Towards Fluid Machine Intelligence: Can We Make a Gifted AI?

Ian Davidson,[1] **Peter B. Walker**[2]
[1]University of California, Davis
[2]Office of Naval Research
davidson@cs.ucdavis.edu, peter.b.walker1@navy.mil

Abstract

Most applications of machine intelligence have focused on demonstrating *crystallized* intelligence. Crystallized intelligence relies on accessing problem-specific knowledge, skills and experience stored in long term memory. In this paper, we challenge the AI community to design AIs to **completely** take tests of *fluid* intelligence which assess the ability to solve novel problems using problem-independent solving skills. Tests of fluid intelligence such as the NNAT are used extensively by schools to determine entry into gifted education programs. We explain the differences between crystallized and fluid intelligence, the importance and capabilities of machines demonstrating fluid intelligence and pose several challenges to the AI community, including that a machine taking such a test would be considered gifted by school districts in the state of California. Importantly, we show existing work on seemingly related fields such as transfer, zero-shot, life-long and meta learning (in their current form) are not directly capable of demonstrating fluid intelligence but instead are task-transductive mechanisms.

Introduction, Motivation and Significance

The long term aim of Artificial Intelligence (AI) is for a machine to exhibit levels of intelligence similar to those possessed by humans. Over the years, this has been *seemingly* achieved in a plethora of different and more challenging circumstances: digit recognition (1980s) (LeCun et al. 1990), 3D object recognition (1990s) (Murase and Nayar 1995), achieving a level of play equivalent to the world champion of checkers (2000s) (Schaeffer et al. 2007), beating a champion at Jeopardy (2010s) (Ferrucci et al. 2013), and more recently beating a world champion at the game of GO (Silver et al. 2016). However, these previous demonstrations of machine intelligence have effectively focused on *crystallized* intelligence which is the ability to use application specific previously learnt knowledge, skills and/or experience. For example, despite AlphaGo's resounding success, none of the knowledge it has demonstrated can be applied to other situations.

The field of cognitive psychology defines *fluid* intelligence as reasoning and/or problem solving that can be applied in a variety of situations including those not seen as yet. Humans use fluid intelligence extensively, so that, unlike machines, they do not need large training sets dedicated to each and every new problem they encounter. Instead, we use problem independent knowledge. However, to date, the AI community has not sought to develop machines with fluid intelligence which would require identifying and using complex patterns beyond those found in any one specific problem. This is not doubt partially due to most work optimizing performance on a specific task.

A number of tests have been developed to assess fluid intelligence in children and adults including the Raven's Progressive Matrices (RPM) and Naglieri Nonverbal Ability Test (NNAT). The latter is given to elementary school-age children in a number of different states to determine their eligibility to enter various gifted and talented education (GATE) programs (see https://www.cde.ca.gov/sp/gt/ for the state of California's program). Though there are variations in these different non-verbal assessments, they all have a similar mode of presentation. For example, in Figure 1 we see examples of the NNAT-like questions with three rows and three columns of tiles with one tile been with-held. The testee is asked to choose which of a list of given tiles best completes the pattern. Consider the left most question, which is an example of reasoning by analogy. Here, we see there is a horizontal pattern (the common color of the back-most tile) which restricts the answer set to be the first, second, or third option. Similarly, a vertical pattern (the common color of the front tile) can be applied also which allows the testee to arrive at the second option. For the purposes of the discussion here, it should be pointed out that the puzzle was solved using knowledge beyond the puzzle itself.

In this paper, we pose several challenges to the AI community to have a machine demonstrate fluid intelligence. Specifically, we focus on non-verbal tests of intelligence such as the RPM and NNAT.

- *Two Year Challenge.* Our short term challenge is for the machine to score 95% accuracy on a forty TRUE/FALSE question version of the NNAT Level D. That is, rather than choose amongst several options, answer TRUE/FALSE to a given tile being a continuation of the patterns in the question. This is a test of **verification** ability. That is, the AI need only verify if a given response provides a correct answer.

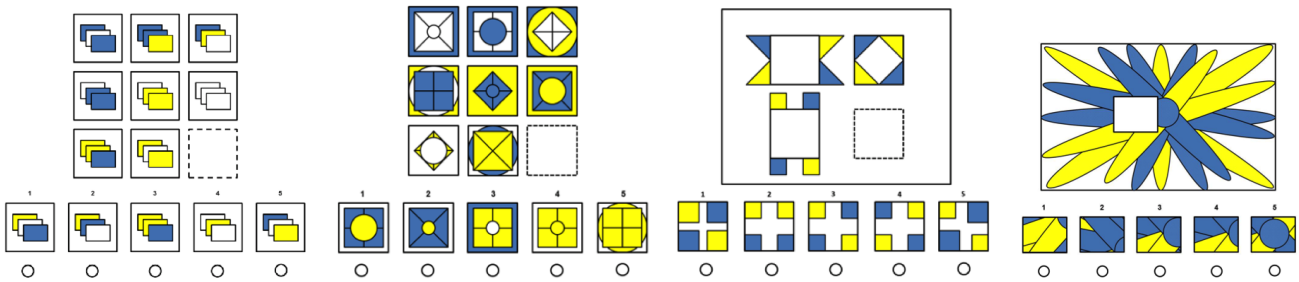

Figure 1: Examples of a non verbal test of fluid intelligence. From left to right reasoning tests of Analogy, Serial, Spatial and Pattern Completion. Note that the NNAT test addresses all four types of reasoning whilst most RPM tests focus on analogy and many systems are based around analogy reasoning methods (i.e. (Correa, Prade, and Richard 2012))

- Our medium term *Five Year Challenge* to the community is for a machine to score 95% accuracy on a forty question version of the NNAT Level D without any **human involvement**.[1] This would be sufficient for the machine to be considered gifted amongst third grade children for the majority of schools in California. This is a test of **recognition** ability to recognize the correct answer from a given list.
- Our long term *Ten Year Challenge* to the community is for a machine to **generate** the correct answer with 95% accuracy on a forty question version of the NNAT Level D. Note, here the potential answers are **not** given to the machine and it must generate the correct answer. This is a test of **recall**. Importantly, it guarantees fluid intelligence since the machine can't rely on secondary mechanisms such as elimination or artifacts to guess the correct answer.

In the next section, we overview the history of intelligence and aptitude testing and explain the significance, differences and the need to focus on non-verbal tests such as RPM and NNAT to assess machine fluid intelligence. In the subsequent section we discuss how existing AI systems, though successful, do not test fluid intelligence. Most importantly areas such as transfer, zero-shot, meta and life-long learning in their current form demonstrate crystallized and not fluid intelligence. In our penultimate section we outline progress on work for machines to take a non-verbal test of intelligence, most commonly RPM. We show that existing work does not completely take the test but instead rely on a human coding the problem for a machine. Finally, we conclude by summarizing our proposal.

Tests of Crystallized and Fluid Intelligence

Spearman (Spearman 1904) conjectured that even though people exhibit intelligence in different ways (i.e. math ability, problem solving or even crossword ability) there exists one underlying type of intelligence, the g-factor. L.L. Thurnstone (Thurstone 1938) pioneered the first alternative theory in the field of psychometric which involved administering 50+ tests and grouping these tests into 7 types of intelligence/skills (spatial ability, verbal comprehension, word fluency, perceptual speed, numerical ability, inductive reasoning and memory). Since this seminal work it has been generally accepted that there are indeed several types of intelligence.

Cognitive scientists have widely argued that g-factor can be broken into two separate but measurable forms of intelligence: *crystallized intelligence* and *fluid intelligence.* Spearman refers to these as eductive (fluid) and reproductive (crystallized) components of intelligence. The names are chosen to create sharply differing visions. Crystallized intelligence refers to intelligence derived from experience, culture and education and is used to solve problems **previously** seen before. The term crystallized then makes reference to well structured (like a crystal structure) intelligence. In contrast fluid intelligence measures abstract reasoning, mental agility and adaptability to solve new problems **not seen before**. The term fluid refers to intelligence that, like a fluid, can fill any vessel/problem.

Here we intertwine a brief history of how children are tested for giftedness and intelligence and aptitude testing in general. The history of this area is complex and at times painful as tests of intelligence and giftedness were used as the basis of eugenics (Chase 1980).

IQ Tests. Early work identified giftedness with measures of high IQ (Dai 2010) such as the Lorge-Thorndike Intelligence Test introduced in 1954. This test was later revised and named the CogAT6 test which measures intellectual ability for children (see https://www.hmhco.com/programs/cogat) that requires vocabulary and quantitative skills. It is has been criticized as being culturally biased (Dai 2010); (Lohman and Rocklin 1995) in that it tests for abilities learnt by only some demographics. A common criticism is an example question: "A light bulb is to a lamp like a flame is to what?" (the correct answer being candle). However, it is argued this is a test of crystallized intelligence that can only be learnt by those who have access to knowledge about candles (either directly or reading about them). However, IQ tests are often used in multi-dimensional assessment such as mathematics. Here, identification is first assessed for generalized intelligence using any one of the general cognitive ability tests mentioned previously and supplemented with more domain

[1] Many papers purportedly already achieve accuracy in excess of 80% but they require significant human involvement to "code" the problem for the machine. For example, in Figure 1 (left), human guidance is required to code the order of the tiles into a symbolic representation.

specific assessments of mathematical aptitude.

Non-verbal Tests. The U.S department of defense since World War I has used non-verbal tests of intelligence. Group administered tests placed new recruits many of whom had poor or limited English ability (McCallum, Bracken, and Wasserman 2001). These were quickly adopted by GATE school programs after various lawsuits argued tests such as COGAT Form 6 were biased against non-native English speakers (Naglieri, Booth, and Winsler 2004).

The NNAT was developed by Jack Nagileri for Pearson. It is administered to many 3rd grade children in the state of California to determine entrance into GATE. The test measures nonverbal reasoning and general problem solving skills in children. Each test has upto four components: i) Pattern completion, ii) Reasoning by analogy, iii) Serial reasoning and iv) Spatial visualization (see Figure 1). Being completely nonverbal it is considered culture neutral since it does not require the **ability to read, speak or write**. Since the test contains only core mathematical shapes such as circles, squares and triangles it is not biased against socio-economically disadvantaged children or children where English is not the primary language and hence is felt to not require crystallized intelligence. This makes it an ideal choice for an AI as there is no need to incorporate such social information. However, this test is not without its flaws it can possess a wide range of score variability and too many score very highly or lowly (Dai 2010) . Other similar tests are available i.e. Otis-Lennon School Ability Test (OLSA) but contain a verbal component.

Towards Fluid Intelligence

Here we will argue that most AI systems are designed to posses crystallized intelligence as they are typically developed to solve a single problem that is well defined. Consider the quintessential method of supervised machine learning (Gress and Davidson 2018; Gilpin, Eliassi-Rad, and Davidson 2013; Wang et al. 2013; Chattopadhyay et al. 2013; Davidson 2009) or even variations such as semi-supervised learning (Qian and Davidson 2010) where the machine is taught from annotated examples to solve a problem such as digit recognition. Since the definition of a classic supervised learning problem requires the classes to be known apriori we are not solving a novel problem rather training a system to solve an existing problem/task (T) given a data set (D). This is evident as most supervised learning is formulated as an optimization function of fit to the training data plus some regularization term to prevent overfitting. Most importantly, performance is optimized on that particular task and the learnt knowledge, to say recognize digits, can not be used used for related tasks (i.e. recognize letters) let alone distant tasks.

Why Transfer, Multi-Task, One-Shot, Zero-Shot, Meta and Life-Long Learning Is Insufficient.

On the surface it may seem that more recent advances in machine learning are sufficient to allow machines to possess the ability to think and reason abstractly. Here we argue that though many methods superficially appear to be possible of attaining fluid intelligence they are instead what we refer to as *task-transductive*, that is they are useful for learning knowledge relevant to an already **given** task but not for learning generalized knowledge about tasks that are yet to be encountered.

Scenario	Setting
Transfer	Learn T_t from T_s, D_t
Multi Task	Learn $T_1 \dots T_n$ from $D_1 \dots D_n$
One Shot	$T_{n+1} \dots T_{n+m}$ from $D_1 \dots D_{n+m}$
Zero Shot	Learn $T_{n+1} \dots T_{n+m}$ from $D_1 \dots D_n$
Meta	Learn *parameters* to train T_{n+1}
Life Long	Learn T_{n+1} from $T_1 \dots T_n, D_1 \dots D_{n+1}$

Table 1: Common new learning scenarios. All of these are task transductive and do not generate the knowledge to solve unseen tasks. Note for one shot learning the training sets for tasks $n+1 \dots n+m$ contain just one instance.

Transfer Learning. Transfer learning in its simplest form allows applying a model learnt for a source task T_s (learnt from data D_s) to a new target task T_t (learnt from data set D_t and T_s). For example in our earlier work (Qian et al. 2014) we showed how a model to rank cars could be used to build a model to rank trucks using fewer examples. However, after the learning task there is no new knowledge learnt/generated that could be applied to another task such as ranking of buses, rather the problem was task-transductive in that the source task was used to learn a specific target task.

Multi Task Learning. Multi-task learning similarly tries to learn multiple related tasks ($T_1 \dots T_k$) at once using data ($D_1 \dots D_k$). For example in our work (Qian and Davidson 2010) we learnt to recognize many types of scenes using far less data than learning one task at a time. However, again, nothing was learnt that could be generalized beyond the existing multiple tasks and we were task-transductively learning each task from the other.

One and Zero Shot Learning in Vision. The area of computer vision has long sought to mimic the ability of humans to recognize images given the ability of a six year old to to recognize in excess of 10,000 different objects (Biederman 1987). The seminal work on one shot learning (Fei-Fei, Fergus, and Perona 2006) motivates one-shot learning to use (like humans) prior knowledge about object categories to classify new objects. However, most one-shot learning work is limited to learning a similarity function between say faces and determining if a new face has been seen before which cannot be used for other tasks. That is, this function cannot be used for any other purpose beyond face recognition. Zero shot learning attempts to learn a predictive function for a set of classes $T_1 \dots T_n$ and use them to predict new unseen classes $T_{n+1} \dots T_{n+m}$. However, it is important to realize that these new unseen tasks are already known.

Meta Learning. The educational psychological definition of meta learning is to learn about ones own learning. However, in the computer science literature it can be viewed as being learning to tune parameters of a learning method and not directly relevant to our work (Vilalta and Drissi 2002).

Life Long Learning in NLP. The natural language process-

ing (NLP) field explores the area of learning many tasks $T_1 \ldots T_n$ and leveraging those results for a new task T_{n+1} which is referred as life long learning (Chen and Liu 2016). These tasks could be classification tasks and differs from one and zero shot learning as each task has its own data set $D_1 \ldots D_n$ including the new task D_{n+1}.

Therefore though these methods use similar terminology to those used in the cognitive science literature they do not have the same aim. Therefore, not surprisingly these methods have **not** been used to address non-verbal tests of fluid intelligence. We now discuss the progress made towards that end.

Progress So Far: A Brief History of Machines Taking Non-Verbal Tests of Intelligence

Work on machines taking non-verbal tests of intelligence began over twenty five years ago starting with the seminal work of Carpenter (Carpenter, Just, and Shell 1990). It is important to realize, that many of these systems were designed by the cognitive science literature to better understand human intelligence and the strength/weaknesses of tests not to maximize performance. Most importantly, to our knowledge, none were created with the specific aim of demonstrating a machine possesses fluid intelligence.

Early work was entirely rule based starting with Carpenter's insight (Carpenter, Just, and Shell 1990) that five core rules/patterns: `all-same`, `all-different`, `pairwise-progression`, `addition` and `two-value distribution` could be used to solve many problem instances in RPM. For example Figure 1 actually has three patterns: all-same (row-wise applied to back tile to and column-wise applied to front tile) and all-different (diagonally applied to front tile). However this work and most early work required a human to hand-code problems in a symbolic form that these rules could be applied to. Later work purports to not require hand coding but instead provides a system where humans can provide annotations to the problem which are then used by the system (Forbus and Usher 2002). However, this work still requires significant human involvement.

Work that directly takes the images and requires minimal human involvement has only recently been explored (Lovett and Forbus 2017). Some innovative work (McGreggor, Kunda, and Goel 2010) have realized that the correct answer most increases the self similarity of the completed matrix and have used fractal geometric calculations to construct such a measure. However, such work would be unable to pass our third challenge since it is a test of secondary mechanisms associated with the problem and not a generation mechanism. In this year's ICML a team from DeepMind (Santoro et al. 2018) attempted to allow a machine to solve RPM problems with no human intervention. They viewed each panel in an RPM question as a greyscale input feature map and tried various deep learning architectures to address the problem. Their results are promising but fall far short of our challenges. When the machine is trained and tested on **similarly** shaped, numbered and positioned objects accuracy was 75%. But this drastically declined if the training and test data varied. For example if the machine was trained on squares (i.e. in Figure 1 left) but the tests involved circles or even if different colors or shades were used the machine performs quite poorly.

In summary some existing work provides a useful platform to address the building of fluid intelligent machines. But there are three main limitations with existing work:

- They are mainly aimed at understanding limitations of tests and human intelligence and not directly at a machine exhibiting fluid intelligence.
- They assume human involvement such as converting the pictorial problem into a symbol representation with one notable exception being the poorly performing DeepMind work (Santoro et al. 2018).
- All existing work (in its current form) cannot address our long term challenge of generating the correct answer [2]. This means we can never be sure if they are truly demonstrating fluid intelligence but instead exploiting some secondary measure to answer questions correctly.

A thorough summary of computers taking many tests of intelligence (not just non-verbal tests) exists (Hernández-Orallo et al. 2016) which summarizes work until mid 2015.

Conclusion

Fluid intelligence is what allows humans to reason quickly and productively in new settings. It can be considered one of the most prized elements of human behavior and is tested for in children in many states to determine entry into GATE programs. However, most machine intelligence development focuses on creation of crystallized intelligence which allows the machine to solve a particular task but no problems unrelated to that task. Whilst new work such as transfer, zero-shot and life long learning have similar names used in the cognitive science literature they do not allow a machine to address as yet unseen problems. Instead, we argue they are performing task-transduction: learning to perform tasks that may have little (one-shot learning) or none (zero-shot learning) training data but which have already been identified. We propose three challenges to the AI community: a two year challenge, a five year challenge and a ten year challenge. If a machine meets these challenges it would be considered gifted (according to a third grade standard) by most schools in the state of California. Most existing work on these non-verbal tests of intelligence are mostly aimed at understanding human intelligence and/or the tests, but most importantly (with one exception from DeepMind (Santoro et al. 2018)) are limited to requiring intensive human involvement to hand code problems.

Another interpretation of the challenges set forward in this paper is as an alternative to the Turing test. There is already a body of literature on using tests of intelligence towards this ends (Clark and Etzioni 2016; Schoenick et al. 2016).

References

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review* 94(2):115.

[2] There are apparent exceptions (Correa, Prade, and Richard 2012) but they do not work completely with image creation.

Carpenter, P. A.; Just, M. A.; and Shell, P. 1990. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review* 97(3):404.

Chase, A. 1980. The legacy of malthus; the social costs of the new scientific racism.

Chattopadhyay, R.; Fan, W.; Davidson, I.; Panchanathan, S.; and Ye, J. 2013. Joint transfer and batch-mode active learning. In *International Conference on Machine Learning*, 253–261.

Chen, Z., and Liu, B. 2016. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 10(3):1–145.

Clark, P., and Etzioni, O. 2016. My computer is an honor studentbut how intelligent is it? standardized tests as a measure of ai. *AI Magazine* 37(1):5–12.

Correa, W. F.; Prade, H.; and Richard, G. 2012. When intelligence is just a matter of copying. In *ECAI*, volume 12, 276–281.

Dai, D. 2010. *" The nature and nurture of giftedness: A new framework for understanding gifted education*. New York, NY, USA: Teachers College Press.

Davidson, I. 2009. Knowledge driven dimension reduction for clustering. In *IJCAI*, 1034–1039.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4):594–611.

Ferrucci, D.; Levas, A.; Bagchi, S.; Gondek, D.; and Mueller, E. T. 2013. Watson: beyond jeopardy! *Artificial Intelligence* 199:93–105.

Forbus, K. D., and Usher, J. 2002. Sketching for knowledge capture: A progress report. In *Proceedings of the 7th international conference on Intelligent user interfaces*, 71–77. ACM.

Gilpin, S.; Eliassi-Rad, T.; and Davidson, I. 2013. Guided learning for role discovery (glrd): framework, algorithms, and applications. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 113–121. ACM.

Gress, A., and Davidson, I. 2018. Human guided linear regression with feature-level constraints.

Hernández-Orallo, J.; Martínez-Plumed, F.; Schmid, U.; Siebers, M.; and Dowe, D. L. 2016. Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence* 230:74–107.

LeCun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; and Jackel, L. D. 1990. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, 396–404.

Lohman, D., and Rocklin, T. 1995. *International handbook of personality and intelligence*. New York: Plenum. chapter Current and recurrent issues in the assessment of intelligence and personality.

Lovett, A., and Forbus, K. 2017. Modeling visual problem solving as analogical reasoning. *Psychological review* 124(1):60.

McCallum, R. S.; Bracken, B. A.; and Wasserman, J. D. 2001. *Essentials of nonverbal assessment.* John Wiley & Sons Inc.

McGreggor, K.; Kunda, M.; and Goel, A. 2010. A fractal analogy approach to the raven's test of intelligence. In *Workshops at the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Murase, H., and Nayar, S. K. 1995. Visual learning and recognition of 3-d objects from appearance. *International journal of computer vision* 14(1):5–24.

Naglieri, J. A.; Booth, A. L.; and Winsler, A. 2004. Comparison of hispanic children with and without limited english proficiency on the naglieri nonverbal ability test. *Psychological Assessment* 16(1):81.

Qian, B., and Davidson, I. 2010. Semi-supervised dimension reduction for multi-label classification. In *AAAI*, volume 10, 569–574.

Qian, B.; Wang, X.; Cao, N.; Jiang, Y.-G.; and Davidson, I. 2014. Learning multiple relative attributes with humans in the loop. *IEEE Transactions on Image Processing* 23(12):5573–5585.

Santoro, A.; Hill, F.; Barrett, D.; Morcos, A.; and Lillicrap, T. 2018. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, 4477–4486.

Schaeffer, J.; Burch, N.; Björnsson, Y.; Kishimoto, A.; Müller, M.; Lake, R.; Lu, P.; and Sutphen, S. 2007. Checkers is solved. *science* 317(5844):1518–1522.

Schoenick, C.; Clark, P.; Tafjord, O.; Turney, P.; and Etzioni, O. 2016. Moving beyond the turing test with the allen ai science challenge. *arXiv preprint arXiv:1604.04315*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484.

Spearman, C. 1904. " general intelligence," objectively determined and measured. *The American Journal of Psychology* 15(2):201–292.

Thurstone, L. L. 1938. Primary mental abilities.

Vilalta, R., and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review* 18(2):77–95.

Wang, X.; Qian, B.; Ye, J.; and Davidson, I. 2013. Multi-objective multi-view spectral clustering via pareto optimization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 234–242. SIAM.