Artificial Intelligence 230 (2016) 74–107

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Computer models solving intelligence test problems: Progress and implications

José Hernández-Orallo [a], Fernando Martínez-Plumed [a,*], Ute Schmid [b], Michael Siebers [b], David L. Dowe [c]

[a] Dept. of Computer Systems and Computation, Universitat Politècnica de València, 46022 Valencia, Spain
[b] Faculty of Information Systems and Applied Computer Science, University of Bamberg, 96045 Bamberg, Germany
[c] Faculty of Information Technology, Monash University, Clayton, Vic. 3800, Australia

A R T I C L E I N F O

Article history:
Received 25 March 2014
Received in revised form 30 March 2015
Accepted 27 September 2015
Available online xxxx

Keywords:
Intelligence tests
Cognitive models
Artificial intelligence
Intelligence evaluation

A B S T R A C T

While some computational models of intelligence test problems were proposed throughout the second half of the XXth century, in the first years of the XXIst century we have seen an increasing number of computer systems being able to score well on particular intelligence test tasks. However, despite this increasing trend there has been no general account of all these works in terms of how they relate to each other and what their real achievements are. Also, there is poor understanding about what intelligence tests measure in machines, whether they are useful to evaluate AI systems, whether they are really challenging problems, and whether they are useful to understand (human) intelligence. In this paper, we provide some insight on these issues, in the form of nine specific questions, by giving a comprehensive account of about thirty computer models, from the 1960s to nowadays, and their relationships, focussing on the range of intelligence test tasks they address, the purpose of the models, how general or specialised these models are, the AI techniques they use in each case, their comparison with human performance, and their evaluation of item difficulty. As a conclusion, these tests and the computer models attempting them show that AI is still lacking general techniques to deal with a variety of problems at the same time. Nonetheless, a renewed attention on these problems and a more careful understanding of what intelligence tests offer for AI may help build new bridges between psychometrics, cognitive science, and AI; and may motivate new kinds of problem repositories.

1. Introduction

Artificial intelligence (AI) is typically defined as “the scientific understanding of the mechanisms underlying thought and intelligent behaviour and their embodiment in machines”.[1] Associated with the notion of natural and artificial intelligent agents is our (human-centred or anthropocentric) belief that intelligence underlies most human behaviour. The origin of

* Corresponding author.
E-mail addresses: jorallo@dsic.upv.es (J. Hernández-Orallo), fmartinez@dsic.upv.es (F. Martínez-Plumed), ute.schmid@uni-bamberg.de (U. Schmid), michael.siebers@uni-bamberg.de (M. Siebers), david.dowe@monash.edu (D.L. Dowe).

[1] See homepage of the Association for the Advancement of Artificial Intelligence, http://www.aaai.org.

http://dx.doi.org/10.1016/j.artint.2015.09.011

AI research was based on the conviction that "intelligence is the computational part of the ability to achieve goals in the world" [1].

Indeed, AI research can claim some impressive milestones. For example, already in 1959 Arthur Samuel presented a self-learning program that could play checkers (or draughts) [2]. In 1997 the 1957 prophecy of Herbert Simon that within 10 years a computer would be world's chess champion (eventually) came true when Deep Blue won against the human chess champion Garry Kasparov [3]. In 2011 IBM's program Watson [4,5] was the winner of the *Jeopardy!* TV quiz. However, one can ask whether the mechanism underlying the behaviour of these programs is the same as or similar to the mechanism underlying human intelligent behaviour. In fact, this success in specialised tasks is a very illustrative demonstration of the *big switch* approach in AI research. If we fix the switch to one particular problem, we can devise, after several years or decades of research, a system that performs better than humans. We can even embed many specialised programs into a system and devise an automated switch. For instance, if we have a good program for checkers, a good program for chess, etc., we can devise a meta-system that is able to recognise which kind of game has to be played and switch to the appropriate program. If AI evaluation is based on specific benchmarks that are known beforehand, non-intelligent systems will be able to thrive. Game playing is a good example where a reaction against this specialisation is beginning to flourish. Since 2005 the performance of game playing systems is evaluated in the game playing competition [6] on a wide variety of games—some invented ones not disclosed to the participants until the competition. Consequently, in the area of game playing, systems using a big switch approach are hardly successful in contrast to systems realising general game playing algorithms.

While successfully playing games can be seen as a special manifestation of intelligent behaviour, intelligence tests assess the underlying ability to act intelligently in many different domains [7]. In psychology research the classical approach to intelligence assessment is to apply psychometric tests measuring intelligence [7]. Some of these tests, the so-called IQ tests, are standardised in such a way that humans can be classified as below, about, or above average intelligence. Nonetheless, there are many other human intelligence tests and cognitive tests that also measure intelligence. In addition, other similar tests from other areas were not originally targeted to humans. In what follows, for simplicity, we will use the term *intelligence tests* for all of them. The intelligence test tasks address a variety of reasoning abilities, for example, solving number series problems, detecting regularities in spatial configurations, or understanding verbal analogies. Some types of problems are rather independent of the subject's educational and cultural background, others depend on background knowledge.

In early AI research, the intelligence test approach was considered as a useful approach for AI programs as well; Newell argued that one of the ways artificial intelligence could be achieved was "to construct a single program that would take a standard intelligence test, say the WAIS or the Stanford–Binet" [8]. And indeed, as early as 1963, Evans devised an AI program that could solve geometric analogy tasks from the WAIS (Wechsler Adult Intelligence Scale) test [9,10] and, in the same year, Simon and Kotovsky [11] presented a program that could solve Thurstone letter series completion problems [12]. Both types of problems address the ability to identify regularities in patterns and generalise over them. This connection between inductive inference and intelligence tests was also identified early on, for instance by Blum and Blum [13]: "Intelligence tests occasionally require the extrapolation of an effective sequence".

After the initial interest of AI research in intelligence test problems, this branch of research sank into oblivion during the next decades. However, in the 1990s, cognitive science research recovered this line of research, and cognitive models were proposed to simulate the human cognitive processes that take place when solving inductive inference in intelligence test problems [14].

In AI, forty years after the work of Evans, Simon and Kotovsky, in 2003, computer programs solving intelligence tests became of interest again. On one hand, Sanghi and Dowe [15] wanted to make a conclusive point about how easy it was to make non-intelligent machines pass intelligence tests. On the other hand, Bringsjord and Schimanski aimed at resuscitating the role of psychometric tests—including not only intelligence tests but also tests about personality, artistic creativity, etc.—in AI [16]. They claimed that psychometric tests should not be dismissed but placed at a definitional, major role for what artificial intelligence is and proposed "psychometric artificial intelligence" (PAI) as a direction of research. While this approach moves towards an ability-oriented categorisation of problems instead of the classical task-oriented categorisation, it is not clear whether it is free from the big switch approach, especially if the kinds of tasks that appear in intelligence tests are known beforehand. In fact, Sanghi and Dowe's system used a big switch approach. Given these two opposite stances on the use of intelligence tests, it is hard to tell yet how influential they have been or will be.

But the fact is that the past ten (and especially five) years have been blooming with computational models aimed at solving intelligence test problems. The diversity of goals and approaches has also widened, including the use of intelligence tests for the analysis of what intelligence is, for the understanding of certain aspects of human cognition, for the evaluation of some AI techniques or systems, including robots, and, simply, to have more insights about what intelligence tests really represent. We will use the term *computer model* for all these approaches, independently of their purpose, of the employed techniques, and of the range of problems they are able to address.

In the current state of research, there exist many systems addressing different intelligence tests. Currently, there is no general framework to characterise all intelligence tests. Therefore, it is unclear whether instances of problems are members of the same or different problem classes. With the exception of *big switch* approaches, like the one of Sanghi and Dowe [15], current systems are based on algorithmic approaches specifically designed to solve a special class of such problems and even only one specific test. It is an open question whether a general algorithmic approach for a diverse set of intelligence tests can be designed in principle. Comparing the current situation of computer models which solve intelligence tests with the