How do you test the strength of AI?

Nikolay Mikhaylovskiy[1, 2][0000-0001-5660-0601]

[1] Higher IT School of Tomsk State University, Lenin Ave, 36, Tomsk, Tomsk Oblast, 634050, Russia
[2] NTR Lab, 12 Proyezd Maryinoy Roshchi, Bldg. 9, Block 1, Second Floor, Moscow, 127521, Russia

nikolaj.mihajlovskij@hits.tsu.ru, nickm@ntr.ai

Abstract. Creating Strong AI means to develop artificial intelligence to the point where the machine's intellectual capability is in a way equal to a human's. Science is definitely one of the summits of human intelligence, the other being the art. Scientific research consists in creating hypotheses that are limited applicability models (methods) of compressing information. In this article, we show that this paradigm is not unique to the science and is common to the most developed areas of human activities, like business and engineering. Thus, we argue, a Strong AI should possess a capability to build such models. Still, the known tests to confirm the human-level AI do not address this consideration. Based on the above we suggest a series of six tests of rising complexity to check if AI have achieved the human-level intelligence (Explanation, Problem-setting, Refutation, New phenomenon prediction, Business creation, Theory creation), five of which are new to the AGI literature.

Keywords: AGI, Strong AI, Epistemology, Turing test.

Creating Strong AI means to develop artificial intelligence to the point where the machine's intellectual capability is in a way equal to a human's or, as Ray Kurzweil [1] put it, machine intelligence with the full range of human intelligence.

A number of cognitive architectures have emerged over time as a result of research on Strong AI. While most of them will never evolve into Strong AI, it is important to have a common ground to judge where do they stand against that goal.

Additionally, I agree with Arthur Franz [2] that Strong AI will be subject to evolution, including both shallow (personal for a single individual) and deep (inherited through reproduction). Thus, even within a single research program or architecture it is important to understand progress towards Strong AI.

In this paper, we explore the question of how to determine if the Strong AI have been achieved. Specifically, scientific activity is the summit of human intelligence and scientific research consists in creating hypotheses that are limited applicability models (methods) of compressing information. In this article, we show that this paradigm is not unique to the science and is common to the most developed areas of human activities, like business and engineering. Thus, we argue, a Strong AI should possess a capability to build such models. Still, the known tests to confirm the human-level AI do not address this consideration. We aim to fill this gap.

To that end, in Chapter 1 we explore existing tests for strong AI. In Chapter 2, we stripe the applicable notion of scientific knowledge from the modern epistemology and define the key features of the scientific knowledge. In Chapter 3, we show that many other human activities, most notably business, engineering and contemporary marketing rely on the similar knowledge structures. Finally, in Chapter 4 we device the tests for the Strong AI in the sense we have previously defined.

Throughout this paper, we use the terms “Strong AI” and “Artificial General Intelligence” (AGI), interchangingly, despite the ongoing terminological discussion within the AI community if these are the same or different notions.

1 Tests for AI

A number of tests have been devised to test if a system has an artificial intelligence. Some of them include:

- The Turing test, suggested by Alan Turing
- Lovelace Test (suggested by Bringsjord, Bello and Ferrucci)
- Psychometric tests (suggested, for example, by Bringsjord and Schimansky – WAIS and ETS, or by Feng and Yong – IQ)
- The Piaget-MacGuyver Room test (suggested by Bringsjord and Licato)
- The Coffee Test (attributed to Wozniak by Goertzel)
- The Robot Student Test (suggested by Goertzel)
- The Employment Test (suggested by Nilsson)

Let us review each of them.

1.1 The Turing test

The test is likely the most prominent AI test and was introduced by Turing [4] as a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. Rather than trying to determine if a machine is thinking, Turing proposed that a human evaluator would judge natural language conversations between a human and a machine designed to generate human-like responses. The evaluator would be aware that one of the two partners in conversation is a machine, and all participants would be separated from one another. The conversation would be limited to a text-only channel such as a computer keyboard and screen so the result would not depend on the machine's ability to render words as speech.

The Turing test follows Denis Diderot formulation [5]: "If they find a parrot who could answer to everything, I would claim it to be an intelligent being without hesitation."

Considerable effort have been put over the years into building this type of behavior on computers, including multiple Loebner prize competitions.

Still, to a wide agreement, the test does not check if the machine can really think (see, for example, the book [6] – notably, e.g. [7]: “The human creators of systems undergoing Turing test know all too well that they have merely tried to fool those people who interact with their systems into believing that these systems really have minds”).

1.2 Lovelace Test

Bringsjord, Bello and Ferrucci in [7] have suggested a Lovelace test:

Artificial agent A, designed by H, passes LT if and only if

1. A outputs o;
2. A‘s outputting o is not the result of a fluke hardware error, but rather the result of processes A can repeat;
3. H (or someone who knows what H knows, and has H ‘s resources — for example, the substitute for H might he a scientist who watched and assimilated what the designers and builders of A did every step along the way) cannot explain how A produced o.

Thus, we can call Lovelace test a requirement of grand intractability. Obviously, though anecdotal in many cases, grand intractability per se is not a sign of intelligence.

1.3 Psychometric tests

Psychometric approach to AI have been suggested by Bringsjord and Schimanski in [13]: “Psychometric Al 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.”

While having a quantitative test is important for tracking progress, there is a wide criticism of psychometric tests even for humans. Being able to pass all the established tests makes the approach more interesting, but still, amenable to fooling just like the Turing test.

1.4 The Piaget-MacGuyver Room test

Bringsjord and Licato introduced the Piaget-MacGyver Room test in [14]. They define the Piaget-MacGyver Room test, “which is such that, an information-processing artifact can credibly be classified as general-intelligent if and only if it can succeed on any test constructed from the ingredients in this room. No advance notice is given to the engineers of the artifact in question, as to what the test is going to be; only the ingredients in the room are shared ahead of time. These ingredients are roughly equivalent to what would be fair game in the testing of neurobiologically normal Occidental students to see what stage within his theory of cognitive development they are at.”

1.5 The Goertzel Tests

Goertzel et al. in [8] lists several potential tests for AGI that are circulating in the AGI community:

- The Wozniak “coffee test”: go into an average American house and figure out how to make coffee, including identifying the coffee machine, figuring out what the buttons do, finding the coffee in the cabinet, etc.
- Story understanding – reading a story, or watching it on video, and then answering questions about what happened (including questions at various levels of abstraction)
- Graduating (virtual-world or robotic) preschool
- Passing the elementary school reading curriculum (which involves reading and answering questions about some picture books as well as purely textual ones)
- Learning to play an arbitrary video game based on experience only, or based on experience plus reading instructions (as it was put in [3]: The goal of this scenario would not be human level performance of any single video game, but the ability to learn and succeed at a wide range of video games, including new games unknown to the AGI developers before the competition.)

Some of these tests are already satisfied by deep-learning systems, for example, MuZero [9]. When evaluated on 57 different Atari games - the canonical video game environment for testing AI techniques - the algorithm scored 20 times better than humans in median and 50 times better on average. When evaluated on Go, chess and shogi, without any knowledge of the game rules, MuZero matched the superhuman performance of the AlphaZero algorithm that was supplied with the game rules. Thus, we can say that to a large extent mastering this specific test turns out to be a focus of specific narrow AI (reinforcement deep learning).

Similarly to the Turing test, “story understanding” and “elementary school reading curriculum” could be passed by software systems simply by manipulating symbols of which they had no understanding.

Wozniak coffee test, considered per se, requires a robot to be able to do several perception and navigation tasks that can be accomplished by a specific-purpose robot.

A test of graduating a pre-school is a more interesting one. Here a lot depends on a country and a specific preschool – requirements differ widely and the cognitive abilities to pass this test deserve a separate article, or even a book.

1.6 The Employment Test

Employment Test have been suggested by Nilsson [10]. He argues that “Machines exhibiting true human-level intelligence should be able to do many of the things humans are able to do. Among these activities are the tasks or “jobs” at which people are employed. I suggest we replace the Turing test by something I will call the “employment test.” To pass the employment test, AI programs must be able to perform the jobs ordinarily performed by humans. Progress toward human-level AI could then be measured by the fraction of these jobs that can be acceptably performed by machines.”

This is definitely a very comprehensive test, actually including the tests that we propose as the tests for the scientists jobs.

Luke Mullenhauser [11] argues that: “This is a bit “unfair” because I doubt that any single human could pass such vocational exams for any long list of economically important jobs. On the other hand, it’s quite possible that many unusually skilled humans would be able to pass all or nearly all such vocational exams if they spent an entire lifetime training each skill, and an AGI — having near-perfect memory, faster thinking speed, no need for sleep, etc. — would presumably be able to train itself in all required skills much more quickly, if it possessed the kind of general intelligence we’re trying to operationally define.”

An interesting subcase of the test have been (somewhat implicitly) suggested by Janelle Shane [12] – AGI should be able to generate (and understand?) humor. From my perspective this test is also a subcase of AGI being able to create art, and also deserves separate consideration.

2 Knowledge and cognition in science

There are multiple concurring definitions of knowledge in both AI and philosophical literature. Our goal is to define the knowledge in a way that is compatible with both contemporary epistemology and (potential) computer implementations.

2.1 Scientific knowledge and its advance

In this paper we take a critical rationalist view on the knowledge and cognition, starting from Karl Popper's view, that the advance of scientific knowledge is an evolutionary process characterized by his formula [15]:

$$PS_1 \rightarrow TT_1 \rightarrow EE_1 \rightarrow PS_2 \qquad (1)$$

In response to a given problem situation (PS_1), a number of competing conjectures, or tentative models (TT_1), are systematically subjected to attempts to define their applicability domain. This process, error elimination (EE_1), performs a similar function for science that natural selection performs for biological evolution where a species tests ecological niches. Models that better survive the process of refutation are not more true, but rather, more "fit"—in other words, more applicable to the problem situation at hand (PS_1). The evolution of models through the scientific method may, reflect a certain type of progress: toward more and more interesting problems (PS_2).

2.2 Models

The model consists in explaining the phenomenon, that is, assuming a mechanism for how it can occur. When building a model, we take a certain point of view on the phenomenon, discarding irrelevant details. Each model in scientific type knowledge has a limited domain of applicability.

In the end, in order to determine which model is better, it would be right to conduct an Experimentum crucis - that is, an experiment that would make it possible to unambiguously determine which theory is correct. In order to conduct such an experiment completely scientifically, it would be best for us to find such facts that the models would predict in different ways, and check which option is actually implemented.

2.3 Theories

As we have already said, building a model requires a certain point of view on the subject. We will call such points of view theories or paradigms. Each such theory determines what is important in the subject for consideration. A look at a person from the point of view of mechanics, electromagnetism, chemistry, and population genetics will be significantly different. In addition, each specific problem, being solved within the framework of the theory, determines what other properties of the subject we should discard when considering it within the framework of this problem.

You can demand more, for example, define a theory as S.V. Illarionov does [16]: “Theory is a holistic conceptual symbolic system, that is, it is based on some conceptual representations and is expressed in a symbolic form, in the form of symbols. Relationships are set in this system so that this symbolic system can be a reflection of a certain circle of natural phenomena or, as they sometimes say, some fragment or aspect of the material world. ”

I completely agree with Illarionov's definition for the case of scientific theories (he would consider this expression to be a pleonasm), but we will consider cognition in a framework broader than science. In applications, for example, in business, conceptual representations in the form of symbols are unnecessary.

To visually imagine theories, let's turn inside out, perhaps, the most famous metaphor in philosophy - the Platonic Cave.

So, let's imagine that a certain object (phenomenon) is in a dark cave, and the walls of this cave are our consciousness. If we illuminate the object with the light of theory on one side, we will see one shadow on the wall. This shadow is a model of the phenomenon built using this theory. If we light from another, the picture on the wall will turn out to be completely different. Moreover, in both cases, the interior of the subject will be hidden from us, and most of the information about the object will be lost.

Within the framework of one theory it is possible to build models of a multitude of phenomena. The laws of Newtonian mechanics are applicable to the motion of the planets, and to the collision of balls on the pool table. In terms of the cave metaphor, this means that you can mark many objects in one beam of light and get their models - shadows on the wall.

The most important property of scientific theories is their ability to predict phenomena that were not known at the time of their formulation. Let me cite Illarionov [16] again: “Everyday knowledge is based on previous observations of repeatedly occurring phenomena and allows you to make predictions that are very important and useful for successful practical activity, although they have the nature of probabilistic expectations. But science can do something completely different: it can predict phenomena that we have never observed. These are specifically theoretical predictions. ”

“When, in 1819, Fresnel (1788-1827) made a report on his wave theory of light at the French Academy, Poisson (1781-1840) stood up and stated that, according to this theory, in the middle of the shadow of a round screen or a ball there should be a bright spot. The next day, Augustin Fresnel and Domenic Francois Jean Arago (1786-1853) reported: there really is a bright spot. Now it is called the Poisson spot in honor of the one who instantly, in the mind, solved this problem. This did not follow from previous observations and is an example of a nontrivial theoretical prediction. ”

Nevertheless, every theory has its own limited domain of applicability. In our metaphor, this means that the light beam of the theory is limited (imagine a movie projector). Only a limited number of phenomena can be placed in our limited beam of light. Moreover, some phenomena are generally flat and turned to this beam by an edge; therefore, from the angle of this theory, the phenomenon is generally invisible or does not exist. It can be seen and understood only in the light of another, completely orthogonal theory.

3 Knowledge and cognition in other areas of human activity

3.1 Knowledge and cognition in business

Startup is most correctly defined as a temporary enterprise created to seek, develop, and validate a scalable business model (a similar definition was probably first coined by Steve Blank [17]). Here, a business model means a way of creating, using an economically sound process, for a certain type of consumers, value for which they are willing to pay money.

This search process can be divided into two separate phases:

- Customer discovery: find customer segments with a problem you can solve. Make sure that customers are willing to pay.
- Testing channels: find channels with enough customers, profitable economy and potential for scalability

A popular and mature methodology for building startups is Lean Startup [14]. With this methodology, during the customer discovery stage a startup defines the target customer segments, their problems, and what is valuable for them. Different value propositions mean different segments. The initial set of segments is considered a hypothesis. It will change. Then startup defines a value proposition for each segment and conducts problem interviews to confirm, refine or reject a hypothesis. During the interviews, they may find new segments or refine existing ones.

As soon as the problem is confirmed, the startup starts modeling economics for the segment. On what conditions do economics become profitable? Are these conditions realistic? How much money is there in this segment, is it worth the effort? If the economics is potentially profitable –the startup starts building a Minimum Viable Product. The goal is to make first manual sales of the product.

When customers are paying and the startup knows why they do so, the startup can begin testing channels. If something goes wrong, the process is repeated.

A sales channel is a combination of 3 items:

- Marketing channel – traffic source
- Sales instrument – landing page, presentation, sales script, sales letter etc.
- Product and its price

At this stage, the goal of the startup is to find scalable channels, and a goal within the channel is its profitable economics. If profitability for a user in a channel is achieved, the next goal is profitability at scale:

- Can you increase sales flow by x10 and keep it profitable?
- Is there enough channel capacity to scale?
- Will the traffic cost grow when scaling?

Thus, Lean Startup is a paradigm where a user problem is solved, this solution becomes a model of the user from the viewpoint of a Lean Startup and then the applicability domain of this solution is found by testing hypotheses

- About the value proposition for the channel
- About significant sales flow
- On the convergence of the economy in the channel
- On the convergence of economies at a scale

This means that a Lean Startup generates scientific-type knowledge.

3.2 Knowledge and cognition in engineering

It is useful to note that in our ordinary life the ability to solve problems comes solely as a result of training.

If similar problems have to be solved by a large number of people, it becomes possible to analyze and generalize the process of solving them: narrow down the scope of the process, standardize its inputs and outputs, as well as the operations performed. This is how technologies are created. Thus, a technology is a model of the process of creating a class of results, and we can deduce that the technology is another type of knowledge of scientific type. Most everything we have discussed about the scientific knowledge applies here.

4 Testing for Strong AI

Following the above on the basis of epistemological classification we can device several tests for human-like cognitive ability of the AI, in the order of their rising complexity:

- Explanation
- Problem-setting
- Refutation
- New phenomenon prediction
- Business creation

- Theory creation

4.1 Explanation test

Given a well-defined scientific theory and an empirical phenomenon, provide an explanation of the phenomenon and compute its quantitative characteristics. An example of test of this type is “Find the minimum speed that basilisk lizard can run over the water”. More problems of this sort from the physics can be found, for example, in the book of Nobel prize winner Pyotr Kapitsa [15]

4.2 Problem-setting test

Given a well-defined scientific theory and the general knowledge of the world create a task of the type mentioned in the previous subsection.

4.3 Refutation test

Given competing models/explanations for a set of empirical phenomena, device an Experimentum crucis to figure out which is better.

4.4 New phenomenon prediction

Given a well-defined scientific theory predict a phenomenon that is not previously known.

4.5 Business creation

Create a successful startup

4.6 Theory creation

Create a theory that is a meaningful improvement over existing noes in one of the scientific fields.

5 Conclusions

It is obvious from the above tests that the current state of AGI is pretty far from being really equal to human, probably as much as it was from being able to satisfy Turing test in 1950, so the paranoia of machines talking over humans in midterm, at least intellectually, seems to be pretty ungrounded. On the other hand, human history have shown that a culture should not necessary be higher or more intellectual to take over a neighboring country/region.

References

1. Kurzweil, Ray: Long Live AI, *Forbes, Aug 15* (2005), https://www.forbes.com/home/free_forbes/2005/0815/030.html, last accessed 2020/02/25
2. Franz A., Will super-human artificial intelligence (AI) be subject to evolution? H+ Magazine, Sep 6, 2013, https://hplusmagazine.com/2013/09/06/will-super-human-artificial-intelligence-ai-be-subject-to-evolution/ , (2013) last accessed 2020/04/05
3. Adams, S., Arel, I., Bach J., Coop R., Furlan R., Goertzel B., Hall J.S., Samsonovich A., Scheutz M., Schlesinger M.,. Shapiro S.C., Sowa .F.: Mapping the Landscape of Human-Level Artificial General Intelligence. AI Magazine, 33(1), 25-42. (2012) https://doi.org/10.1609/aimag.v33i1.2322
4. Turing A. M.: Computing Machinery and Intelligence. Mind, Vol. LIX. No. 236., pp. 433-460 (1950) doi:10.1093/mind/LIX.236.433
5. Diderot, D.: Pensees Philosophiques, Addition aux Pensees Philosophiques, Flammarion, p. 68, (2007) ISBN 978-2-0807-1249-3
6. Moor J. H. (editor) The Turing Test. The Elusive Standard of Artificial Intelligence. COGS, vol. 30, Kluwer Academic Publishers (2003)
7. Bringsjord, S., Bello, P. and Ferrucci D.: Creativity, the Turing Test, and the (Better) Lovelace Test, in: Moor J. H. (editor) The Turing Test. The Elusive Standard of Artificial Intelligence. COGS, vol. 30, Kluwer Academic Publishers (2003)
8. Goertzel B., Iklé M., Wigmore J. The Architecture of Human-Like General Intelligence. In: Wang P., Goertzel B. (eds) Theoretical Foundations of Artificial General Intelligence. Atlantis Thinking Machines, vol. 4. Atlantis Press, Paris (2012)
9. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T.: Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. (2019) arXiv:1911.08265.
10. Nilsson, N. J.: Human-Level Artificial Intelligence? Be Serious!. AI Magazine, 26(4), 68. (2005) https://doi.org/10.1609/aimag.v26i4.1850
11. Muehlhauser, L.: What is AGI? https://intelligence.org/2013/08/11/what-is-agi/ (2013) last accessed Feb 29th, 2020
12. Shane J.: Why did the neural network cross the road? https://aiweirdness.com/post/174691534037/why-did-the-neural-network-cross-the-road , (2018), last accessed Feb 29th, 2020
13. Bringsjord, S. and Schimanski, B.: What is artificial intelligence? Psychometric AI as an answer, Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI– 03), Morgan Kaufmann, San Francisco, CA, pp. 887–893 (2003)
14. Bringsjord S., Licato J. Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room. In: Wang P., Goertzel B. (eds) Theoretical Foundations of Artificial General Intelligence. Atlantis Thinking Machines, vol. 4. Atlantis Press, Paris (2012)
15. Popper, K.: The myth of the framework: in defence of science and rationality. Editor: Notturno M.A., Routledge. pp. 2–3. (1994) ISBN 9781135974800.
16. Illarionov S.V.: Theory of knowledge and philosophy of science. M.: ROSSPEN (2007) (In Russian)
17. Blank, S.: The Four Steps to the Epiphany: Successful Strategies for Products That Win. K&S Ranch (2013).
18. Ries E.: The Lean Startup, Crown Business (2011)
19. Kapitsa P.L.: Problems in Physics. M., Znaniye, (1966) (In Russian)