The AAAI 2017 Spring Symposium on
Science of Intelligence: Computational Principles of Natural and Artificial Intelligence
Technical Report SS-17-07

Human Learning in Atari

Pedro A. Tsividis
Department of Brain and Cognitive Sciences
MIT

Thomas Pouncy
Department of Psychology
Harvard University

Jacqueline L.Xu
CSAIL
MIT

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
MIT

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

Abstract

Atari games are an excellent testbed for studying intelligent behavior, as they offer a range of tasks that differ widely in their visual representation, game dynamics, and goals presented to an agent. The last two years have seen a spate of research into artificial agents that use a single algorithm to learn to play these games. The best of these artificial agents perform at better-than-human levels on most games, but require hundreds of hours of game-play experience to produce such behavior. Humans, on the other hand, can learn to perform well on these tasks in a matter of minutes. In this paper we present data on human learning trajectories for several Atari games, and test several hypotheses about the mechanisms that lead to such rapid learning.

Introduction

Reinforcement learning algorithms using deep neural networks have begun to surpass human-level performance on complex control problems like Atari games (Guo et al. 2014; Mnih et al. 2015; Van Hasselt, Guez, and Silver 2015; Schaul et al. 2015; Stadie, Levine, and Abbeel 2015). However, these algorithms require hundreds of hours of gameplay to achieve human levels of performance, while our experiments show that humans are able to learn these tasks in a matter of minutes. This suggests that these algorithms may employ different representations and learning mechanisms than humans. Many decades of research in cognitive science has shown that humans have early-arising "start-up" software – rich representations about objects and physics (Spelke 1990; Baillargeon 2004; Baillargeon et al. 2009; Rips and Hespos 2015) and about agents (Johnson, Slaughter, and Carey 1998; Tremoulet and Feldman 2000; Csibra et al. 2003; Schlottmann et al. 2006; Spelke and Kinzler 2007; Csibra 2008; Kiley Hamlin et al. 2013), and that humans have the capacity to rapidly acquire new concepts (Carey 1978; Landau, Smith, and Jones 1988; Markman 1989; Bloom 2000; Xu and Tenenbaum 2007; Lake, Salakhutdinov, and Tenenbaum 2015) and build intuitive theories (Murphy and Medin 1985; Carey 1985; Gopnik, Meltzoff, and Bryant 1997), which they can use to explain (Lombrozo 2009; Williams and Lombrozo 2010), predict (Rips 1975; Murphy and Ross 1994), and imagine (Ward 1994; Jern and Kemp 2013). For a thorough review of the human "start-up" software, see Lake et al. (2016).

In addition to bringing early-arising representations to bear on Atari tasks, humans come equipped with rich prior knowledge about the world – for example, knowledge about keys, doors, ice, birds, and so on. While such knowledge could give humans an edge over AI algorithms, we believe that it plays a minimal role in humans' ability to rapidly master these tasks. Instead, we believe that strong priors specified at a more general level are what give rise to rapid learning. This includes priors about objects as spatiotemporally contiguous entities whose properties can be learned from experience; an imperative to explore these objects and to observe available evidence to rapidly build theory-like models of the Atari worlds they encounter; and an ability to use these models to simulate possible future worlds and generate effective plans.

In this paper we present some of the first systematic data on human Atari gameplay. Going beyond simply observing gameplay, we experimentally manipulate (1) prior knowledge about specific objects, (2) prior knowledge about the game environment and rules, and (3) observational learning experience. These manipulations allow us to test different hypotheses about the nature of human learning in Atari.

Human Atari Gameplay

Part I - Gameplay

Methods We selected two games in which humans outperform the asymptotic

Double Deep Q-Network (DDQN) (Frostbite and Venture), and two games in which the DDQN outperforms humans (Amidar and Stargunner). Participants found through Amazon Mechanical Turk were assigned to one game that they had not previously played, and were told that they should play for a minimum of 15 minutes. All participants were paid $2.00 and were promised bonus pay of up to $2.00 extra depending on their cumulative score performance. Prior to playing the game, participants were told only that they could use the arrow keys and the spacebar, and that, beyond that, they should try to figure out how the game worked in order to play well. Participant numbers are as follows: (Frostbite: 71, Venture: 18, Amidar: 19, Stargunner: 19).

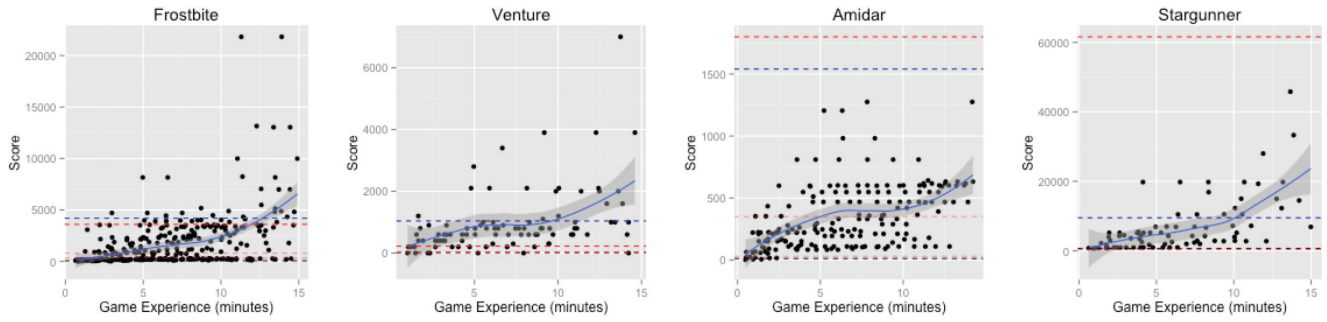

Figure 1: Human learning curves for four Atari games. Black horizontal line: random play. Blue horizontal line: 'expert' play. Red horizontal lines: DDQN after 10, 25, and 200 million frames of game-play experience (46, 115, and 920 hours, respectively).

Learning Curves Figure 1 shows learning curves for the four games. Each points represents a (human, time, score) tuple; the score reported at each point is the highest score obtained by that particular subject after that amount of cumulative gameplay experience. For comparison, we have also plotted random performance (black horizontal), human 'expert'[1] play (blue horizontal line), and DDQN performance after 10, 25, and 200 million frames of game-play experience (46, 115, and 920 hours, respectively), in red (bottom to top)[2]. We highlight a few qualitative observations: human performance is above random performance within the first minute of play; in three out of the four games, humans reach 'expert' performance within the allotted 15 minutes (on Amidar they are well on their way); in all of the games, humans exceed the DDQN's 10- and 25-million-frame score within just a few minutes; in Frostbite and Venture, as mentioned earlier, humans exceed even the DDQN's asymptotic performance.

Of course, the 10-million (46-hr) and 25-million-frame (115-hr) comparisons are unfair – after all, in addition to having potential cognitive advantages, humans come to these tasks with a working visual system, while the DDQN has to learn a visual system from scratch. With this in mind, we can also look at the DDQN's rate of improvement and compare it to human rates of improvement at score-matched points – including points at which the DDQN is doing relatively well. Figure 2 shows such a comparison for three of our four games; a comparison for Venture is missing because humans immediately exceeded the DDQNs performance in the 10-40 million frame range. Note that the rate of improvement is measured in log units.

[1] 'Expert' refers to the human game tester employed by DeepMind and used as the human benchmark in the original DQN paper. After the tester trained on each game for two hours, their scores over the 20 subsequent game episodes that lasted over 5 minutes were averaged and reported.

[2] Data taken from Schaul (2015). As of this writing, this was the highest-performing Deep RL model for which we could obtain learning-curve-like data for a large range of games. The model significantly outperforms the original DQN, as well as several subsequent variants.

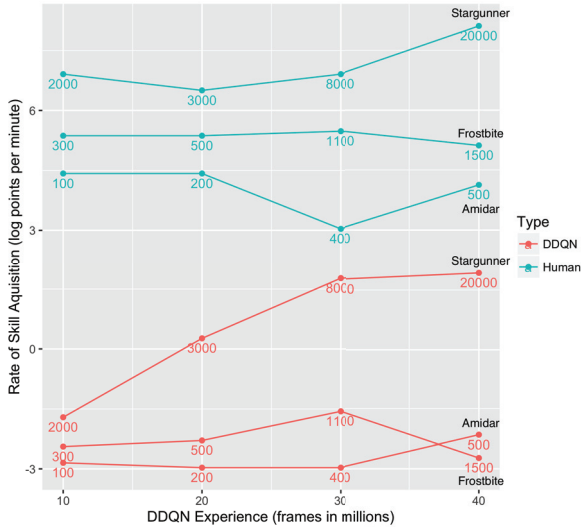

Figure 2: Learning rate comparison. X-axis: DDQN experience, in millions of frames. Y-axis: Rate of improvement in log (points per minute), estimated using finite differences. Human rate of improvement is taken from score-matched points (shown as numbers annotating the curves). A comparison for Venture is missing because humans immediately exceeded the DDQN's performance in the 10-40 million range. DDQN estimates are made from Figure 7 in Schaul et al. (2015).

Part II - Experimental Manipulations

The game of Frostbite is one of several games worth analyzing closely, as humans exhibit particularly impressive performance relative to the DDQN. In what follows we present several experimental manipulations. Our hope is that by understanding the representations that facilitate human performance in this game (and eventually in other games), we can help pave the way toward designing more human-like AI agents.

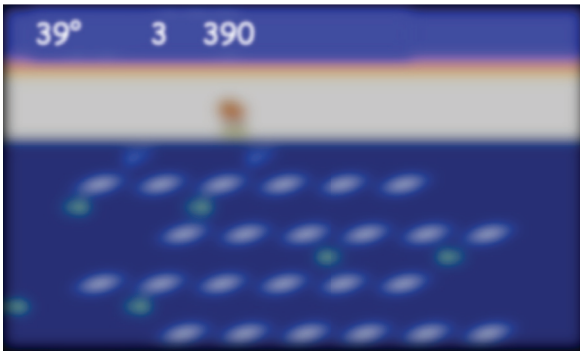

Figure 3: Screenshot of a blurred version of Frostbite. The light blue items are birds; the green items are fish.

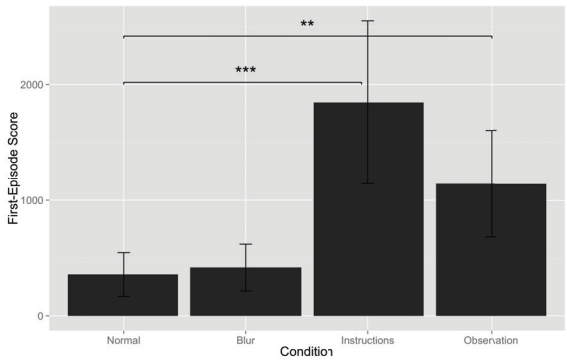

Figure 4: First-episode mean scores and 95% confidence interval for normal, 'blurred', 'instructions', and 'observation' condition. Note that blurring the screen has no effect, whereas the instructions and observing another player allow humans to capture approximately 1000 points in their first episode – this corresponds to human performance after about 5 minutes of play under normal conditions.

Obscuring Object Identity One specific hypothesis as to why humans rapidly learn to perform well at Frostbite is that they come equipped with strong priors about the objects they encounter. Parsing the game screen into platforms, igloos, birds, and fish must clearly make the task easier for humans, as humans know things about the properties of each of these objects: platforms provide support; igloos provide shelter; fish and birds can be eaten. To test whether such knowledge is in fact a benefit to humans, we created a blurred version of the game, in which objects could be only be identified as generic objects – that is, we masked the semantic identity of the objects.[3] Figure 3 shows an example game screen. We present data from this condition after describing the other experimental manipulations.

Reading the Instruction Manual If people's rapid learning is due to the formation of a model-like representation, then anything that would enable them to learn this representation should result in an increase in early performance. To test this hypothesis, we provided participants with the opportunity to read the game's original instruction manual prior to playing. The intent was to provide players with a short description of critical aspects of gameplay, which was mostly informative about objects and their roles in the game, as well as the goal of the game. Subjects read the manual, answered a short questionnaire intended to check that they understood the rules, and then played for 15 minutes.

Learning from Observation Humans' ability to form a theory-like representation of the game should also be aided by observing others play. We randomly selected an episode in the 75-85th percentiles of episodes from the first round of experiments, and had all participants in this condition watch a video of that episode prior to playing. The episode corresponded to the 79th percentile of all Frostbite episodes, and lasted 1 minute, 26 seconds.

Results Figure 4 shows means and 95% CIs of first-episode scores for normal, 'blur', 'instructions', and 'observation' conditions (participant Ns are 71, 63, 72, 72, respectively). The difference in first-episode performance between normal and blurred conditions is not significant (p=0.663. Mean, CIs: Normal: 356, [167, 545]. Blur: 417, [216, 618]). This is unsurprising: a bird, a priori, could be useful (it can be hunted and eaten) or it could be harmful (it can attack you). In the actual game the interaction between the agent and the bird is, in fact, quite implausible a priori – real-life birds are much lighter than humans and are generally not capable of pushing them, and yet, in the game they do so. While priors about object identity and the resulting behaviors of objects may be minimally useful to a novice player, most of the important properties of objects in these games come from their role in the particular game. It is after observing objects' movements and interactions that humans rapidly form theories of the game dynamics.

As we expected, reading the instruction manual and observing competent players provided participants with a significant first-episode advantage over normal play (Instructions vs. Normal: p=.0001. Observation vs. Normal: p=.002. Mean, CIs: Normal: 356, [167, 545]. Instructions: 1848, [1144, 2552]. Observation: 1144, [683, 1605]). Had players only learned from observation, one might posit that they simply copied a successful policy. However, the fact that the instruction manual was helpful suggests that this is not the case, and hints at the possibility that humans used this information to build a model of the game dynamics, which they then used to play successfully.

Discussion

The real power of human intuitive theories is that they enable humans to explain the world, generalize from few examples, think counterfactually, and generate effective plans. The experiments above show that humans are capable of learning complex Atari tasks from just a few minutes of gameplay, and that their behavior is aided by information that would be helpful in theory-building. Future efforts will address the nature of this theory-building.

[3]This resulted in blurred birds, fish, crabs, and clams, but clearly-identifiable water, ice floes, and igloos.

Acknowledgments

The authors would like thank Greg Hale for his helpful web-programming advice, and Paulo Peccin, author of Javatari. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Baillargeon, R.; Li, J.; Ng, W.; and Yuan, S. 2009. An account of infants physical reasoning. *Learning and the infant mind* 66–116.

Baillargeon, R. 2004. Infants' physical world. *Current directions in psychological science* 13(3):89–94.

Bloom, P. 2000. *How children learn the meanings of words*. Number Sirsi) i9780262523295. MIT press Cambridge, MA.

Carey, S. 1978. 8 the child as word learner.

Carey, S. 1985. Conceptual change in childhood.

Csibra, G.; Bıró, S.; Koós, O.; and Gergely, G. 2003. One-year-old infants use teleological representations of actions productively. *Cognitive Science* 27(1):111–133.

Csibra, G. 2008. Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* 107(2):705–717.

Gopnik, A.; Meltzoff, A. N.; and Bryant, P. 1997. *Words, thoughts, and theories*, volume 1. Mit Press Cambridge, MA.

Guo, X.; Singh, S.; Lee, H.; Lewis, R. L.; and Wang, X. 2014. Deep learning for real-time atari game play using offline monte-carlo tree search planning. In *Advances in neural information processing systems*, 3338–3346.

Jern, A., and Kemp, C. 2013. A probabilistic account of exemplar and category generation. *Cognitive psychology* 66(1):85–125.

Johnson, S.; Slaughter, V.; and Carey, S. 1998. Whose gaze will infants follow? the elicitation of gaze-following in 12-month-olds. *Developmental Science* 1(2):233–238.

Kiley Hamlin, J.; Ullman, T.; Tenenbaum, J.; Goodman, N.; and Baker, C. 2013. The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science* 16(2):209–226.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2016. Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Landau, B.; Smith, L. B.; and Jones, S. S. 1988. The importance of shape in early lexical learning. *Cognitive development* 3(3):299–321.

Lombrozo, T. 2009. Explanation and categorization: How why? informs what?. *Cognition* 110(2):248–253.

Markman, E. 1989. *Categorization and Naming in Children: Problems of Induction*. Bradford Books. MIT Press.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Murphy, G. L., and Medin, D. L. 1985. The role of theories in conceptual coherence. *Psychological review* 92(3):289.

Murphy, G. L., and Ross, B. H. 1994. Predictions from uncertain categorizations. *Cognitive psychology* 27(2):148–193.

Rips, L. J., and Hespos, S. J. 2015. Divisions of the physical world: Concepts of objects and substances. *Psychological bulletin* 141(4):786.

Rips, L. J. 1975. Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior* 14(6):665–681.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2015. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Schlottmann, A.; Ray, E. D.; Mitchell, A.; and Demetriou, N. 2006. Perceived physical and social causality in animated motions: Spontaneous reports and ratings. *Acta Psychologica* 123(1):112–143.

Spelke, E. S., and Kinzler, K. D. 2007. Core knowledge. *Developmental science* 10(1):89–96.

Spelke, E. S. 1990. Principles of object perception. *Cognitive science* 14(1):29–56.

Stadie, B. C.; Levine, S.; and Abbeel, P. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

Tremoulet, P. D., and Feldman, J. 2000. Perception of animacy from the motion of a single object. *Perception* 29(8):943–951.

Van Hasselt, H.; Guez, A.; and Silver, D. 2015. Deep reinforcement learning with double q-learning. *CoRR, abs/1509.06461*.

Ward, T. B. 1994. Structured imagination: The role of category structure in exemplar generation. *Cognitive psychology* 27(1):1–40.

Williams, J. J., and Lombrozo, T. 2010. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science* 34(5):776–806.

Xu, F., and Tenenbaum, J. B. 2007. Word learning as bayesian inference. *Psychological review* 114(2):245.