The Comparison and Measurement of Artificial Intelligence

Pei Wang

May 30, 2022

1 Introduction

According to most theories and people's intuition, it makes sense to say that one person or system is "more intelligent" than another one. According to some opinions, it even makes sense to use a number to indicate the degree of intelligence.

In general, if a certain property can be properly measured, it can be compared between any two entities with the property, but comparability does not entail measurability, because the relation can be partial (i.e., not defined between any two entities), or there is no natural unit for a measurement. Therefore, comparison is more fundamental than measurement for evaluating a property.

The comparison and measurement of intelligence have been studied in psychology and other related fields (education, anthropology, philosophy, etc.) for decades [Sternberg, 2000]. In particular, intelligence quotient (IQ) has been widely used for various purposes, though the controversies around it never stop [Gottfredson, 1997, Grigorenko and Sternberg, 1998].

The situation in AI is even more complicated, as the comparisons of intelligence may happen in three scopes:

1. among different types of intelligence (computers, humans, animals, etc.),
2. among different AI systems (e.g., NARS, SOAR, CYC, AlphaGo, etc.),
3. among different variants of the same system (e.g., OpenNARS, ONA, etc.).

In each scope, the purposes served by the comparison and measurement are not exactly the same, and consequently, the focuses of the research differ, and so do the results.

In the following, I first surveys the major issues and the various opinions on each of them. After that, I will analyze the issues and made several conclusions on the general topic of comparison and measurement in artificial intelligence. Finally, I will propose some concrete ideas in the NARS-related works.

2 Opinions and Approaches

To compare or measure a property, a precondition is to have a clear definition of the property. However, this is exactly what is lacking on "intelligence", especially in the context of "artificial intelligence" [Legg and Hutter, 2007a, Monett and Lewis, 2018, Wang, 2019]. Obviously, different definitions of intelligence require different ways of comparison and measurement, which is a major reason for the large number of approaches on this topic. For a comprehensive survey, see [Hernández-Orallo, 2017].

For instance, each of the five types of AI defined in [Wang, 2019] suggests different standard for comparison and measurement of intelligence:

Structure-AI: To compare and even measure the similarity of an AI model with the human brain. Many models have been justified in this way [Hawkins and Blakeslee, 2004, Markram, 2006], though each of them tends to be similar to the brain in different aspects, and there is little consensus on the necessary aspects of the brain to be simulated in an AI, not to mention commonly agreed measurements.

Behavior-AI: Turing Test [Turing, 1950] uses "indistinguishable from human in conversation" as the criterion of being intelligent. Though this idea has been insightful and influential, it has raised various criticisms and controversies [Hayes and Ford, 1995, French, 2000, Marcus et al., 2016]. Using human IQ tests to evaluate AI systems [Bringsjord and Schimanski, 2003] is in a similar situation.

Capability-AI: Many claims on the achievements of AI, including IBM's Watson [Ferrucci et al., 2013] and DeepMind's AlphaGo [Silver et al., 2016], are based on their extraordinary abilities in solving hard problems. Similarly, it is a common practise in machine learning to evaluate models according to their performance on benchmark datasets, which has also attracted criticisms. [Raji et al., 2021].

Function-AI: As a cognitive function is often specified as a type of computation [Russell and Norvig, 2010, Poole and Mackworth, 2017], intelligence is widely taken as abstract problem solving abilities, described by a set of desired features [Anderson and Lebiere, 2003, Laird et al., 2009], which can be checked one-by-one when evaluating a system's intelligence. There is no agreement on the relevant features yet.

Principle-AI: Such an approach evaluates the intelligence of an AI system according to its accordance with certain domain-independent principle, usually in the form of rationality or optimality [Legg and Hutter, 2007b, Wang, 2010]. Here the primary issue is the validity and applicability of the principle itself.

Though it is possible for a research project to pursue multiple objectives (such as being both brain-like and practically useful), only one of them can be considered as primary, as they are not always correlated.

Accurately speaking, projects guided by different understandings of intelligence are not comparable or commensurable in terms of which one is "more intelligent," though such comparisons nevertheless occur often. This is caused not only by the common tuition that there is only one "true intelligence", but also by the reality that all of the above schools (probably except Structure-AI) still evaluate a system's intelligence via "task accomplishing" (or call it "goal achieving" or "problem solving") tests, which is also how human intelligence is usually evaluated, compared, and measured [Hernández-Orallo, 2017]. We can say that such tests are the "greatest common factor" among the schools, though the tests are designed and evaluated according to different considerations.

In this situation, the questions commonly discussed include

(1) What tasks should be used to test a system's intelligence?

(2) What is the relationship between the testing scores and intelligence?

For the first question, the answers can be

Single task: For application-oriented systems, it is natural to use the practical problem to be solved as the sole testing case for the system's intelligence. From the very beginning of AI, there has been the belief that certain tasks are proper indicator of intelligence, such as theorem proving, game playing, natural language conversation, etc. [Feigenbaum and Feldman, 1963]. This tradition continues until today, as shown in the suggestion of looking for "Good Challenges" [Cohen, 2005] and the practise of using benchmark problems [Raji et al., 2021].

Multiple tasks: For people who see intelligence as a collection of multiple capabilities or functions [Gardner, 1983], it is necessary to use a carefully selected collection of tasks to evaluate the different aspects of intelligence, then the individual scores are summarized into a total score, like the evaluation of all-around athletes [Adams et al., 2016, Mueller et al., 2007].

All tasks: Some theoretical models assume that it is both necessary and possible to consider all problems solvable by the system, and so the size of this set or the average quality of the solution can be used to indicate the system's intelligence [Legg and Hutter, 2007b, Hernández-Orallo, 2017].

Meta-level task: Some theoretical models consider intelligence as a meta-level capability that is independent of the concrete problem-solving capabilities (i.e., the skills or expertise), so intelligence should be evaluated at the meta-level, such as the ability and efficiency for the system to learn new skills of a certain complexity or structure [Wang et al., 2018, Chollet, 2019].

After deciding the testing problems, there are still issues on how the testings should be arranged and how the scores should be used. Among the factors considered, a major distinction is made between the following two types:

Static testing. Directly use the testing scores of single tests as measurements of intelligence. This is the common practice.

Dynamic testing. Use the increasing rate of the score over a period of time as measurements of intelligence [Grigorenko and Sternberg, 1998]. This is a relatively new approach.

3 Analysis and Clarification

Most inspirations of AI research come from the study of human intelligence, though different people get their inspirations at very different levels of abstraction of the human mind/brain complex. The same is true for the comparison and measurement of intelligence.

In psychology, the study of IQ Test started to meet the practical needs to handle the difference in intellectual ability among human beings when solving selected problems [Gottfredson, 1997]. However, to directly use this approach in the comparison and measurement of AI systems runs into obvious problems. Computers have been able to solve many problems far better than human brains can do, though people intuitively do not consider that as intelligence, as analyzed in [Wang, 2019].

In [Wang et al., 2018], a simple diagram is used to intuitively illustrate different ideas about the measurement of intelligence. As a starting point, assuming we have established a reasonable measurement $S(t)$ as an indicator of an AI's problem-solving capability at a certain moment t. Using it, we can distinguish different types of "AI", as shown in Figure 1:

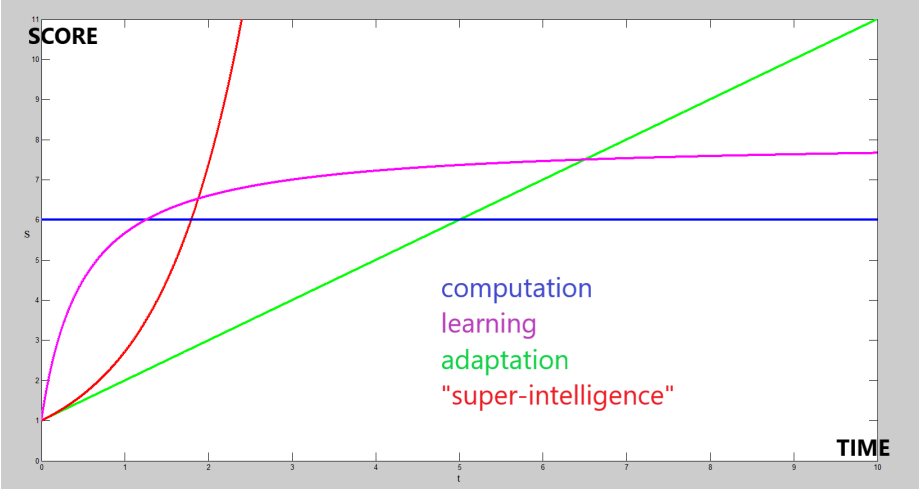

Figure 1: Different types of problem-solving ability.

Computation. If an AI is theoretically equivalent to a Turing Machine, then its problem-solving capability does not change over time, as it returns to the same initial state after each run, so the same input always leads to the same process, output, and resource expense [Hopcroft et al., 2007]. Therefore, $S(t) = c, S'(t) = 0$, meaning that its problem-solving capability remains a constant and never changes in time.

Learning. In the current Machine Learning study, "learning" is usually taken as a process in which the system's capability (of solving a problem) increases as the training goes, and eventually converges to a fixed input-output mapping [Flach, 2012]. Therefore, $S'(t) > 0, S'(t) \to 0, S(t) \to c$, meaning that its problem-solving capability increases initially, then, when learning finishes, it roughly remains a constant.

Adaptation. For a system like NARS, adaptation (or "learning" in the original and broad sense) is a never-ending process, so $S(t)$ does not necessarily converge to a stable input-output mapping. Since the system's adaptation mechanism is designed at the meta-level and largely independent of the system's experience, it can be roughly considered as a constant, that is, $S'(t) = c$ [Wang, 2006].

Super-intelligence. This is a possibility suggested by some researchers, who take intelligence as a ladder that extends beyond the "human-level", so some future AI may work according to mechanisms beyond our comprehension [Kurzweil, 2006, Bostrom, 2012]. In Figure 1, this possibility is represented by a function where $S(t)$ and $S'(t)$ are both increasing functions of time.

The major difference among these types of systems is in their directives, $S'(t)$, rather than in $S(t)$ themselves. Actually, this is the case even when human intelligence is measured, as IQ is a *quotient* obtained by dividing the subject's "mental age," which correlates to $S(t)$, by the person's chronological age, which correlates to t. $S'(t)$ is defined by the limit of $(S(t) - S(t_0))/(t - t_0)$ when $(t - t_0) \to 0$. Therefore, what IQ measures is adaptation/learning capability, rather than problem-solving capability.

In its daily usage, the term "intelligence" is indeed often used to indicate human problem-solving ability, because among humans, their innate capabilities, i.e., the initial values $S(t_0)$, are relatively similar to each other, so their values at a later time $S(t_n)$ roughly correlate to their derivatives $S'(t)$ in the period $[t_0, t_n]$. However, this rough correlation does not exist among computer systems, because a system with a high score $S(t_n)$ on certain tests may have no adapting/learning capability at all (e.g., in Turing computation $S(t_n) = S(t_0)$). From the value of $S(t_n)$ and t_n, $S'(t)$ cannot even be estimated without known the $S(t_0)$ for that specific system.

The above analysis leads to the conclusion that "problem-solving capability" and "adapting/learning capability" need separate measurements, and the latter is closer to the intuitive meaning of "intelligence", while the former is better called "skill" or "expertise."

Between $S(t)$ and $S'(t)$, the former usually heavily depends on the system's knowledge on the specific testing problem, while the latter being less problem-specific, and depends more on the system's general-purpose cognitive capability, which is another reason why it better fits the label "intelligence." In general, skills come mainly from the system's *nurture*, while intelligence comes mainly from the system's *nature*.

Turing computation specifies functions and algorithms independent of their usage history, while the current machine learning research interprets "learning" as "to learn a computation or function", which is arguably different from the open-ended learning processes in the human mind [Wang and Li, 2016]. What is labeled as "adaptation" is closer to what the term "learning" means in psychology and everyday life. Of course, it does not mean that intelligence remains unchanged in the system's lifetime, but that its change is relatively much smaller than the change of the skills, so in Figure 1 it is only roughly taken as a constant.

As for the systems where $S'(t)$ changes as much as $S(t)$, like the "super-intelligence" in Figure 1, I do not consider it as a realistic possibility that deserve analysis and discussion, but only a theoretical one [Wang et al., 2018], because such a system does not even have a nature to be specified except being "superhuman" and incomprehensible.

In conclusion, the above analysis shows that testings for intelligence should be *dynamic* [Grigorenko and Sternberg, 1998] and on the system's (meta-level and domain-independent) *skill acquisition* capability, rather than about the specific skills it has at a moment [Chollet, 2019], even though the tests are inevitably about specific skills.

4 Proposals for NARS

Based on the previous analysis, some suggestions are made about the comparison and measurements of intelligence between (an implementation of) NARS and other systems.

4.1 Between different versions of NARS

What is immediately needed in NARS research is the guideline for comparison between different design decisions with respect to their impacts to the system's intelligence.

Since all the designs follow the basic principles of NARS, including its working definition of intelligence and development strategy, some comparisons can be done via theoretical analysis, with respect to the objective of "being adaptive under AIKR (the Assumption of Insufficient Knowledge and Resources)." These comparisons will be at the meat-level, in the sense that they are only about the built-in components of the system, and independent of the system's experience or content in memory.

Everything else being the same, a more intelligent NARS will be "more adaptive" by interacting with its environment in more complicated manners,

with

- higher expressive power (i.e., a richer Narsese grammar),
- higher inferential power (i.e., more NAL rules),
- more input/output channels (i.e., more sensorimotor capabilities).

The above factors are all about the *logic* part of the system.

For the *control* part of NARS, to be "more intelligent" means to have higher resource efficiency, especially with respect to computational time and space. This is a more complicated topic than that in the logic part. Here some decision decisions can still be justified via theoretical analysis, if they do not depend on special properties of the situations or tasks. For those that cannot be decided theoretically, practical testings will be needed, though there are still general guidelines:

- The performance improvements obtained by proper training and education should not be considered as intelligence improvements, as they are at the object-level.
- The performance improvements in certain types of tasks may come at a cost in other types of tasks (such as bias in resource allocation among channels, buffers, and the memory). They will produce NARS implementations with different "personalities" and each adapts well in different domains or environments, with roughly the same level of intelligence.
- The testing tasks should be as domain independent as possible. For instance, if the desired performance can be provided by the acquiring of a certain type of compound operation, then the tests should be able to use components from different domains to get the same effects. In this way, systems' intelligence can be compared even if they have disjoint operations, such as among robots with different sensorimotor mechanisms.
- When testing a design change, diverse tasks are more suitable than uniform tasks to show the effects of the change in various aspects of the system, where intelligence can be considered as the overall effect, like the "g-factor" that is based on the assumed correlation among cognitive functions.
- Testing for resource efficiency can be carried out by comparing answers to questions or memory snapshots at different moment, and/or with difference space parameters.

In general, it is possible to compare the difference of intelligence among different versions of NARS according to the above guidelines, though the results only establish a partial order among the versions, rather than a total order, since there may still be dependency on environments and tasks. It is possible to define a numerical IQ as the quantile among the comparable versions, though it probably will not be very useful at this stage of the research.

4.2 Between NARS and other AI systems

In principle, NARS cannot be compared with AI systems developed to achieve a very different understanding of "intelligence," as they are actually aim at different goals.

For AI systems whose working definitions of intelligence is close enough with NARS, comparisons can be approximately carried out, both theoretically and empirically. Again, here the comparisons should be at the meta-level, though the tasting problems will be at object-level.

It makes sense to compare NARS with other AI systems in their capabilities at a specific application, though those comparisons are not about their intelligence, but their skills and applicability. The skill of NARS on solving a specific problem can be improved by proper training and education without a redesign, and therefore may have little to do with the system's intelligence.

For example, Chollet wrote "The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty" [Chollet, 2019] that has overlap with the definition of intelligence of NARS, though the ARC dataset he proposed still looks too problem-specific to be used on NARS, as it is completely focused on vision, and does not encourage hierarchical perception. In that aspect, ARC is even less proper than the Bongard Problem [Hofstadter, 1979].

4.3 Between NARS and other types of intelligence

It makes sense to compare NARS with various forms of natural intelligence (of human, animal, or their collections). Since these systems can be considered as being adaptive under AIKR, the comparisons between NARS and human or animal will actually be closer to the comparison between variants of NARS than between NARS and other "AI systems".

As discussed above, the comparisons should be at the meta-level (cognitive ability and mechanism), not object-level (problem-solving capability or performance), even though the testing problems must be concrete. Even when the system's performance can be measured with a score S, what matters most is not $S(t)$, but $S'(t)$, that is, how the score changes with experience. Therefore, some type of dynamic testing will be more proper here.

According to my working definition of intelligence [Wang, 2008, Wang, 2019], the similarity among various forms of intelligence is in the relationship between a system's experience and behavior (adaptation under AIKR), rather than in the content of the specific experience and behavior. In particular, AGI is not designed and developed to replace humans, as such a system does not necessarily have human-like experience, nor human-like behaviors or problem-solving capabilities. For example, a robot's sensors and actuators may be completely different from that of human beings, while still be highly intelligent. To test the intelligence of such a robot using human problems will be totally pointless.

5 Conclusions

The comparison/measurement of intelligence is both necessary and possible, though there are many misconceptions in the current practise.

Between AI systems designed according to very different understandings of "intelligence," no common comparisons and measurements are meaningful, and so are milestones, benchmark problems, etc.

Intelligence should be considered as a meta-level property, though all practical testings must use object-level problems. Therefore, the selection of problems will inevitably introduce bias into the results. Consequently, each test contributes evidence to the evaluation of intelligence, but cannot provide conclusive result once for all. It means we cannot expect to develop a single test to settle the issue once for all, but to depend on many such tests.

For intelligence, *comparison* is more fundamental than *measurement*, and the latter can be established on the former as a relative rank, that is, if a system is more intelligent than p percent of the other system that have been compared to it, then p can be used as its "IQ." As the scope of comparable systems change, such a measurement will only have a temporal value. Even so, it still makes more sense than applying a human IQ test to AGI, or to depend on a fixed set of problems, no matter how they are selected or designed.

References

[Adams et al., 2016] Adams, S. S., Banavar, G., and Campbell, M. (2016). I-athlon: Toward a multidimensional Turing Test. *AI Magazine*, 37(1):78–84.

[Anderson and Lebiere, 2003] Anderson, J. R. and Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Sciences*, 26(1):587–640.

[Bostrom, 2012] Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*.

[Bringsjord and Schimanski, 2003] 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, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Chollet, 2019] Chollet, F. (2019). On the measure of intelligence. *CoRR*, abs/1911.01547.

[Cohen, 2005] Cohen, P. R. (2005). If not Turing's Test, then what? *AI Magazine*, 26:61–67.

[Feigenbaum and Feldman, 1963] Feigenbaum, E. A. and Feldman, J., editors (1963). *Computers and Thought*. McGraw-Hill, New York.

[Ferrucci et al., 2013] Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, 199:93–105.

[Flach, 2012] Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, New York, NY, USA.

[French, 2000] French, R. M. (2000). The Turing Test: the first fifty years. *Trends in Cognitive Sciences*, 4(3):115–121.

[Gardner, 1983] Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Book.

[Gottfredson, 1997] Gottfredson, L. S. (1997). Mainstream science on intelligence: an editorial with 52 signatories, history, and bibliography. *Intelligence*, 24:13–23.

[Grigorenko and Sternberg, 1998] Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124(1):75–111.

[Hawkins and Blakeslee, 2004] Hawkins, J. and Blakeslee, S. (2004). *On Intelligence*. Times Books, New York.

[Hayes and Ford, 1995] Hayes, P. and Ford, K. (1995). Turing Test considered harmful. In Mellish, C. S., editor, *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, pages 972–977.

[Hernández-Orallo, 2017] Hernández-Orallo, J. (2017). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, Cambridge.

[Hofstadter, 1979] Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books, New York.

[Hopcroft et al., 2007] Hopcroft, J. E., Motwani, R., and Ullman, J. D. (2007). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Boston, 3rd edition.

[Kurzweil, 2006] Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. Penguin Books, New York.

[Laird et al., 2009] Laird, J. E., Wray, R. E., Marinier, R. P., and Langley, P. (2009). Claims and challenges in evaluating human-level intelligent systems. In Goertzel, B., Hitzler, P., and Hutter, M., editors, *Proceedings of the Second Conference on Artificial General Intelligence*, pages 91–96.

[Legg and Hutter, 2007a] Legg, S. and Hutter, M. (2007a). A collection of definitions of intelligence. In Goertzel, B. and Wang, P., editors, *Advance of Artificial General Intelligence*, pages 17–24. IOS Press, Amsterdam.

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

[Marcus et al., 2016] Marcus, G., Rossi, F., and Veloso, M. M. (2016). Beyond the Turing Test. *AI Magazine*, 37(1):3–4.

[Markram, 2006] Markram, H. (2006). The Blue Brain project. *Nature Reviews Neuroscience*, 7(2):153–160.

[Monett and Lewis, 2018] Monett, D. and Lewis, C. W. P. (2018). Getting clarity by defining artificial intelligence - a survey. In Müller, V. C., editor, *Philosophy and Theory of Artificial Intelligence 2017*, pages 212–214. Springer, Berlin.

[Mueller et al., 2007] Mueller, S. T., Jones, M., Minnery, B. S., and Hiland, J. M. (2007). The BICA Cognitive Decathlon: A test suite for biologically-inspired cognitive agents. In *Proceedings of the Behavior Representation in Modeling and Simulation Conference*.

[Poole and Mackworth, 2017] Poole, D. L. and Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, Cambridge, 2 edition.

[Raji et al., 2021] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *CoRR*, abs/2111.15366.

[Russell and Norvig, 2010] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 3rd edition.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489.

[Sternberg, 2000] Sternberg, R. J. (2000). *Handbook of intelligence*. Cambridge University Press.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX:433–460.

[Wang, 2006] Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Springer, Dordrecht.

[Wang, 2008] Wang, P. (2008). What do you mean by "AI". In Wang, P., Goertzel, B., and Franklin, S., editors, *Proceedings of the First Conference on Artificial General Intelligence*, pages 362–373.

[Wang, 2010] Wang, P. (2010). The evaluation of AGI systems. In Baum, E. B., Hutter, M., and Kitzelmann, E., editors, *Proceedings of the Third Conference on Artificial General Intelligence*, pages 164–169.

[Wang, 2019] Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37.

[Wang and Li, 2016] Wang, P. and Li, X. (2016). Different conceptions of learning: Function approximation vs. self-organization. In Steunebrink, B., Wang, P., and Goertzel, B., editors, *Proceedings of the Ninth Conference on Artificial General Intelligence*, pages 140–149.

[Wang et al., 2018] Wang, P., Liu, K., and Dougherty, Q. (2018). Conceptions of artificial intelligence and singularity. *Information*, 9(4).