 and X, where X increases progressively as the artificial pigeons become more adept at solving the task: pigeons "graduate" to the next distance once they achieve good performance (100 successive successes in a maximum duration 50 timesteps each) on the task. Other than box-spot distance, the position of the various elements of the task is randomized for each shaping and test instance. For jump-and-peck, the box is fixed in place underneath a banana, and for push-extinction the box is present with no reward is available. The three shaping tasks are interleaved.

Artificial pigeons trained in this way did not succeed at the test on their first try in an "insightful" manner, unlike real pigeons. Instead, we present results for repeated tests, in which, after training, the pigeons face a succession of randomized test problems (with the box and banana placed randomly).

"Pigeon Insight" Model

Learning is modeled using deep Reinforcement Learning (Sutton & Barto, 2018), specifically an actor-critic algorithm. Reinforcement Learning is learning what to do in order to maximize a reward signal, where obtaining a reward often requires multiple successive actions. To know whether an action was good, it is therefore useful to evaluate the resulting situation, without waiting for the reward itself: if the new situation is promising (as opposed to dire), the tendency to repeat that action in similar contexts should be reinforced (as opposed to weakened). Many reinforcement learning algorithms exploit these ideas by making use of an actor which selects actions, and a critic which evaluates situations and generates a learning signal.

A technical description of these ideas and their implementation is given below in order to make the present work reproducible. Readers who wish to familiarize themselves further with Reinforcement Learning are encouraged to consult the article by Kaelbling, Littman, and Moore (1996) or the more expansive book by Sutton and Barto (2018). For a discussion of the connections between Reinforcement Learning approaches in AI and in psychology, see chapters 14 and 15 of Sutton and Barto (2018, accessible online).

The simulated environment is a Markov Decision Process, where images count as states s from a set $\mathcal{S}$ ($s \in \mathcal{S}$), pigeon behavior as actions $a \in \mathcal{A}$, with rewards $r \in \mathcal{R}$ (10 on successful completion, 0 otherwise), and a transition function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ defining the dynamics of the environment. In an actor-critic algorithm, the agent, with no prior knowledge of the environment dynamics, learns from experience a policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0,1]$ (mapping states to a probability of selecting each action, based on the parameters θ of the *actor*) and a value function $v_w : \mathcal{S} \to \mathbb{R}$ (which denotes the agent's future prospects, or return, assuming it follows its policy from the current state; it is approximated as $\hat{v}_w$ based on parameters w of the *critic*). Actor-critic systems are considered more plausible models for biological agents (Sutton & Barto, 2018, pp395-402).

Two convolutional neural networks are used to approximate the value function v as $\hat{v}_w$ (critic network) and to implement the policy (actor network). The architecture is shown for the actor network in figure 3; the critic network is identical save for the last layer, which has only one output and no nonlinearity. Learning proceeds online by gradient descent, according to the update rules:

$$w \leftarrow w + \alpha_w \delta \nabla \hat{v}_w(S')$$
$$\theta \leftarrow \theta + \alpha_\theta \delta \nabla \log \pi_\theta(A|S)$$

Where S is the state, A is the action chosen (according to the policy π), R is the reward, S' the following state, and $\delta = R + \gamma \hat{v}_w(S') - \hat{v}_w(S)$ is the one-step time-difference error. We use a discount γ (0.9) and learning rates α_w and α_θ (0.003 and 0.0003). Thus, by way of the time-difference error, the critic adjusts its estimate of the value of a state based on that of the next state. Meanwhile, the actor learns to preferentially select actions which lead to surprisingly high-valued states (states with positive time-difference errors). The interleaved processes of estimating the value of states and improving the policy leads (demonstrably under certain conditions) to a locally optimal policy. In our implementation, the actor was regularized based on the entropy of its output to ensure continued adequate exploration (as in Mnih et al. (2016)), and learning and acting was parallelized (16 concurrent agents) to accelerate computation time.

A first cohort of 20 agents was given shaping training up to a performance of 90% completion within 50 time-steps, and then continued learning in the test condition; we call this *condition 1*. A second cohort of 20 was given more extensive training (150,000 additional timesteps after meeting the criteria for condition 1); we call this *condition 2*. The expectation was that additional training would result in overfitting and render transfer more difficult (as observed for human insight in the work of Wiley (1998)). A third cohort was directly given the test without any prior training; we call this *condition 3*. In all cases, the primary measure is the rate of success: how likely each simulated pigeon is to succeed at its task within 50 time-steps. This is measured as a running average (cf. figure 4).

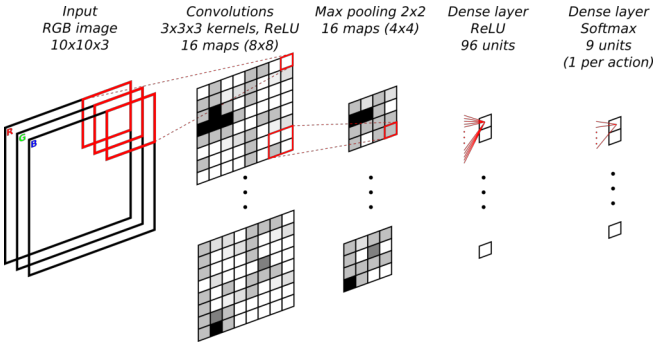

Figure 3: The neural network architecture used for the actor. For illustrative purposes, example activations are given in shades of grey, and example connections in red.

Results

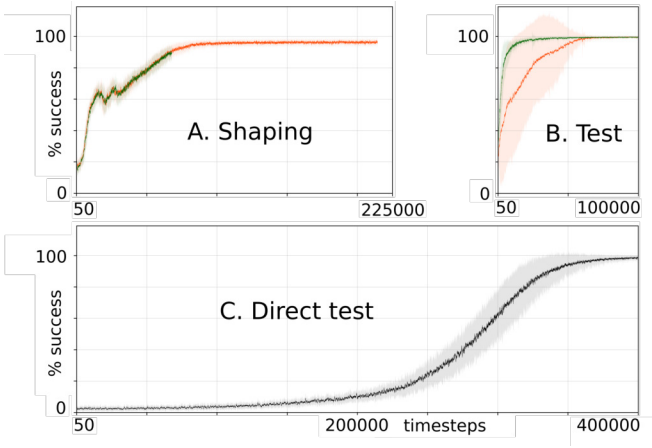

Figure 4: Performance of the actor-critic model. All graphs show the success rate for 20 runs, smoothed over 100 timesteps; color bands show the standard deviation. Note that the success rate is initially high during shaping because the shaping tasks are easy in the beginning, and progressively made more difficult as performance increases. Graphs A and B show the performance for conditions 1 (dark green) and 2 (lighter orange) for the shaping and test, whereas Graph C shows the performance for condition 3 (naive agents). Condition 2 had worse average performance on the test, with greatly increased variance.

The shaping program was successful in improving performance. Agents in conditions 1 and 2 transferred successfully to the final task, rapidly learning the new task in condition 1, although there was often a delay for those of condition 2 who had been given more extensive training. Agents in condition 2 showed considerable variance in the transfer - some of them necessitating a much longer time than others. Condition 1 and 2 both showed substantially better performance than condition 3 on the test. These results are shown in figure 4.

In condition 1, agents adapted rapidly to the new task. However, in condition 2 there often was a period of "impasse" during which the agents displayed low performance; individual curves are shown in figure 5. These impasses remained short compared to condition 3, but were substantial compared to condition 1 (see figure 5b); impasse was followed by a rapid performance increase, which was accompanied by an increase in expected value as estimated by the critic components of the agents. There was also an increase in positive time-difference errors, which correspond to unexpected progress, from the agent's perspective.

Discussion

Did the simulated pigeons experience "insight"? Unlike the real pigeons, few solved the test situation on their first try, suggesting that out-of-the-box RL is not sufficient for insight. However, especially for condition 2, they displayed patterns that are reminiscent of findings on the insight process. Recall the characteristics of insight enumerated in the background section. Many of them are reflected in the behavior of the deep RL agents:

1. The insight sequence: in condition 2 especially, one can distinguish a fruitless search/impasse phase from a sudden resolution.

2. Sudden and surprising solution: the sudden increase of "feelings of warmth" in humans Metcalfe and Wiebe (1987), i.e. their subjective appreciation of how close they are to solving the problem, resembles the sudden increase of the estimated value function in the agents. (Recall that the value function, estimated by the critic component of the agents, measures their expectation of acquiring reward; it is thereby analogous to the "feelings of warmth" measure.) The steepness of the learning curve for shaped agents (conditions 1 and 2) is sudden by comparison to naive agents (condition 3).

3. Restructuring: the agents ought to behave "as if" the yellow objective is the green spot with which they trained. We conjecture that when the agent learns this, the rest of the correct solution "falls into place" rapidly due to prior learning[4].

4. Role of experience: "insight" is made possible by prior experience, with extensive experience having an ambiguous role – too much experience being detrimental to performance, as in Wiley (1998).

Additionally, we note several associations between the concepts of reinforcement learning and those of psychology, which are known in RL and cognitive psychology, but have received little attention in the insight literature. Readers familiar with RL may have recognized transfer and curriculum learning techniques used for instance in robotics; those well-read in psychology noticed that the overfitting of condition 2

[4]The distributed nature of neural networks makes this difficult to verify; we reserve such investigations to future work.

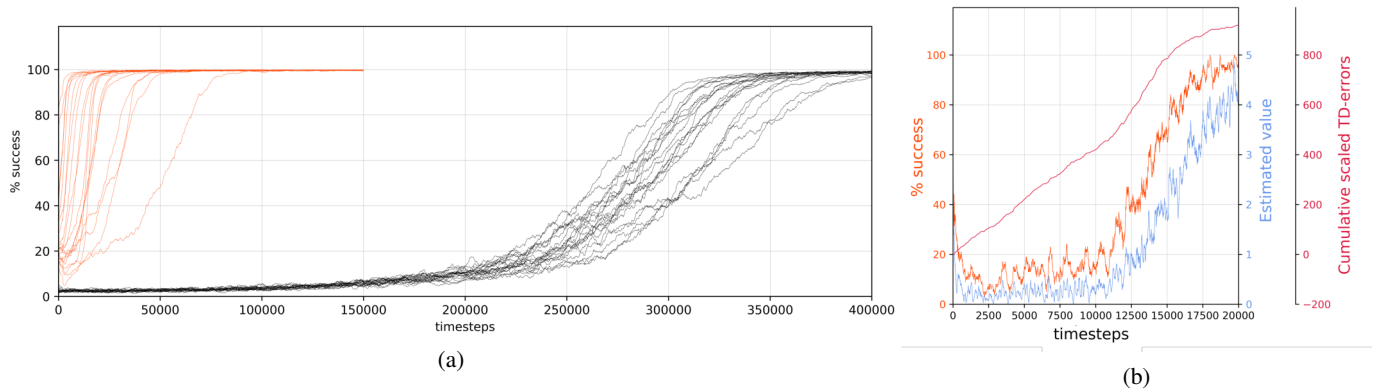

Figure 5: **(a)** All "overfitting" transfer curves (orange, left), compared to learning from scratch (black, right), over 20 runs. (All curves have been smoothed for readability, showing the average over 4000 timesteps.).
(b)A single learning curve on the test (one of the 20 shown in orange in subfigure (a)). The top curve is the cumulative TD-error, the middle curve is the % of success, the bottom curve is the estimated value.

is reminiscent of the Einstellung effect, by which prior experience gets in the way of finding an optimal solution to a new problem (Luchins, 1942).

Thus although the artificial pigeons needed a considerable amount of interaction with the test by trial and error (note that both pigeons (Epstein et al., 1984) and chimpanzees (Köhler, 1921; Birch, 1945) also showed some amount of trial and error even during the test), they also presented learning patterns resembling those of insight: namely (1) a comparatively sudden increase of performance, accompanied by (2) an increase in expected return, which (3) was made possible by a "just-right" amount of prior experience.

The proposed model thus displays some characteristics of insight while being limited in other respects. The most notable of these limitations is the time needed to discover the full solution during the test. This might be a matter of learning quickly from limited data during the test (this is the solution favored by Epstein (2014)), or of making use of more profound regularities in the shaping tasks, e.g. via temporal abstraction as suggested by Colin, Belpaeme, Cangelosi, and Hemion (2016). Alternatively, they might identify new regularities between old and new tasks on the fly (Friston et al., 2017), or use off-policy learning to make use of prior experience (as suggested by Richard Sutton in personal communication; cf. Tolman and Honzik (1930)). Finally, perhaps the use of model-based reinforcement learning allows for trial and error to occur in subconscious simulation "in the agent's mind" (Hamrick et al., 2016; Hélie & Sun, 2010). These various approaches are not mutually exclusive - indeed, all of them are compatible, and perhaps only some (yet-to-be-realized) combination of all of these methods can produce behavior truly comparable to animal and human insight.

Conclusion

Insight problem-solving was historically presented by Köhler as a challenge for Thorndike's concepts of animal learning. Nowadays Aha!-moments, due to the sheer speed of the phenomenon in human beings and animals, remain puzzling for modeling approaches that rely on statistical trial-and-error. However, their apparent reliance on learning and thereby generalization, and their representational component, has made them equally challenging for traditional cognitive models. Both symbolic and statistical approaches have difficulty explaining insight.

We suggest that the statistical approaches offer, after all, a promising avenue of research for explaining insight. The established importance of learning for insight (Birch, 1945; Wiley, 1998) suggests a model based on learning. Our results show how transfer learning can accelerate the resolution of a new problem to the point of making it seem, in contrast to solving it "from scratch", rather sudden. This and the focus of contemporary machine learning techniques on representation designates them as clear candidates for modeling insight.

We have presented a simulation of a psychological experiment on insight, with the aim of proposing a model of the cognitive processes underlying animal behavior in the experiment. Our artificial pigeons were not a match for the real pigeons performance-wise: they required more experience to solve a simplified version of the task; their "insights" were slower and clumsier. However the proposed model showed qualitative properties reminiscent of those seen in pigeons. It is a long way to recreating the insights of chimpanzees, let alone humans; we have given some directions for future research, and we hope that the methodology presented here (replicating insight studies on non-human animals) can serve as a basis for future investigations of the creativity of great apes - such as ourselves.

Acknowledgments

This work was completed as part of Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN, CogNovo, grant number 604764. We would like to thank Dr. Robert Epstein for

his helpful comments, and the reviewers whose constructive criticism helped make this a better article.

References

Birch, H. G. (1945). The relation of previous experience to insightful problem-solving. *Journal of Comparative Psychology*, *38*(6), 367–383.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods*, *35*(4), 634–639.

Colin, T. R., Belpaeme, T., Cangelosi, A., & Hemion, N. (2016). Hierarchical reinforcement learning as creative problem solving. *Robotics and Autonomous Systems*, *86*, 196–206.

Cook, R. G., & Fowler, C. (2014). "Insight" in pigeons: absence of means–end processing in displacement tests. *Animal cognition*, *17*(2), 207–220.

Epstein, R. (2014). On the orderliness of behavioral variability: Insights from generativity theory. *Journal of Contextual Behavioral Science*, *3*(4), 279–290.

Epstein, R., Kirshnit, C., Lanza, R., & Rubins, L. (1984). "insight" in the pigeon: antecedents and determinants of an intelligent performance. *Nature*, *308*, 61–62.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural computation*, *29*(10), 2633–2683.

Hamrick, J. B., Pascanu, R., Vinyals, O., Ballard, A., Heess, N., & Battaglia, P. (2016). Imagination-based decision making with physical models in deep neural networks. In *Proceedings of the NIPS 2016 workshop on intuitive physics.*

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: a unified theory and a connectionist model. *Psychological review*, *117*(3), 994–1024.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive psychology*, *22*(3), 374–419.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*(7), 1000–1009.

Köhler, W. (1921). *Intelligenzprüfungen an menschenaffen [the mentality of apes]*. Berlin: Springer-Verlag.

Kounios, J., & Beeman, M. (2015). *The eureka factor: Creative insights and the brain*. Random House.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological monographs*, *54*(6), i–95.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 176–201.

MacLellan, C. J. (2011). An elaboration account of insight. In *AAAI fall symposium: Advances in cognitive systems* (pp. 194–201).

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & cognition*, *15*(3), 238–246.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 1928–1937).

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge University Press.

Poincaré, H. (1909). *Science et méthode*. Flammarion.

Rothenberg, A. (1995). Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule. *The American Journal of Psychology*, *108*(3), 419–438.

Schilling, M. A. (2005). A" small-world" network model of cognitive insight. *Creativity Research Journal*, *17*(2-3), 131–154.

Shettleworth, S. J. (2012). Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *66*(4), 217-226.

Sprugnoli, G., Rossi, S., Emmerdorfer, A., Rossi, A., Liew, S.-L., Tatti, E., ... Santarnecchi, E. (2017). Neural correlates of eureka moment. *Intelligence*, *62*, 99–118.

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, *37*(8), 1132–1149.

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction*. MIT Press. (Accessible at `http://incompleteideas.net/book/the-book-2nd.html`)

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109.

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J., ... Windischberger, C. (2018). Ultra-high-field fmri insights on insight: Neural correlates of the aha!-moment. *Human brain mapping*, *39*(8), 3241–3252.

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, *4*, 257–275.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*(6972), 352–355.

Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, *21*(1), 5–39.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & cognition*, *26*(4), 716–730.