City Research Online

City, University of London Institutional Repository

Citation: Besold, T. R. (2014). A note on chances and limitations of psychometric AI. Lecture Notes in Computer Science, 8736, pp. 49-54. doi: 10.1007/978-3-319-11206-0_5

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/18673/

Link to published version: https://doi.org/10.1007/978-3-319-11206-0_5

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

A Note on Chances and Limitations of Psychometric AI

Tarek R. Besold

Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany
tbesold@uni-osnabrueck.de

Abstract. Human-level artificial intelligence (HAI) surely is a special research endeavor in more than one way: In the first place, the very nature of intelligence is not entirely clear; there are no criteria commonly agreed upon necessary or sufficient for the ascription of intelligence other than similarity to human performance (and even this criterion is open for a plethora of possible interpretations); there is a lack of clarity concerning how to properly investigate HAI and how to proceed after the very first steps of implementing an HAI system; etc. In this note I assess the ways in which the approach of Psychometric Artificial Intelligence [1] can (and cannot) be taken as a foundation for a scientific approach to HAI.

1 Introduction

From a certain perspective AI seems to stand out between the modern sciences for more than one reason: Neither is there agreement upon what shall be AI's overall objective (i.e., whether the purpose of AI is the implementation of technical systems supporting humans in their everyday tasks and facilitating human intellectual activity, or if the purpose of AI is the creation of a computer system exhibiting general intelligence — in doing so possibly outperforming humans in tasks requiring reasoning and thought —, or something in between these two extremes), nor is there a commonly accepted methodology for conducting research in AI, nor is there consensus concerning the valuation of previous developments and of the actual status quo in AI as a story of success or perpetual failure.

These and related observations repeatedly caused philosophers of science and even some researchers from within AI to wonder about AI being a special type of science, or to even question (and occasionally finally deny) the status of AI as a science. In this note, specifically focussing on the subbranch of AI dealing with human-level AI (HAI), I want to undertake a critical review of Psychometric AI (PAI) [1] which has been proposed as a strictly scientific approach and conceptual framework suitable for measuring, evaluating, and guiding progress during the development of an HAI system.

2 Introducing Psychometric AI

Psychometric AI [1] aims to apply the full battery of techniques from psychometrics to an HAI 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.[1] PAI as a dedicated

[1] This definition of PAI actually only is approximate and partially incomplete. For the actual detailed phrasing see the following section introducing and discussing PAI in detail.

research program was first explicitly institutionalized a decade ago in [1] and has been actively worked on since (as, e.g., documented by the articles collected in [2]).

2.1 Psychometrics and Intelligence

Psychometrics as a field is by definition concerned with analyzing and developing means of psychological measurement as, for instance, personality assessments or questionnaires for measuring certain abilities or knowledge. For doing so, psychometrics engages in the study of theoretical approaches to measurement as well as in the active development and implementation of the corresponding concrete instruments and procedures. For a long time special interest has been taken in the measurement of the phenomena commonly subsumed under the term "intelligence": There is a wide variety of psychometric tests of intelligence, ranging from tests with only one type of item to varied batteries of different questions, combining verbal and non-verbal items and requiring subjects to perform qualitatively very different tasks (e.g., spatial tasks as opposed to language-related tasks).

2.2 Psychometrics and Artificial Intelligence

The goal of PAI now is to carry over the quantitative assessment of intelligence from the realm of classical psychometrics into artificial intelligence. In his programmatic papers [1, 3, 4], Bringsjord introduces PAI as a research program and discusses various objections critics might have.

Naive Psychometric AI: The first attempt at defining what it means for an agent (human or artificial) to be intelligent, given in [1] and repeated in the later publications, reads as follows: "*Some agent is intelligent if and only if it excels at all established, validated tests of intelligence*."

So in this account, which I will refer to as "naive PAI" (nPAI) in the following, AI in its entirety is reduced to a strictly psychometric core, namely the construction of a system which outperforms most humans in all currently available intelligence tests. Clearly, as also noticed by Bringsjord, this definition is overly narrow and insufficient for the purpose of building an HAI. Even the most advanced and broad battery of items has to be considered as too narrow when compared to the full range of cognitive capacities seen in humans. And also from a system engineer's perspective, given that at any point in time there will only be a finite number of commonly accepted tests of intelligence available, the maxim underlying nPAI seems dubious: Each of the individual capacities could be addressed by a specifically dedicated standalone subsystem, so that intelligence would actually be reduced to correctly selecting and executing the respective module from a finite number of available subprograms based on a finite number of possible input categories. But this could hardly be considered satisfactory as an answer to the intelligence puzzle for anyone but diehard followers of Descartes: At least the Cartesian would be comforted in that the resulting AI — being an almost ideal instantiation of a "type-c machine" [5] — might be able to pass any test for a particular set of cognitive capacities, whilst still failing a test for any mental power whatsoever.

General Psychometric AI: A second, more commonly used definition of PAI, to which I will henceforth refer as "general PAI" (gPAI), is also introduced in [1]: "*Psychometric*

AI is the field devoted to building information-processing entities capable of at least solid performance on 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."

Clearly, gPAI intuitively feels a lot like the classical Turing Test [6]. And this impression is not unfounded: Having a look at Harnad's Total Turing Test (TTT; [7]) as modernized and (possibly) broadened version of Turing's original proposal, we notice that — due to the simple mass of possible test methods, presentation modalities and required modes of interaction — a solution to gPAI would most likely also serve as an important step towards solid performance on the TTT. Conversely, a system convincingly solving the TTT most likely would also encompass the ability to solve gPAI.

So instead of focusing exclusively on the classical psychometric battery of tests for intelligence, gPAI widens its focus to also encompass all other available tests for cognitive capacities whatsoever. As there are now at least two possible reading to this claim. In order to avoid misunderstandings, let us clarify what in all probability is the intended meaning of this statement: Although the definition could be read as universally quantified statement in a strong sense, demanding from a system trying to accomplish gPAI to solve all possible tests for all conceivable mental capacities which might be validated at some point, this does not seem meaningful as a standard for evaluation.[2] In light of these considerations what is meant by the description of gPAI quoted above is the requirement for an AI system to pass, at a given point in time, all available validated psychometric tests for any kind of mental ability. So although the number of possible tests would still be enormous, gPAI would in this reading (contrary to the first possible interpretation) nonetheless be dealing with a finite and well-defined set of tests; the significant difference to nPAI resulting from opening up the scope of the tests from a strict focus on tests of intelligence to also include other mental abilities in general.

3 (Dis)Advantages of Psychometric AI for Guiding HAI

Let us return to the initial considerations concerning the use of PAI as means of guidance and assessment for the development of human-level artificial intelligence.

3.1 The Beauty of Numbers: What PAI Can Do

The great advantage of an approach such as PAI over, for instance, the (Total) Turing Test is the quantitative nature of the used evaluation method. Psychometric methods and tests are metric by their very nature, they are designed to provide assessments in terms of numbers on a scale. This clearly has already two advantages by itself: On the one hand, if psychometrics is applied in the evaluation of an artificial agent the result is a number which is at first glance directly comparable to the results of other systems on the same test(s). There is no question of how to compare outcomes of different evaluation

[2] Especially not since gPAI claims superiority over the Turing Test/Total Turing Test amongst others also for being — from an engineering perspective — in its goals and conditions of success less elusive and for almost automatically enforcing a more feasible and manageable approach by gPAI's test-based nature (see [1] for details).

runs, the standardized nature of the testing technique is meant to take care of these and similar issues. On the other hand, this also allows for a fairly unquestionable measure of progress, both due to the applied evaluation method as well as due to the overall goal of the PAI approach. A higher score on a particular battery of tests simply and straightforwardly indicates that the examined agent has advanced towards the target of performing decently on the respective tests. And, in turn, the term "decently" is (at least for HAI) also clearly defined by, for example, the average outcome human test subjects achieve on the tests under consideration; so even problems surrounding optimality criteria and/or the choice of an appropriate normative dimension seem to be addressable by the framework.

These advantages should by no means be underappreciated. Research in HAI has in many ways suffered from a lack of quantitative assessment methods for its systems: Neither could researchers easily compare the performance of their different agents and AIs, nor could they themselves know whether and how well they were advancing in the outcomes of their work. In summary, PAI does offer several very pleasant properties which many other paradigms lack and, I believe, can with a clear conscience be recommended for consideration as a solution to everyone whose main concern lies with having a research program with completely transparent and hard-to-question means of evaluation. Still, as will be discussed in the following subsection, this comes at a price which casts serious doubt on the applicability of PAI in HAI, and even the satisfiability of PAI's own goals by means of PAI.

3.2 A (Possibly) Fatal Flaw: What PAI Cannot Do

HAI has been characterized as the endeavor to create computer systems that exhibit intelligence at a level similar to humans. This by itself is ambiguous in many ways as neither "exhibit" nor "intelligence" are well-defined in the given context. Against this background, PAI[3] now offers clear specifications for both terms, equating a system's exhibition of intelligence with its performance on a huge variety of psychometric tests for different mental capacities — the act of solving the tests becomes the active demonstration, the resulting scores define the (level of) intelligence. But it is here where a quite troublesome conceptual flaw in the PAI framework comes to light.

[1] justifies the reliance on psychometrics as crucially defining criterion of an agent's intelligence as follows: "*What's intelligence? (...) [M]ost thinkers seem to forget that there is a particularly clear and straightforward answer available, courtesy of the field that has sought to operationalize the concept in question; that field is psychometrics. Psychometrics is devoted to systematically measuring psychological properties, usually via tests. These properties include the one most important in the present context: intelligence.*"

But this reading of psychometrics and the relation between its tests and the respective objects of investigation unfortunately turns out to be overly simplistic. Psychometric tests do not directly measure any mental ability as, for example, intelligence or

[3] As already discussed above, nPAI (when seen in an HAI context) clearly seems to be overly simplistic in many ways. As, moreover, it is trivially subsumed under the notion of gPAI I will in the following argument without loss of generality only refer to the latter notion.

creativity. There even is no unanimously agreed upon definition of what exactly intelligence is as a mental faculty and what its defining characteristics would be; and the same holds for many other high-level cognitive capacities. Psychometric measures are correlational measures, measuring the performance of traits which are commonly associated (and assumed to be strongly positively correlated) with what is considered intelligence — and even these correlations are in part justified merely by plausibility arguments and general agreement within the respective research community.

Therefore, what PAI actually trains its systems on (and measures them against) is not a standard for human-like intelligence of what form soever, but are benchmarks for better defined, more or less clearcut cognitive capacities which are quite plausibly intimately related and/or part of what is considered human intelligence. But even when taking the entirety of available psychometric tests available at a specific point in time (as proposed by gPAI) and having a system succeed on them, we still would only be dealing with placeholders and parts without any guarantee at all that an immediate measure of "human intelligence" would implicitly or explicitly have emerged from within the collection of different tests (leaving . In a way similar to the case of a person suffering from the savant syndrome [8], when an AI should exceed human performance on all available psychometric means of assessment of intelligence or other high-level mental abilities, all we would know for sure is that the system performs better on the tested correlated tasks — but nothing would be revealed about whether the system is "truly intelligent" in the general way typically developing humans are deemed to be.

Clearly, a strict behaviorist would disagree with this conclusion as at least from the exclusively behavior-oriented point of view human and machine would be indistinguishable in their performance on the applied psychometric tests. But at the same time, even from this perspective, it could not be excluded that there might still be the possibility of some aspect of intelligence that simply had not yet been accounted for in the available battery of measures — so even under a behaviorist angle the verdict that solving PAI guarantees that human-level intelligence has been achieved by an artificial intelligence would require a considerable leap of faith.

4 Comparison to Related Work and Conclusion

Trying to establish an adequate standard for ascertaining the level of human-like intelligence an artificial agent has achieved goes back to the very beginning of AI research — one might even say that the field started out from this question in [6]. Thus it should not come by surprise that by now there is a remarkable variety of tests on the market, two of them being the above mentioned original Turing Test and Harnad's expansion of it, the Total Turing Test. As already discussed, Total Turing Test and gPAI seem to share a close relationship in terms of their generality and requirements. Unfortunately, this closeness also makes them share a weakness usually brought forth against the Turing Test (and thus even more applicable to the more general TTT): Amongst others, [9, 10] remark that the Turing Test might be effectively useless as a test of machine intelligence as for passing it something similar to a "*human subcognitive substrate*" [9] would be needed. For a task like the Turing Test this seems fairly straightforward, as the machine would have to make sure that it can reflect all the behaviors produced by low- and mid-

level cognitive structures in humans. And to a lesser degree, this also holds for PAI: Amongst the battery of eligible psychometric tests there also are items which assess not only high-level capacities, but which move down in the cognitive hierarchy.

Still, PAI indeed has at least one advantage over the Turing Test, the TTT, and most (if not all) other versions and variants thereof: Contrary to them, due to its reliance exclusively on quantitative psychometrics tests, PAI offers a well-defined goal and the option of quantifying progress towards meeting it, getting rid of the seemingly unavoidable vagueness and ambiguity in the evaluation of a system's performance. From a purely engineering-oriented perspective, this argument should not be dismissed too easily — it should just always be kept in mind at which price this clarity is obtained.

Going back to the original undertaking of this note, trying to address the question whether PAI can serve as a guiding force for HAI, tying it even closer to standard scientific procedures, the answer as so often is not a simple "Yes." or "No.", but rather a "Up to a certain point.". On the one hand, it surely will have to be the case that an actual HAI system eventually must be able to perform well on psychometric tests, so using a wide variety of the latter as means of quantitatively assessing progress in HAI has to be seen as beneficial. On the other hand, it cannot be presumed that excelling on all available validated psychometrical measures at any point in time will guarantee that an artificial system has reached human-level intelligence. In summary, PAI unfortunately still will not serve as a final methodological answer deciding the question for human-level intelligence in artificial systems and, thus, falls fundamentally short of its own promises.

References

1. Bringsjord, S., Schimanski, B.: What is Artificial Intelligence? Psychometric AI as an Answer. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03), Morgan Kaufmann (2003)
2. Bringsjord, S.: Psychometric artificial intelligence. Volume 23(3) of Journal of Experimental & Theoretical Artificial Intelligence (JETAI). (2011)
3. Bringsjord, S., Schimanski, B.: "Pulling it All Together" via Psychometric AI. In: AAAI Technical Report of the AAAI Fall 2004 Symposium on Achieving Human-Level Intelligence Through Integrated Systems and Research, The AAAI Press (2004) 9–17
4. Bringsjord, S.: Psychometric artificial intelligence. Journal of Experimental & Theoretical Artificial Intelligence **23**(3) (2011) 271–277
5. Wheeler, M.: God's Machines: Descartes on the Mechanization of Mind. In Husbands, P., Holland, O., Wheeler, M., eds.: The Mechanical Mind in History. MIT Press, Cambridge, MA (2008) 307–330
6. Turing, A.: Computing Machinery and Intelligence. Mind **LIX**(236) (October 1950) 433–460
7. Harnad, S.: Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. Minds and Machines **1** (1991) 43–54
8. Treffert, D.A.: The savant syndrome: An extraordinary condition. A synopsis: Past, present, future. Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1522) (2009) 1352–1357
9. French, R.: Subcognition and the Limits of the Turing Test. Mind **99** (1990) 53–565
10. Cullen, J.: Imitation Versus Communication: Testing for Human-Like Intelligence. Minds and Machines **19** (2009) 237–254