Künstliche Intelligenz manuscript No.
(will be inserted by the editor)

Can machine intelligence be measured in the same way as human intelligence?

Tarek Besold · José Hernández-Orallo · Ute Schmid

Received: date / Accepted: date

Abstract In recent years the number of research projects on computer programs solving human intelligence problems in Artificial Intelligence (AI), Artificial General Intelligence (AGI), as well as in Cognitive Modelling, has significantly grown. One reason could be the interest of such problems as benchmarks for AI algorithms. Another, more fundamental, motivation behind this area of research might be the (implicit) assumption that a computer program that successfully can solve human intelligence problems has human-level intelligence and vice versa. This paper analyses this assumption.

Keywords Intelligence Tests · Strong AI · Psychometric AI · Cognitive Modelling

1 Introduction

As early as the possibility of machine intelligence was considered, the role of human intelligence tests in the development and evaluation of AI was linked to the understanding of what intelligence is and how it should be measured. However, the question of *whether human intelligence tests are valid for the evaluation of machines* has had very different (and opposed) answers, including absolute indifference of or neglecting the question. Why do we find this diversity of answers? Is it because of disparate conceptions of what a machine is and the nature of computation? Is it because of divergent views of what intelligence is and what human intelligence tests measure? Or is it because various breadths and types of psychometric tests are being considered?

In order to find answers, in section 2 we first analyse the conceptual notions of machine intelligence that have been adopted in (strong) AI and Cognitive Science, also introducing the notion of a computational theory of mind, the Church-Turing thesis, and the Physical Symbol System Hypothesis (PSSH) as necessary theoretical foundations. In section 3, we rely on the conception of (human) intelligence and its measurement, mostly but not only from psychometrics. We highlight the relevance of the set of *subjects* to be measured and the set of *problems* that are considered. Section 4 presents previous experience about the use of psychometric tests in AI and some explicit claims about their convenience for measuring machine intelligence and for stimulating the progress of AI. In section 5 we are ready to unravel the question by the use of several *arguments*, and clarify whether and, if so, under which circumstances human intelligence tests are necessary and sufficient for the evaluation of machine intelligence. Finally, we close with some remarks about how human intelligence tests can still be useful for AI (and AI for psychometrics) and future directions for intelligence evaluation.

Tarek R. Besold
Institute of Cognitive Science, University of Osnabrück, Germany, E-mail: tbesold@uni-osnabrueck.de

José Hernández-Orallo
Dept. de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain, E-mail: jorallo@dsic.upv.es

Ute Schmid
Faculty Information Systems and Applied Computer Science, University of Bamberg, Germany, E-mail: ute.schmid@uni-bamberg.de

2 Conceptual foundations of strong AI

In AI as well as in cognitive science research, there are many projects which explicitly or implicitly aim at recreating human higher-level cognitive or intellectual capacities with computational means. Three conceptual notions are shared by most, if not all endeavours in

"strong" artificial intelligence and cognitive systems research, namely the concept of a computational theory of mind, the Church-Turing thesis, and the Physical Symbol System Hypothesis (PSSH).

The computational theory (Pylyshyn, 1980) foundationally bridges the gap between humans and computers by advocating that the human mind and brain can be seen as an information processing system and that reasoning and thinking correspond to processes that meet the technical definition of computation as formal symbol manipulation. The Church-Turing thesis (Turing, 1969) adds an account of the nature and limitations of the computational power of such a system by establishing Turing computability as valid characterisation of computability in general. The PSSH (Newell, 1980) operates on a different level, characterising the nature of the computations: taking the computational characteristic of cognition and intelligence as given, it proposes a general criterion for a system to display intelligence by stating that "[t]he necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system", with "general intelligent action" referring to rational behaviour which, in turn, can be understood as an agent's ability and determination to select a certain action if, given her goals, it is known that this precise action will lead to achieving the goal(s) (cf. Newell, 1982).

The implications of the PSSH are twofold. In its necessity, it states that also human thinking and higher-level cognition is a kind of symbol processing. In its sufficiency, it opens up the way to machine intelligence also on a paradigmatic level with regard to the type of computation. Whilst in the meantime also alternative readings have been proposed, in the classical account of the hypothesis, the symbols are physical objects representing things in the world, having a recognisable semiotic meaning or denotation but exhibiting arbitrary shapes unrelated to their meanings, and allowing for recursive composition with other symbols by rule, thus forming a combinatorial representation (Harnad, 1990). Arbitrariness of shape and combinatorial nature of the representation are thereby two key properties: the former allows symbols to designate anything at all by not prescribing to the symbol what expressions it can designate, but instead by leaving the determination of the designated object of a complex expression to the interaction between respective symbol tokens, whilst the latter characteristic accommodates for the generality of the intended intelligent action by establishing a representation language.

While proponents of strong AI and Artificial General Intelligence (AGI) (Kühnberger and Hitzler, 2009; Besold, 2013a) explicitly claim that the recreation of human-level intelligence on computer systems is possible, researchers in "standard" (weak) AI and Cognitive Modelling work on the weaker assumption that computability provides an appropriate conceptual apparatus for theories of the mind. That is, computational models can be used to simulate human information processes thereby providing detailed and consistent generative descriptions of different areas of cognition (Johnson-Laird, 1988). However, in both cases, the methodological challenge is to provide empirical evidence that the behaviour generated by a computer simulation is based on principles similar to human cognition. In other words, whenever we create an AI system or a computational cognitive model, we implicitly or explicitly propose that the computer program realises or models (aspects of) human-like intelligence.

3 Defining and measuring (human) intelligence

But what is meant by the term "intelligence"? In everyday life we use this term intuitively to evaluate our fellow human beings. Ascribing intelligence in this context typically means that this person is better than average in intellectually demanding areas such as mathematics, physics or chess. On the other hand, the term intelligence is used to contrast and compare human abilities with that of other species, typically animals. In the context of strong AI and AGI, intelligence usually is ascribed by observation or interviewing of a system, i.e. a kind of implicit or unsystematic conduction of the Turing test (Turing, 1950). In the context of Cognitive Modelling, performance parameters of computer models (such as number of iterations to solve a task) are compared to empirical data (such as solution times) gained from empirical studies with human subjects.

Systematic research on the conceptualisation and measurement of intelligence is the realm of psychometrics —a branch of psychology established at the beginning of the last century. Intelligence is defined as the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with the environment (Wechsler, 1944). In psychometrics, specific test batteries are designed to capture intelligence by assigning an intelligence quotient to a human based on his or her performance in a series of tasks. Typically, such IQ tests consist of sub-tests addressing different aspects of intelligence, such as visual-spatial, verbal-linguistic, and logical-mathematical abilities (Sternberg, 2000). Many researchers assume that all branches of intellectual activity are based on a common fundamental function (Spearman, 1904). Based on factor analytical models, there is some evidence that there is a

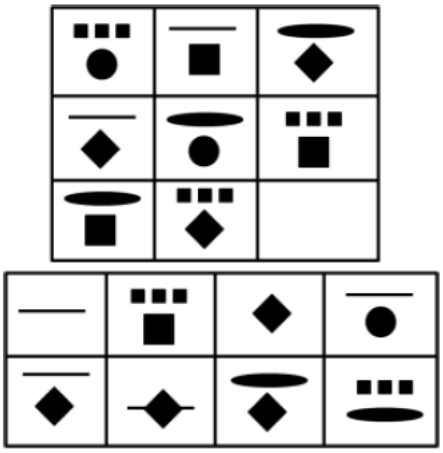

Fig. 1 A visual pattern problem similar to Raven Progressive Matrix problems (Lovett *et al.*, 2010).

general factor (the *g* factor) which determines performance in all more specific domains. However, there is high controversy about the so-called "theories of intelligence", each elaborating a different "taxonomic structure of abilities" (Sternberg, 2000), and its correspondence with theories of mind in psychology and cognitive science.

The most prominent example of a test measuring general intelligence is the Raven Progressive Matrix test —a non-verbal test where matrices of visual patterns are given and the pattern of the given regularities has to be identified (see Fig. 1). Inductive reasoning with numbers is another example for a test that is included in several IQ tests (Amthauer *et al.*, 1999). For example, given the series *3, 7, 15, 31, 63*, the pattern $2 \times f(n-1)+1$ can be identified and used to calculate the next number *127* (Hofmann *et al.*, 2014).

As mentioned above, theories of intelligence and tests have always dealt with two intertwined dimensions: the landscape of abilities (the *problems* that make up the tasks) and the kind of individuals (the *subject* populations). For instance, different sets of tasks (and not only different ranges of difficulty) are used for children and disabled people than for the rest of the normative adult population. How does this duality between tasks and subjects behave for machine intelligence? The discussion in section 2 is relevant to the consideration of the taxonomy of the sets of tasks (symbolic, computational, etc.) and a hierarchical taxonomy between sets of subjects in the order humans, animals and machines.

Given that psychometrics provides a well-established methodological approach to capture human intelligence, why not adopt a similar procedure to decide whether AI programs are intelligent? This idea could offer a further method to assess intelligence of artificial systems besides the Turing test approach and the performance comparison of humans and computer models.

4 Psychometrics and Psychometric AI

"Psychometric AI" (PAI) (Bringsjord and Schimanski, 2003) as a research program aims to apply the full battery of techniques from psychometrics to a strong AI context, setting its internal standard by declaring an agent as intelligent if and only if it does well in all established, validated tests of intelligence and mental ability, and subsequently setting out to use the results of the respective tests in a dedicated effort to build agents meeting the aforementioned criterion. PAI has by now become a line of research in its own right as, e.g., documented by the articles collected in Bringsjord (2011).

Still, the use of psychometrics in an evaluation-related AI context is wider. For example, Detterman (2011) has challenged the field of AI to assess an artificial system's performance by using classical IQ tests (but, unlike PAI, without the commitment to and focus on the engineering of agents adhering to the introduced standard). Detterman's call is issued to AI researchers working on approaching the level of intelligence exhibited by humans, challenging them to prove the validity of their claims by administering an IQ test (precompiled by experts in human psychometrics) to their computer systems and comparing the actual IQ scores — with the restriction that in the full version of the challenge only "a priori algorithms" are admissible, i.e., algorithms not previously specialised for a battery of tasks.

Over the last years, we can observe a growing number of publications presenting computer programs solving IQ test problems, mostly addressing tests with high loading of the *g*-factor such as Raven Progressive Matrices and number series problems. In standard AI research, IQ test problems have been identified as challenging application domains for algorithmic approaches to inference (Siebers and Schmid, 2012). In AGI, algorithms for solving IQ test problems are designed with the aim to surpass average human performance on these tests (Strannegård *et al.*, 2013a,b).

Some researchers want to demonstrate that computer programs which can solve IQ test problems definitely are not based on principles underlying human intelligent behaviour. A computer program can be handcrafted to perform well on a specific set of tasks, such as an IQ test (Sanghi and Dowe, 2003), instead of covering a wider scope of problems. On the other hand, there are some applications of general approaches from theorem proving (Burghardt, 2005) and inductive programming (Hofmann *et al.*, 2014) which were applied to number

series problems to demonstrate that the proposed algorithms are general enough to be applied to problems outside the original domain.

In Cognitive Science, researchers are interested in providing cognitive models simulating cognitive principles of pattern identification that are assumed to be underlying human general intelligence (Lovett *et al.*, 2010). Cognitive models for IQ test problems can relate the concept of item difficulty of psychometrics with the complexity of information processing. While item difficulty is defined by the percentage of subjects in a validation sample who solve this item, a cognitive model can give an explanation of item difficulty with respect to the necessary effort for an algorithmic solution.

These different perspectives on developing algorithmic approaches to solve IQ test problems have different underlying assumptions about how to determine whether a computer program is intelligent: even if AI uses IQ tests as a challenging area of application, there is no need to ascribe intelligence to such programs — evaluation can be done by comparing the performance of different algorithms. Nevertheless, many researchers may still have the implicit assumption that a program which performs better is more intelligent.

When AI algorithms based on general principles of inference are tested for applicability over different domains, it is harder to define a rank order over approaches. It is open to discussion which system is more powerful —one that outperforms others in some domains but has weak performance or is not even applicable in other domains or one with average performance over a broad variety of domains. One is tempted to use the analogy to humans which have one isolated exceptional skill but are intellectually impaired otherwise —so called idiot savants (Miller, 1999)— versus humans with average abilities over many areas. When a cognitive model is proposed, the crucial question is not whether this model is intelligent in itself but to what extent it mimics the cognitive processes performed by humans when solving such problems, that is, whether the model can provide an explanation of human behavior. How to determine similarity between processing features of a computer model and empirically observable behavior in humans is a challenging methodological problem in its own (Cooper *et al.*, 1996). If processing parameters of model and human performance are judged as sufficiently similar, the model is assumed to capture the core characteristics of human information processing. In consequence, it might be claimed that the model captures relevant aspects of human intelligence.

From all these approaches, only AGI models explicitly make the claim to aim at creating human-level (or even super-human) intelligence. In this context, PAI can be seen as an alternative approach to the previously more dominant observational assessments (like the Turing test).

5 Can human intelligence tests work for machines?

The validity of any discriminative test is mostly based on its necessity and sufficiency. In the case of intelligence, this is not different, and a similar argument has been applied to the Turing test and other proposals for measuring (artificial) intelligence. Namely, necessity means that if a system is intelligent then it must pass the test and sufficiency means that if a system passes the test then it must be intelligent.

In order to analyse whether human intelligence tests are valid for machines we will introduce six arguments (or characteristics of human intelligence tests) that can be used to explore whether they are necessary and sufficient, as summarised in Table 1. Many of these arguments rely on machines and humans (and non-human animals) being different *subject* populations and on the breadth and variety of *problems* in the test.

1. *Human intelligence tests are anthropocentric.* As humans are the quintessential example of intelligence, it makes sense to derive the concept and the measurement tools from them. Human tests have just been tuned to become necessary and sufficient for humans (or as much as possible). For instance, they evolved into culture-fair tests to solve problems about necessity. They also evolved to have a range of tasks to solve problems about sufficiency. So, this argument can be used to question whether human intelligence, as measured by IQ tests, is a particular (anthropocentric) type of intelligence, instead of a universal one. Also, even if it is human-level intelligence what we want to measure, it is not clear that the tests can work for other kinds of subjects. Hence, there can be important concerns about the necessity and sufficiency of these tests for non-human subjects, such as machines (and animals).

2. *Human intelligence tests are administered in a particular way.* Intelligence tests are practical for humans. A reliable measurement can be obtained with a short test. As a result, only AI systems that are specialised to the particular test interface and choice of symbolic representation can be evaluated, including those tests that require the understanding of language. In addition, many tests require the extrapolation of sequences with no feedback whatsoever about what is correct or not (e.g., just follow the sequence in a "natural" way). This happens even if the language or the milieu (e.g., tests for blind people) are adapted to the examinee. Consequently, the interface raises doubts about the necessity

#	Argument	Necessity	Sufficiency
1	Human intelligence tests are anthropocentric.	Effect	Effect
2	Human intelligence tests are administered in a particular way.	Effect	-
3	Human intelligence tests are specialised to kinds of subjects.	Effect	Effect
4	Human intelligence tests are usually normalised to a population.	Effect	Effect
5	Human intelligence tests are robust to training.	-	Effect
6	Human intelligence tests are composed of many different abilities and factors.	Effect	Effect

Table 1 Six arguments that can be used to determine the (in)adequacy of human intelligence tests for machines and whether they can have effect on the necessity and sufficiency of these tests.

of human intelligence tests for other kinds of subjects, such as machines (and animals, small children and disabled people). As a sign of this, tests for animals are administered with rewards, not instructions.

3. *Human intelligence tests are specialised to kinds of subjects.* The two previous arguments can be responded by the fact that there are specialised human intelligence tests for disabled people, for children of different ages, etc. It could also be argued that the same could be done for machines. Actually, Detterman suggests that a "unique battery of intelligence tests" (Detterman, 2011) could be designed on purpose for machines developed by "the editorial board of *Intelligence* and members of the International Society for Intelligence Research". However, which are the criteria for the inclusion in this 'unique' battery? And if some tests are developed anew, what would be the guidelines to specialise these tests for machines? Would they be still useful for humans? And what if the battery is finally passed by a machine? Would the commission be tempted to look for a different or more machine-unfriendly battery that humans can pass but the program cannot, à la CAPTCHA[1] (Von Ahn *et al.*, 2003)? In the end, any adaptation or specialised selection of tasks would raise questions about the necessity and sufficiency for machines, even more than a standard human intelligence test.

4. *Human intelligence tests are usually normalised to a population.* The result from a human intelligence test is just a number that can be compared to other numbers, in an ordinal, but not a quantitative way. In other words, human intelligence is measured at "the ordinal level", or is "weakly measurable" (Bartholomew, 2004, pp. 145). For instance, to make the number meaningful, IQ is normalised to have a mean of 100 and a standard deviation of 15. Clearly, this normalisation is not going to work for other populations, as it does not work for children or even some subgroups of the human population. However, we cannot re-normalise for machines, as the mere notion of an 'average' machine is ridiculous, because there is no normative population of machines. This means that even if the human normalisation is used, we have problems about sufficiency, i.e., it is unclear what it means if a machine scores an IQ of 20 or an IQ of 524, as these values are clearly anomalous for humans. Conversely, from the necessity point of view, what are machines far below or far above human intelligence expected to score?

5. *Human intelligence tests are robust to training.* Having a public and transparent measurement benchmark is always a good thing for science. Some human intelligence tests are well-known and public, but not all. Actually, some academic and professional psychological tests are never made public, because otherwise people could practice on them and game the evaluation. Even if care is taken not to disclose or repeat exercises to avoid rote learning, humans can prepare for the *kinds* of tests. Fortunately, the improvement of training, even if significant, is limited (Bors and Vigneau, 2001), although the reasons and the permanent effect beyond the particular test are not very well understood. Aware of this, Detterman talks about two levels in his challenge, the second where tests are not disclosed previously to the evaluation (so programmers cannot implement specialised approaches). However, it is not clear that this robustness to humans trying to game the evaluation by systematic practising should also hold for machines. As already discussed in the previous section, specialised systems can be built whose only purpose is to score well in intelligence tests (e.g., Sanghi and Dowe, 2003), without being able to do well in other tasks. This raises strong doubts on sufficiency.

6. *Human intelligence tests are composed of many different abilities and factors.* There are broad test batteries, including many abilities and skills. This is good, but how far should we go? For instance, PAI states that machines should be evaluated with "all established, validated tests of intelligence and mental ability, a class of tests that includes not just the rather restrictive IQ tests, but also [...] tests of artistic and literary creativity, mechanical ability, and so on" (Bringsjord, 2011). However, some humans score poorly on some of these tests (e.g., Stephen Hawking). Conversely, we clearly see that some tasks that are discriminative for humans (e.g., arithmetic or even reaction time) are meaningless for machines. Even if we agree that results for different tests should be analysed separately, it is not clear how

[1] CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) are tests (e.g., distorted letters) to detect bots in Internet applications.

the relations between abilities and factors could be interpreted for machines, or how different tasks could be weighted for an aggregated value. For instance, if the *g* factor has been found in human evaluation, should it appear for machines as well? More blatantly, should we expect well-established tests of choice reaction time to be correlated with intelligence in machines as they are correlated in humans (Deary *et al.*, 2001)? Most of the research in human intelligence is then, in principle, not extrapolatable to machines (not even to animals), raising delicate issues about the breadth, composition and interpretation of separate results of any test battery for machines. Actually, it is unclear whether we can come up with a right set of problems by making the battery larger or smaller such that they are sufficient and necessary for machines.

From the above analysis of arguments, there seem to be serious doubts about the necessity and sufficiency of *current* human intelligence tests for the evaluation of machines. This is consistent with other previous papers (Dowe and Hernández-Orallo, 2012; Besold, 2014).

6 Conclusions

In the previous sections we have discussed whether human intelligence tests are valid for the evaluation of machines. As a result, we have found several issues about their sufficiency and necessity. Even if we argue that intelligence tests are not valid, in principle, for the evaluation of machines, it is important to highlight that this does not mean that human intelligence tests are useless for artificial intelligence and cognitive science. More on the contrary, it is precisely the analysis of what is lacking in our AI systems to score well in a range of human intelligence tests and what is lacking in human intelligence tests to properly discard those systems that are not intelligent what can give us insight about the nature of human intelligence tests and also about the progress of AI *if the tests are generalised*. In other words, the use of human intelligence tests in AI research is useful, provided we are cautious about the semantic and quantitative interpretation of results.

The PAI methodology for AI, however, does not advocate for any generalisation or improvement of psychometric tests. This can actually be one of the most useful outcomes of this process, by the development of brand-new tests based on (algorithmic) information theory (Hernández-Orallo, 2000; Legg and Hutter, 2007) or on models about cognition (Mueller *et al.*, 2007), and their hybridisation with (cognitive) generalisations of the Turing Test (Besold, 2013b) or generalisations of psychometrics (Hernández-Orallo *et al.*, 2014), where tests are devised for any kind of cognitive system independent of its type or nature (individual or collective, artificial, biological, or hybrid). This suggests that an interesting way of looking at the question is by reversing it: *can human intelligence be measured as a very special case of machine intelligence?*

References

Amthauer, R., Brocke, B., Liepmann, D., and Beauducel, A. (1999). *Intelligenz-Struktur-Test 2000 (I-S-T 2000)*. Hogrefe, Göttingen.

Bartholomew, D. J. (2004). *Measuring Intelligence: Facts and Fallacies*. Cambridge University Press.

Besold, T. R. (2013a). Human-level artificial intelligence must be a science. In K.-U. Kühnberger, S. Rudolph, and P. Wang, editors, *Artificial General Intelligence*, volume 7999 of *LNCS*, pages 174–177. Springer Berlin Heidelberg.

Besold, T. R. (2013b). Turing revisited: A cognitively-inspired decomposition. In V. C. Müller, editor, *Philosophy and Theory of Artificial Intelligence, SAPERE 5*, pages 121–132. Springer.

Besold, T. R. (2014). A note on chances and limitations of Psychometric AI. In *KI 2014: Advances in Artificial Intelligence*, pages 49–54. Springer.

Bors, D. A. and Vigneau, F. (2001). The effect of practice on raven's advanced progressive matrices. *Learning and Individual Differences*, **13**(4), 291–312.

Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, **23**(3), 271–277.

Bringsjord, S. and Schimanski, B. (2003). What is Artificial Intelligence? Psychometric AI as an Answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pages 887–893. Morgan Kaufmann.

Burghardt, J. (2005). E-generalization using grammars. *Artificial Intelligence*, **165**, 1–35.

Cooper, R., Fox, J., Farringdon, J., and Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, **83**, 3–44.

Deary, I. J., Der, G., and Ford, G. (2001). Reaction times and intelligence differences: A population-based cohort study. *Intelligence*, **29**(5), 389–399.

Detterman, D. (2011). A challenge to Watson. *Intelligence*, **39**(2-3), 77–78.

Dowe, D. L. and Hernández-Orallo, J. (2012). IQ tests are not for machines, yet. *Intelligence*, **40**(2), 77–81.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, **42**, 335–346.

Hernández-Orallo, J. (2000). Beyond the Turing Test. *J. Logic, Language & Information*, **9**(4), 447–466.

Hernández-Orallo, J., Dowe, D. L., and Hernández-Lloreda, M. V. (2014). Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research*, **27**(0), 50–74.

Hofmann, J., Kitzelmann, E., and Schmid, U. (2014). Applying inductive program synthesis to induction of number series a case study with IGOR2. In *KI 2014: Advances in Artificial Intelligence*, pages 25–36. Springer.

Johnson-Laird, P. N. (1988). *The Computer and the Mind: An Introduction to Cognitive Science*. Fontana Press, London.

Kühnberger, K. U. and Hitzler, P. (2009). Facets of artificial general intelligence. *KI*, **23**(2), 58–59.

Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, **17**(4), 391–444.

Lovett, A., Forbus, K., and Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. In *Proceedings of CogSci-10*, pages 2761–2766.

Miller, M. (1999). The savant syndrome: intellectual impairment and exceptional skill. *Psychological Bulletin*, **125**(1), 31–46.

Mueller, S. T., Jones, M., Minnery, B. S., and Hiland, J. M. H. (2007). The BICA cognitive decathlon: A test suite for biologically-inspired cognitive agents. In *Proceedings of Behavior Representation in Modeling and Simulation Conference, Norfolk*.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, **4**, 135–183.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, **18**, 87–127.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundation of cognitive science. *The Behavioral and Brain Sciences*, **3**, 111–132.

Sanghi, P. and Dowe, D. L. (2003). A computer program capable of passing I.Q. tests. In P. P. Slezak, editor, *Proceedings of ICCS/ASCS-2003*, pages 570–575, Sydney, AU.

Siebers, M. and Schmid, U. (2012). Semi-analytic natural number series induction. In *KI 2012: Advances in Artificial Intelligence*, pages 249–252. Springer.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, **15**, 201293.

Sternberg, R. J., editor (2000). *Handbook of Intelligence*. Cambridge University Press.

Strannegård, C., Amirghasemi, M., and Ulfsbäcker, S. (2013a). An anthropomorphic method for number sequence problems. *Cognitive Systems Research*, **22-23**, 27–34.

Strannegård, C., Cirillo, S., and Ström, V. (2013b). An anthropomorphic method for progressive matrix problems. *Cognitive Systems Research*, **22-23**, 35–46.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**, 433–460.

Turing, A. M. (1969). Intelligent machinery. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 5, pages 3–23. Edinburgh University Press.

Von Ahn, L., Blum, M., Hopper, N. J., and Langford, J. (2003). CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology - EUROCRYPT 2003*, pages 294–311. Springer.

Wechsler, D. (1944). *The measurement of adult intelligence*. Williams & Wilkins, Baltimore.

Tarek R. Besold is a member of the AI Research Group at the Institute of Cognitive Science of the University of Osnabrück (Germany). He studied mathematics and computer science at the University of Erlangen-Nuremberg (Germany) and at the University of Zaragoza (Spain), and logic at the University of Amsterdam (The Netherlands). In 2014 he was an academic visitor at the Centre for Intelligent Systems and their Applications (CISA) of the University of Edinburgh (Scotland, UK). Main fields of research and scientific interest are: Analogies and analogical reasoning, rationality and creativity, human-level AI, and neural-symbolic integration.

José Hernández-Orallo is a reader in the Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València (UPV, Spain). He holds a M.Sc. in Computer Science from UPV, partly completed at the École Nationale Supérieure de l'Électronique et de ses Applications (France), and a Ph.D. in Logic from Universitat de València. His main research areas are in machine learning, artificial intelligence, data mining, machine evaluation and inductive programming.

Ute Schmid is professor of Cognitive Systems at the Faculty Information Systems and Applied Computer Science of the University of Bamberg (Germany). She received a diploma degree in Psychology as well as a diploma degree in Computer Science. She holds a doctoral and a habilitation degree in computer science, both from Technical University Berlin. During her doctoral and postdoctoral time she worked as assistant at TU Berlin, as research stipend at Carnegie-Mellon University, and

as lecturer at the University of Osnabrück. Her main research interest are approaches to learning on the knowledge-level and application of methods of cognitive AI to various domains.