Journal of Artificial General Intelligence 10(2) 1-37, 2019 Submitted 2019-03-29
DOI: 10.2478/jagi-2019-0002 Accepted 2019-08-12

On Defining Artificial Intelligence

Pei Wang PEI.WANG@TEMPLE.EDU
Department of Computer and Information Sciences, Temple University
1925 North 12th Street, Philadelphia, PA 19122-1801, USA

Editors: Dagmar Monett, Colin W. P. Lewis, Kristinn R. Thórisson

Abstract

This article systematically analyzes the problem of defining "artificial intelligence." It starts by pointing out that a definition influences the path of the research, then establishes four criteria of a good working definition of a notion: being similar to its common usage, drawing a sharp boundary, leading to fruitful research, and as simple as possible. According to these criteria, the representative definitions in the field are analyzed. A new definition is proposed, according to it intelligence means "adaptation with insufficient knowledge and resources." The implications of this definition are discussed, and it is compared with the other definitions. It is claimed that this definition sheds light on the solution of many existing problems and sets a sound foundation for the field.

> *"Would you tell me, please, which way I ought to go from here?"*
> *"That depends a good deal on where you want to get to," said the Cat.*
> *"I don't much care where–" said Alice.*
> *"Then it doesn't matter which way you go," said the Cat.*
> *"–so long as I get SOMEWHERE," Alice added as an explanation.*
> *"Oh, you're sure to do that," said the Cat, "if you only walk long enough."*
> — Lewis Carroll, *Alice's Adventures in Wonderland*

1. The Problem

1.1 Why define AI

It is well known that there is no widely accepted definition of Artificial Intelligence (AI) (Kirsh, 1991; Allen, 1998; Hearst and Hirsh, 2000; Brachman, 2006; Nilsson, 2009; Bhatnagar et al., 2018; Monett and Lewis, 2018). Consequently, the term "AI" has been used with many different senses, both within the field and outside it.

Many people do not consider it a big problem. After all, many scientific concepts get good definitions only after the research matures, rather than at the beginning of the study. Given the complexity of intelligence, it is unrealistic to expect a commonly accepted definition of AI at the current stage of the research. Instead of spending time in a debate on definitions, many researchers would rather pursue whatever objective that is fruitful either in theory or in practice, no matter whether it is labeled as "AI" or not.

The above opinion is agreeable to an extent. We can neither suspend the research until a definition is accepted by the community, nor expect a consensus to be arrived merely by theoretical analysis. Nevertheless, there are still reasons to pay attention to this topic at the current time.

With the recent achievements of deep learning (LeCun, Bengio, and Hinton, 2015; Silver et al., 2016), AI has become a hot topic that attracts a lot of public attention. The business world is making strategies to deal with this opportunity and challenge (Goldman Sachs, 2016), and there are even legal and political regulations and policies proposed to deal with AI (Executive Office of the President, USA, 2016). However, without a clear definition of the term, "it is difficult for policy makers to assess what AI systems will be able to do in the near future, and how the field may get there. There is no common framework to determine which kinds of AI systems are even desirable" (Bhatnagar et al., 2018).

The situation is no better within the AI community. "Theories of intelligence and the goal of Artificial Intelligence (A.I.) have been the source of much confusion both within the field and among the general public" (Monett and Lewis, 2018) – among people with different opinions on what "AI" means, there is little chance for them to agree on how to build one, or to agree on the evaluation criteria, benchmark tests, milestones, etc., which are crucial for the healthy growth of a research community (Hernández-Orallo, 2017). It also makes cooperation difficult among different groups.

Even for a single research project, it is common to meet conflict of ideals, where some design decisions are based on one interpretation of AI, and others on a different interpretation. When the incompatibility of these interpretations becomes significant, the project could have fatal troubles that cannot be dealt with technical remedies. For example, in brain-inspired architectures (Reeke and Edelman, 1988; Hawkins and Blakeslee, 2004), there is usually a tension between making it more biologically realistic and realizing more useful functions. The initial design of Soar attempted to meet the needs of both artificial intelligence and cognitive psychology (Newell, 1990), while in the recent years the project has been mainly shaped by AI considerations (Laird, 2012).

Though a well-defined concept is not easy to obtain, its benefits are hard to overstress. It will prevent implicit assumptions from misguiding a research project and avoid many misunderstandings in discussions and debates. Therefore, it is worth the effort to give this topic a treatment it deserves.

This paper is a summary of my previous opinions and arguments on this topic (Wang, 1994, 1995, 2006a, 2008, 2012; Wang, Liu, and Dougherty, 2018), with additional discussions to give it a more systematic and comprehensive treatment. In the following, I will start at the meta-level by discussing definitions in general, then move to the specific case of defining intelligence and AI. After summarizing the proposed definitions, I will introduce my own definition, and compare it with the others, so as to clarify some assumptions in this discussion that are often implicit or hidden.

1.2 What is a definition

In its most common sense, a definition specifies the meaning or significance of a word or phrase. Though this sounds plain, there are still subtle issues to be noticed in the context of defining AI.

First, a definition can be about either the usage of a word or the content of a concept expressed by a word, and in most situations the debate on the meaning of AI is more about the latter than about the former, though it is the former that is directly mentioned. In the current discussion, if "artificial intelligence" was replaced by "computer intelligence" or "machine intelligence," then the underlying problem would not change much. The same is true if the concept is expressed not in English, but another human language. After all, the key issue is not in the choice of the words, but in the idea expressed by them, so this discussion is largely language independent. When the concept to be expressed becomes relatively well-defined, which words are chosen to express it is still a non-trivial problem, but it is secondary. Therefore, issues like the word "artificial" may be associated with "fake" are not what I want to discuss there. Instead, the focus will be on the concepts involved.

By specifying its sufficient and necessary conditions, the definition of a concept draws its boundary, and therefore regulates its usage in thinking and communication. However, even with these obvious advantages, we cannot expect every concept to be well-defined from the very beginning, even for scientific concepts, because every phenomenon studied by science is initially conceptualized as a vague idea, and concepts are fluid in nature (Hofstadter and FARG, 1995). For instance, rather than being defined at the beginning of research fields such as physics, chemistry and biology, their boundaries have formed gradually over time. In general, to have a clear definition is not a precondition for a concept to be used in both scientific research and discussions, though it is indeed highly desired.

It is often neglected that in scientific discussions there are (at least) two types of definitions with different properties: a *dictionary definition* is descriptive as it summarizes the existing usage of the term, while a *working definition* is prescriptive as it specifies a proposed usage of the term. Both are useful, but for different purposes. The former represents a widely accepted standard, while the latter is initially proposed by a single researcher or research team, which may or may not gradually become the common opinion, but whose main purpose is to guide in the research that is being undertaken. It is important to distinguish these two types of definition. For instance, a new theory often uses an existing term in a novel way, which cannot be simply criticized as violating its definition.

With respect to the concept of AI, its dictionary definition is relatively clear – it is nothing but what the AI researchers have been doing. Such a definition is useful for certain purposes, such as for a journal or conference reviewer to decide whether a submission is within the scope of acceptance. On the other hand, a working definition of AI sets the research objective for an AI project – it is a clarification on "what I/we mean by *AI*," which may not agree with the dictionary definition. Given the diverse usages of the term at the current time, to take "what the AI researchers have been doing" as a working definition would lead a research project into chaos. In the following, the discussion is focused on the working definition of intelligence initially from the perspective of AI research, rather than on its dictionary definition, though the latter is still relevant.

1.3 What is a good working definition

The task of choosing a proper working definition is not unique to AI, but is in all branches of science, as well as in many other domains, though in most cases the choice is relatively obvious, so the decision is often simply declared, rather than justified with detailed arguments.

One commendable exception is Carnap's treatment of the concept of "probability" (Carnap, 1950). When attempting to provide a solid foundation for probability theory, Carnap needed to start with a proper definition of probability, or in his word, he wanted to provide an *explicatum* for the *explicandum* embedded in the common usage. Instead of simply throwing out a definition that looked good to him, he first set up the following four requirements:

1. Similarity to the explicandum,
2. Exactness,
3. Fruitfulness,
4. Simplicity.

Of course, Carnap is not the only one who has proposed requirements for working definitions in scientific theories, though these four seem to fit the situation of an AI definition quite well. Therefore, in this section I will discuss what each of the requirements means in the context of AI, though some other features of definitions will also be addressed later in the article.

1.3.1 SIMILARITY TO THE EXPLICANDUM

In our terminology, this requirement asks a working definition to be similar to the dictionary definition of the concept. In "AI," how to interpret the "A" is not a big issue, and the troubles come mostly from the "I."

Though "intelligence" has been used without a well-defined boundary, there are still some common usages that can be taken as basic, which indicate what the concept should include, and what it should exclude.

First, the concept began as an attribute of human beings, and is especially about the mental or intellectual capability displayed by humans. Therefore it is historically *anthropocentric*, and if a working definition of intelligence could even exclude a normal (average) human being, it would not be acceptable – no matter how good such a definition is in other aspects, it is not about the intelligence as we intuitively understand, but about something else.

On the other hand, it is meaningful to talk about non-human intelligence. AI is certainly such a case, and there also have been studies on animal intelligence (Tomasello, 2000; Goldstein, Princiotta, and Naglieri, 2015), collective/group intelligence (Hofstadter, 1979; Leimeister, 2010), and alien/extraterrestrial intelligence (Regis, 1985; Cabrol, 2016). Though there are many controversies in each of these discussions, as far as we take such a discussion as meaningful, we have already accepted the usage of intelligence as a general concept with multiple special cases that can be different from each other here or there while still maintaining certain common nature (Bhatnagar et al., 2018).

According to this consideration, a working definition of intelligence cannot be too anthropocentric to the extent that non-human intelligence becomes impossible *by definition*. It follows that the definition cannot depend on human-specific properties, which can be biological, historical, social, etc. An intelligent being does not have to be human-like in all aspects, otherwise "intelligence" and "human intelligence" would be the same concept.

However, it does not mean that we want a concept to be defined as broadly as possible, since that will make it vacant and trivial. In the case of intelligence, that would also violate the common usage of the concept, since people do not consider everything as being intelligent. Most people do not consider a conventional computer program for sorting or arithmetic calculation to be intelligent, though it does carry out certain "intellectual" activities, and is useful and valuable.

Finally, a working definition does not need to cover all common usages of a concept. For example, in the commercial world the label "intelligent" is often used to mean "more powerful" or "better," which is a usage that can be neglected for the current purpose, since it is not part of the core meaning of the concept, though may be a derived sense.

1.3.2 EXACTNESS

The demands for a definition are raised to resolve the ambiguity of the ordinary concepts in their common usages. Ideally, a definition should provide an accurate sufficient and necessary condition for deciding the applicability of the concept in all situations.

For this reason, "intelligence" should not be defined in terms of other vague concepts, such as "mind," "thinking," "cognition," "wisdom," "consciousness," etc. without defining them first (which is no less complicated than defining intelligence). Such a definition is not wrong, but fails to draw a sharp line between intelligent beings and unintelligent ones, as a definition is supposed to do.

This requirement is still meaningful even if intelligence is (as it should be) considered as a matter of degree (Hernández-Orallo, 2017). In this situation, the definition should provide guidance for this degree to be determined.

It is why formal definitions are preferred, as they are generally more accurate and less ambiguous by using symbols with the intended meanings to replace words in a natural language. However, it should be kept in mind that since the concept of intelligence has empirical content, its definition cannot be completely formal. In (empirical) science, a formal definition only captures the relatively stable aspects of a concept, so is hard to get before the field becomes relatively mature. Before that, the "propensity for premature formalization" may hurt the progress in the field (Thórisson, 2013).

Even a formal definition still needs interpretation when it is applied to a practical situation, and the existence of different interpretations may undermine the exactness of the definition. For example, though the mathematical meaning of probability is fully specified by the axioms of probability theory, its applications still have controversies (Carnap, 1950; Hájek, 2012).

Therefore, the demand for exactness can only be relatively satisfied, as there is no way to completely remove ambiguity in a definition. This is also because some concepts used in the definition may not have exact definitions themselves, and to demand their definitions will cause an infinite regression. No matter how hard we have tried, we have to stop somewhere and depend on some common understanding about some concepts as the starting point of a definition process.

1.3.3 FRUITFULNESS

This is the requirement of being fruitful that distinguishes a working definition from a dictionary definition. When a researcher or a research team defines AI, normally it is not taken as something that already fully exists, but something to be built. To serve the role of being a research objective, a working definition of intelligence, and the derived definition of AI, should set a clear goal for the research, as well as to provide guidance for the following work.

There are many justifiable descriptions about intelligence, but most of them cannot play the role of a working definition well, as they do not provide clear instructions and restrictions for the design decisions when building an AI system. Of course, a definition by itself is not enough to solve all the problems in research, though it nevertheless provides the most fundamental postulations for the project. In particular, the definition distinguishes the features of human intelligence that need to be reproduced in an AI system from those that can be omitted as irrelevant.

Another function of a working definition is to shed light on the solving of the existing problems in AI. Contrary to a popular belief, in scientific research the introduction of a new concept is not encouraged, unless it contributes to the research in a unique way. Many definitions of AI are disliked by researchers, not because they are wrong, but because they are not useful.

Finally, a working definition of AI should give the field a proper identity by specifying its subject matter and scope, which will decide its relationship with other fields, such as computer science and cognitive psychology. This should establish AI as a domain with its unique problems and solutions, and show is as neither a collection of heterogeneous instances nor a novel label of an existing domain.

1.3.4 SIMPLICITY

It is widely agreed that a scientific concept should be as simple as possible. This requirement also appears in other forms, such as the preference of elegance and beauty, which can be interpreted as conceptual simplicity.

Though the favoring of simplicity is uncontroversial, it has been given different reasons and interpretations. For example, in the tradition of Solomonoff (1964), simpler hypotheses are assumed to have higher probability to be correct (Hutter, 2005), while to some other researchers (including me), the preference for simplicity is derived from the requirement of efficiency and economy of cognition (Wang, 2006b), and it is not correlated with correctness.

This requirement does not deny the complexity of the processes involving intelligence. Here the hope is to identify certain essential features of intelligence, from which many other features can be implied.

1.3.5 OVERALL EVALUATION

For a given working definition, usually whether it satisfies each of the above requirements is not a matter of yes or no, but a matter of degree. It is hard to establish a general and practical method to measure the degrees, but it does not mean that they cannot lead to meaningful conclusions. Usually they are used relatively, that is, we can compare two working definitions with respect to a requirement to see which one is better. Actually, this is exactly what we can expect: what is needed is the best definition among the available candidates, no matter what "score" it gets.

What makes the comparison tricky is the conflicts among the requirements. It is often the case that one definition is better by one standard (say, simpler) but not as good by another one (say, less fruitful). Consequently, the final choice may be a compromise, or a "weighted sum" of the individual scores on each dimension, and the weights are decided subjectively. Different researchers value the requirements differently, though they usually agree on the relevance of them.

In conclusion, even though intelligence is hard to define, an AI researcher still inevitably gives it a working definition, as far as he/she claims to be "doing AI." The difference is only at whether the definition is carefully deliberated and explicitly announced, or implicitly accepted or revealed by the choice of research goal, approach, evaluation standard, etc. This is true even for the researchers who consider all discussions on defining AI to be a waste of time, who may eventually suffer from the unanticipated consequences produced by their unquestioned assumptions.

AI may indeed have no widely-accepted definition of the phenomenon it claims to be studying, but this is not a good reason to accept an arbitrary (working) definition as a foundation to carry out one's research.

2. Practices in Defining AI

2.1 Historic development

The research goals in the field of AI have been changing over the years. After the invention of the computer in the 1940s, people soon realized that its capability is not limited to numerical calculation, and that it can be used to carry out many intellectual tasks that are usually considered as demanding human intelligence. Computers were called "giant electronic brains," and several visionary researchers proposed theories that stress the common features of the machines and the minds, including McCulloch and Pitts (1943), Wiener (1948), Shannon and Weaver (1949), Turing (1950), and von Neumann (1958).

Though the above researchers can be considered as pioneers of AI research, and their works have influenced generations of researchers, the research field known as AI today was mainly founded by McCarthy, Minsky, Newell, and Simon. This is not merely because they participated in the Dartmouth meeting (McCarthy et al., 1955) where the phrase "Artificial Intelligence" was coined, but because they established three leading research centers, and their ideas have largely shaped the path of mainstream AI for decades. The following are their opinions about what AI is about:

> "By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity" (Newell and Simon, 1976).
>
> Intelligence usually means "the ability to solve hard problems" (Minsky, 1985a).
>
> "AI is concerned with methods of achieving goals in situations in which the information available has a certain complex character. The methods that have to be used are related to the problem presented by the situation and are similar whether the problem solver is human, a Martian, or a computer program" (McCarthy, 1988).

As mainstream AI has been guided by these intuitive but vague conceptions of intelligence, it has grown into a field that lacks not only a common theoretical foundation, but also a consensus on the overall research objective. Consequently, there is much disagreement on evaluation criteria, progress milestones, benchmark problems, etc. It is normal for a scientific domain to have competing paradigms (Kuhn, 1970), but researchers in other domains at least agree on the problems to be studied.

To the larger community of computer science and information technology, AI is usually identified by the techniques grown from it, which at different periods may include theorem proving, heuristic search, game playing, expert systems, neural networks, Bayesian networks, data mining, agents, and recently, deep learning. Since these techniques are based on very different theoretical foundations and are applicable to different problems, various subdomains have been formed within AI, such as knowledge representation, reasoning, planning, machine learning, vision, natural language processing, robotics, etc. Many researchers identify themselves much closer with these subdomains than with AI, and treat the latter like an optional label that can be added to or dropped from their projects depending on the current public image of AI, which has been on a roller coaster.

Different attitudes towards this diversity can be perceived from two AAAI Presidential Addresses:

> "I want to consider intelligence as a collective noun. I want to see what we in AI have thought of it and review the multiple ways in which we've conceived of it. ... to conceive of AI as the study of the design space of intelligences" (Davis, 1998).
>
> "It has been hypothesized that whatever intelligence is (and we obviously have not been able to fully define it so far), it is a multidimensional thing. ... We must consider the integration and synergies of components in an overall system to really approach some form of artificial intelligence" (Brachman, 2006).

While Davis took the diversity within the field to be an admirable feature, Brachman was concerned about the fragmentation of the research community. Though many attempts have been made in recent years to integrate the subdomains, there is still no consensus on many major issues, including what AI or intelligence means. In conferences, journals, and textbooks, "AI" has been interpreted in a very broad and loose manner, and consequently covers a wide variety of theories and approaches (Luger, 2008; Russell and Norvig, 2010; Poole and Mackworth, 2017).

2.2 Different abstractions of human intelligence

A recent survey (Monett and Lewis, 2018) identified hundreds of definitions of intelligence. In this section I will not analyze individual definitions, but the major perspectives from which the definitions are proposed.

AI is conceived as computer systems that are similar to the human mind in a certain sense, though a computer and a human mind cannot be identical in all aspects. Therefore, here the key issue is *where* the two are similar or even the same. We will see that every working definition of AI corresponds to an abstraction of the human mind that describes the mind from a certain point of view, or at a certain level of abstraction, under the belief that it is what intelligence is really about. This abstraction guides the construction of a computer system that is similar to a human mind in that sense, while neglecting other aspects of the human mind as irrelevant or secondary.

To more clearly show the similarities and differences among these abstractions, in the following both humans and computers are described in a very simple formal framework (Wang, 2008). An agent or system and its interaction with the environment is specified as a tuple $\langle P, S, A \rangle$, where each of the three components describes the agent's *input signals*, *internal states*, and *output actions*, respectively, in a series of moments. For the sequence of moments $0, \ldots, t$, $P = \langle p_0, \ldots, p_t \rangle$ is the sequence of percepts acquired (as input), $A = \langle a_0, \ldots, a_t \rangle$ is the sequence of actions executed (as output), and $S = \langle s_0, \ldots, s_t \rangle$ is the sequence of internal states the agent has gone through. When a human is written as $H = \langle P^H, S^H, A^H \rangle$ and an intelligent computer as $C = \langle P^C, S^C, A^C \rangle$, a working definition of intelligence corresponds to a way to define $C \approx H$ in terms of their components, that is, it explains *in what sense* C and H are similar.

2.2.1 STRUCTURE-AI

The rationale of this perspective seems self-evident. After all, intelligence starts as a notion describing the mental capability produced by the human brain, so the most reliable way to reproduce it is to faithfully simulate the brain. Representative expressions of this position include "the ultimate goals of AI and neuroscience are quite similar" (Reeke and Edelman, 1988) and "At a more fundamental level, any computational model of learning must ultimately be grounded in the brain's biological neural networks" (Lake et al., 2017).

I call this type of definition "Structure-AI," since it requires an AI system to go through isomorphic states or structure changes as the brain does when they are given similar input, which will produce similar output, so the three components of the two are pairwise similar to each other:

$$P^C \approx P^H, \; S^C \approx S^H, \; A^C \approx A^H$$

From this perspective, similarity to the brain is the main standard and justification of the design, rather than merely as a source of inspiration. Therefore, this group includes the brain modeling projects (Hawkins and Blakeslee, 2004; Markram, 2006; Koene and Deca, 2013), but not the artificial neural networks (McCulloch and Pitts, 1943; Rumelhart and McClelland, 1986).

Given the complexity of the human brain, such a project must be very difficult, and it will heavily depend on the progress of neuroscience, but feasibility is not the concern of this article, since the focus is on the identity each definition gives to AI. In that aspect, the major criticism is that this definition is too anthropocentric. As explained above, a fundamental intuition behind AI is that human intelligence is a special form of a general notion of intelligence, which have other forms. Using a well-known metaphor, "to fly" and "to fly as a bird" both can be taken as engineering goals, but they are different goals. The latter is possible (though difficult), but if the former is understood as identical to the latter, many valuable designs will be omitted or even disqualified, simply because they are not similar to the original.

If it turns out to be the case that the only way to get intelligence is doing exactly what the human brain does, then AI should be considered as an ill-conceived concept. Instead, we would better talk about brain modeling or emulation. A related issue is whether "mind" can be completely reduced into "brain." If it is not the case, then a good model of the brain and a good model of the mind are not the same, and the intuitive meaning of intelligence is closer to the latter than to the former.

2.2.2 BEHAVIOR-AI

One way to acknowledge a human-like mind without demanding a human-like brain is to associate intelligence to the external behaviors of the agent. After all, if an agent behaves like a human, it should be considered as intelligent, no matter whether it looks like a human, either inside or outside.

In the agent framework, it means that C is similar to H in the sense that:

$$P^C \approx P^H, \; A^C \approx A^H$$

that is, the two should have similar input-output streams, without requiring any corresponding internal structures and states.

The best-known example of this perspective is the Turing Test (Turing, 1950), which states that if a computer system's verbal behaviors are indistinguishable from that of a human being, it can be considered as intelligent, or a thinking machine, for all practical purposes.

The Turing Test is intuitively appealing, and has been widely taken as the definition of AI by the public. However, within the field most projects do not aim at pretending to be human beings. Actually, work on chatbot (which is the closest to the Turing Test among all subdomains of AI) had not been taken seriously by the mainstream until recent years, and the Turing Test has been criticized as a distraction or even harmful by some influential AI researchers (Hayes and Ford, 1995; Laird et al., 2009; Marcus, Rossi, and Veloso, 2016).

The most ironic point on this matter is that Turing himself did not propose this test (he called it the "imitation game") to be the definition of thinking machines, but a *sufficient condition* for a machine to be considered as intelligent. He explicitly acknowledged that it is not a *necessary condition* by writing "May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection" (Turing, 1950). His intention was merely to get a behavior-based standard for "thinking," though it does not have to be the only standard.

To expect an AI to behave exactly like a human is too anthropocentric to non-human intelligence, since human behaviors not only depend on our intellectual competence and mechanisms, but on biological, evolutionary, and cultural factors that are unique to humans. For example, we cannot expect an extraterrestrial being to pass the Turing Test, though we can expect them to be similar to us in certain aspects (Minsky, 1985b).

A milder version of this perspective aims at computer models whose behaviors are *similar to* that of human beings, though do not have to be *indistinguishable from* them. Such a project may take inspirations from psychology in its architecture or mechanisms (Newell, 1990; Franklin, 2007; Bach, 2009), or use psychological data to train a machine learning model to replicate the behavior (Flach, 2012). Though they usually do not aim at passing the Turing Test, these projects still use psychological data for evaluation and justification.

2.2.3 CAPABILITY-AI

For people whose interest in AI mainly comes from its potential applications, the intelligence of a system should be indicated by its problem-solving capability. For instance, Minsky uses the word "merely means what people usually mean—the ability to solve hard problems" (Minsky, 1985a). It certainly makes sense, as people do judge the intelligence of each other by their problem-solving capability. In the agent framework, it means that C is similar to H in the sense that there are moments i and j that:

$$p_i^C \approx p_j^H, \; a_i^C \approx a_j^H$$

that is, the action (solution) the computer produces for a percept (problem) is similar to the action produced by a human to a similar percept. To make the discussion simple, here we assume that a single percept can represent the problem, and a single action can represent the solution.[1]

In this way, the intelligence of a system is identified by a set of problems it can solve, while whether they are solved in the "human way" does not matter. Now the question is: which problems require intelligence, and which do not?

There have been various opinions:

> "I sometimes think of AI as 'the current frontier of computer science.' ...Then, in that view, AI is simply finding ways to make computers do the useful things that no one yet knows how to make them do" (Minsky, 1983).
>
> A useful definition of intelligence should "span an intellectual range from that of an insect to that of an Einstein, from that of a thermostat to that of the most sophisticated computer system that could ever be built" (Albus, 1991).
>
> "There are challenge problems in planning, e-commerce, knowledge discovery from databases, robotics, game playing, and numerous competitions in aspects of natural language" (Cohen, 2005).
>
> "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" (Nilsson, 2005).

[1] The same interpretation of percept-action as problem-solution can be made in the previous two perspectives, though it is not necessary as in those situations the whole I/O streams are similar to each other. It is only in this perspective that a single percept-action pair needs to make independent sense.

These problems indeed have practical values, but why do they need intelligence while others do not?

One extreme position is to consider every problem-solving process as requiring intelligence, though to different extents. As stated by Hayes and Ford (1995): "So, which parts of computer science are part of AI? We suggest a rather radical answer to this question: all of them." However, in that case why do we need a new concept? In recent years "AI" has been often used to refer to the works that were traditionally called "computer application" or "automation," but is this a good practice?

A less radical position is to define AI as using computers to solve problems that are only solvable by the human mind, as suggested by Minsky (1983). This treatment will indeed separate some problems from the others, but it leads to an iconic result: as soon as a computer system is built to solve a problem successfully, the problem is no longer "only solvable by the human mind," so does not need intelligence anymore. Consequently, "AI is whatever hasn't been done yet" (Hofstadter, 1979; Schank, 1991), which is known as "the AI Effect" (McCorduck, 2004).

Many researchers are not bothered by this situation. To them, as far as their results are valuable, whether they are labeled as "AI" does not matter. This attitude is partially responsible for the lack of a "theory of AI"—as the solving of different problems usually requires different theories and techniques, there is little hope to find a non-trivial foundation for all of them, and consequently, AI will not become a branch of science or engineering, but a "suitcase word" (Minsky, 2006) that has no core meaning.

Another issue of defining intelligence in this way is that it conflicts with an important aspect of the common usage of the word. Intuitively, intelligence is not associated with all types of problem-solving processes. Many people still have the feeling that although today's ordinary computer systems can solve many problems, the way they do so is too rigid and inflexible to be considered as intelligent. For example, people usually do not consider solving a problem by exhaustively considering each possibility to be intelligent, even though this method solves many problems perfectly. The attribute "intelligent" is intuitively associated with "creative," "autonomous," "flexible," and so on. Such associations are dismissed by some researchers as unrealistic expectations, but it nevertheless reveals a mismatch between what is called "AI" inside the field and the public expectation and imagination on what the research should be about.

2.2.4 FUNCTION-AI

One common way to distinguish AI from the other branches of computer science is to associate this field with the cognitive functions identified in the human mind. Currently most textbooks of AI are organized in this way, with chapters on searching, reasoning, learning, planning, perceiving, acting, communicating, etc. (Luger, 2008; Russell and Norvig, 2010; Poole and Mackworth, 2017).

For each function, the typical treatment is to follow the computational paradigm: "a result in Artificial Intelligence consists of the isolation of a particular information processing problem, the formulation of a computational theory for it, the construction of an algorithm that implements it, and a practical demonstration that the algorithm is successful" (Marr, 1977).

In the agent framework, this "Function-AI" perspective takes C to be similar to H in the sense that there are moments i and j that:

$$a_i^C = f^C(p_i^C)\,,\ a_j^H = f^H(p_j^H)\,,\ f^C \approx f^H$$

that is, the function that maps a percept (input problem) into an action (output solution) in the computer is similar to that of a human. Here the function can correspond to searching, reasoning, learning, etc., and since the focus is on the functions (i.e., input-output mappings), the concrete input and output values of the two agents do not have to be similar to each other. Naturally, a system with higher intelligence should implement more such functions efficiently and use them in multiple domains.

Compared to Structure-AI and Behavior-AI discussed previously, this perspective of intelligence is less anthropocentric (though the functions are still abstracted from the human mind), and it gives the field a better identity than Capability-AI by generalizing its capability from specific problems to abstract functions.

One issue of this perspective is the fragmentation of AI (Brachman, 2006) that has been addressed previously. Since each function can be specified in isolation, there is little motivation to take the other functions into consideration, as this will complicate the situation, and may violate the basic assumptions shared by the researchers working on the function.

Another issue is that when a function is specified in this way, it may become very different from its *natural* form in the human mind where it is tightly coupled with the other cognitive processes. One example is that in the current machine learning studies, “learning” has been commonly specified as the process of using a meta-algorithm (learning algorithm) to produce an object-level algorithm (model for a domain problem) according to the training data (Flach, 2012). This working definition is exact and simple, as well as fruitful in many domains, though is arguably only a restricted version if compared to the learning processes in the human mind (Wang and Li, 2016), even compared to the initial diverse approaches within the field (Michalski, Carbonell, and Mitchell, 1984).

These issues have been widely recognized within the field, as shown by the calls for integration (Brachman, 2006), the notion of “AI Complete Tasks” that stresses the dependency among the functions (Shapiro, 1992), and the attempts to organize the functions into a cognitive architecture (Newell, 1990) or agent framework (Nilsson, 1998). Even so, the problem is still far from being solved, mainly because the functions have been specified and developed according to different, even incompatible, assumptions and considerations, and therefore cannot be easily combined. This theoretical incommensurability (Kuhn, 1970) has important practical consequences, as revealed by the attempts of building integrated AI systems, where “component development is crucial; connecting the components is more crucial” (Roland and Shiman, 2002), since the difficulties are mainly theoretical, not technical.

2.2.5 PRINCIPLE-AI

As in any field, there are researchers in AI trying to find fundamental principles that can uniformly explain the relevant phenomena. Here the idea comes from the usage of intelligence as a form of *rationality* (Simon, 1957; Russell, 1997; Hutter, 2005; Wang, 2011) that can make the best-possible decision in various situations, according to the experience or history of the system.

In the agent framework, it means that C is similar to H in the sense that

$$A^C = F^C(P^C)\,,\ A^H = F^H(P^H)\,,\ F^C \approx F^H$$

that is, the function that maps the whole stream of percepts (experience) into the whole stream of actions (behavior) in the computer is similar to that of a human. As in Function-AI, here the focus is on the function, not the actual percepts and actions, except that the input-output relationship in Principle-AI is more general. The above F is often not formally specified, but described informally as a certain "principle," which is not merely about a single type of problem and its solution, but about the agent's life-long history in various situations, when dealing with various types of problems.

This position is widely doubted and sometimes criticized as "physics envy," as the phenomena associated with intelligence seem too complicated and heterogeneous to get a "neat" explanation (Minsky, 1990). Until a system built according to such a definition is widely acknowledged as intelligent, most people will not be convinced that a good definition of AI can be obtained in this way.

2.2.6 RELATIONS AMONG THE PERSPECTIVES

It is well-know that the AI researchers have been taking different approaches, though these approaches have been classified differently. For example, Russell and Norvig (2010) clusters the definitions of AI using a "humanly vs. rationally" distinction and an "acting vs. thinking" distinction. I do not make those distinctions, because according to the above analysis, all the AI definitions can be seen as abstractions of human intelligence, with rationality corresponding to a certain level. On the other hand, it can be argued that every AI system has both a *thinking* aspect and an *acting* aspect, corresponding to their internal and external activities, respectively.

Beside properly classifying different perspectives in defining AI, an important issue is the relationship among the perspectives. Instead of discussing how one of them relates to another one pair-by-pair, here I focus on the overall relation among them, as it is more crucial in this discussion.

The most common opinions on this matter can be expressed by the proverb "All roads lead to Rome" and the parable of "the blind men and an elephant," respectively. According to the former, all the approaches eventually lead to the same goal, and their differences are merely caused by the paths taken. Such a picture is described by Nils Nilsson in his talk "Routes to the Summit" in the "AI@50" conference,[2] in which he took different approaches as starting from different "base camps." According to the latter, each approach only addresses part of the picture, and eventually they should be combined together to get a whole solution, and probably can be organized into an "atlas of intelligence" (Bhatnagar et al., 2018). Both metaphors consider the approaches as complements of each other, though in different ways. According to the elephant metaphor, even one path leads to the summit, it is still only part of the story, as the objective is to explore the whole mountain, not merely the summit.

Though these opinions are not completely groundless, I think they get the situation wrong, and a more suitable metaphor is the Mount Lu in the poem of Su Shi (1037–1101, also known as Su Dongpo): "Viewed horizontally a range; a cliff from the side; It differs as we move high or low, or far or nearby."[3] Here Mount Lu is not the elephant described by the blind men, as the range and the cliff are not *parts* but *views* of the mountain. It is true that the previously mentioned structure, behavior, capability, function, and principle are all features of *human intelligence*, but abstractions according to each of them lead to different notions of *intelligence*, and the *artificial intelligence* systems designed accordingly are even more different, since these concepts often (though not always) require

[2] The conference website is at `http://www.dartmouth.edu/~ai50/homepage.html`, in which Nilsson's talk was listed.

[3] The English translation is from Zhang Longxi, "Lessons from Mount Lu: China and cross-cultural understanding", Cultural Dynamics, Vol. 27(2), 285–293, 2015.

different design decisions, so it is impossible for all of them to be satisfied to the same extent in a single computer system.

It is possible for an AI project to aim at more than one research objective. For example, when working on a model of the mind, it will be nice if some results can find practical applications; when the direct goal is to solve a real-life problem, it may be a good idea to study how it is handled by the human mind. However, there should be an objective to be considered as primary, otherwise the project would suffer from the conflicts among the objectives.

Because each definition sets a separate objective, the research paradigms established accordingly are not compatible with each other. In particular, the achieving or progressing toward one of them does not necessarily imply the same effect for another one. For example, a common belief is that brain modeling is a more fundamental approach, because as soon as the human brain is accurately simulated, human behaviors and so on will appear as consequences. This is not necessarily true, because human behaviors are not only determined by the human brain, but also by the human body and human experience, to say the least. To simulate all of those will not only be a technical challenge, but also different from the AI as we know it.

Therefore, accurately speaking, these perspectives should be considered as different research fields, though with overlapping parts here or there, and all called "AI" for historical reasons. Using Nilsson's metaphor, AI researchers are actually climbing different mountains, and will eventually reach different summits, even though they may take shared paths in certain ranges, and can learn from each other.

2.3 The ultimate aim of AI

The different working definitions of AI correspond to not only different ways to abstract from human intelligence, but also different expectations about the destination of this research.

In the early years of AI research, the works were clearly targeted at computers that are generally comparable with the human mind (Turing, 1950; McCarthy et al., 1955; Feigenbaum and Feldman, 1963). There were ambitious projects like General Problem Solver (Newell and Simon, 1963), the Fifth-Generation Computer Systems (Feigenbaum and McCorduck, 1983), and the Strategic Computing Program (Roland and Shiman, 2002), but none of them reached their declared goal, which led to widespread doubt about the feasibility of the "grand dream of AI," and contributed to the following "AI Winter."

Driven by motivations including to avoid the impossible missions, to obtain the necessary resources, and to improve its public image, the AI community shifted its aim to more realistic tasks, like solving practical problems and carrying out individual cognitive functions. This shift is praised as "AI becomes a science (1987 – present)" (Russell and Norvig, 2002), which was later changed to "AI adopts the scientific method (1987 – present)" (Russell and Norvig, 2010), because "It is now more common to build on existing theories than to propose brand-new ones, to base claims on rigorous theorems or hard experimental evidence rather than on intuition, and to show relevance to real-world applications rather than toy examples" (Russell and Norvig, 2010). It sounds wonderful, but as it turned out this has come at the price of giving up the initial dream of the field. For a long time afterward, topics like "general-purpose intelligence" and "thinking machine" became taboos, and were judged as not serious or even as pseudoscience. The aim of AI had been degraded to the building of "smart tools" (Nilsson, 2009). For people who still believe in the original dream, this goal shifting is disappointing. In an interview with Wired in August 2003, Minsky criticized AI as "brain-dead," as "Only a small community has concentrated on general intelligence."[4]

In recent years, a renaissance has been happening in AI, partly due to the hope raised by the success of new techniques such as deep learning, and partly due to the realization that the old problems cannot be sidestepped. To distinguish these types of research from the conventional works, new names have been introduced, including "Human-level AI" (Minsky, Singh, and Sloman, 2004; Nilsson, 2005; McCarthy, 2007) and "Artificial General Intelligence" (Goertzel and Pennachin, 2007; Wang and Goertzel, 2007). Some people also call this type of work "Strong AI", though it was not Searle's original meaning of the term (Searle, 1980). These new labels carry a common message: as "AI" no longer means what it used to mean within mainstream AI, a new brand is in need, even though what it refers to is not really new.

Though none of the new names has a commonly accepted working definition, each choice of word adding into AI does have intuitive implications and associations. The addition of "human-level" suggests that the current AI is inferior to that of humans; "general" suggests that mainstream AI is special purpose; and "strong" suggests that the conventional AI is weak. Though all of these feelings are justifiable, they provide different reasons when departing from mainstream AI.

Intelligence is widely taken as having different degrees or levels. For instance, most people could agree that though certain animals should be considered as intelligent, they are at lower levels in the "ladder of intelligence" compared to the "human-level," though such an idea is not without controversy (McDermott, 2007). Human-level AI and Strong AI are usually understood as more advanced than conventional AI, though they are all assumed to be moving in the same direction.

Artificial General Intelligence (AGI) could also be interpreted in this way, if a general-purpose system is nothing but many special-purpose systems combined together. However, many AGI researchers share the belief that general-purpose systems and special-purpose systems have certain fundamental differences that are qualitative, not quantitative (Wang and Goertzel, 2007; Goertzel, 2014). AGI research may produce results that look like the results of conventional AI, but they will be by-products, as the research is not aimed at domain-specific solutions. Compared to the works under the names of "Human-level AI" and "Strong AI," the paths explored in the AGI community are generally more unorthodox, since conventional AI is disproved here not because it has not moved far enough, but because it has been moving in a direction that is different from that of AGI. This dissatisfaction of mainstream AI is not enough to give AGI a common working definition. Actually, within the current AGI domain, the previous perspectives of AI can still be recognized. For instance, some projects attempt to achieve AGI by using an architecture that integrates various cognitive functions (Franklin, 2007; Goertzel, 2009), while some others by following a single rational principle (Hutter, 2005; Wang, 2006b), though the researchers on each side of this distinction still have different architectures and principles in mind.

Besides the attempts to restore the original aim of AI (though under new names), there are also speculations like "singularity" (Kurzweil, 2006) or "superintelligence" (Bostrom, 2014) that set the aims even higher. I do not consider any of these concepts well-conceived, so will not analyze them in this paper. My arguments against them can be found in (Wang, Liu, and Dougherty, 2018).

[4] The interview was accessed at `https://www.wired.com/2003/08/why-a-i-is-brain-dead/` in March 20, 2019.

3. My Definition of Intelligence

3.1 Intuitions and motivations

My own opinion about the aim of AI started from the vague feeling that traditional computational systems are based on a design principle that makes them very different from the human mind, and that this principle can explain many other differences between the machine and the mind: A program is traditionally designed to do something in a predetermined *correct* way, while the mind is constructed to *do its best* using whatever it has. Consequently, absolute correctness or optimality of solutions should not be used as the design criteria for a mind-like system, though it is still possible to talk about what is the right thing to do in each situation, and there are guiding principles across situations.

This opinion is obviously not novel—the ideas from the previously mentioned Principle-AI perspective all come from similar intuitions. The real challenge is to turn this opinion into a working definition to guide research, satisfying the previous mentioned requirements for good definitions as much as possible.

To be similar to the common usage of the word, intelligence should take the human mind as the most typical example, while still leave room for various types of non-human intelligence. It means the definition should not be based on human-specific features, nor to demand them to be emulated in detail, otherwise it would be a definition of *human intelligence*, rather than its generalization. In this aspect, it shares motivation with the "universal intelligence" of Hutter (2005), though does not move away from human intelligence as far as Hutter's—by his definition an ordinary human is not very intelligent.

On the other hand, the definition cannot be so broad that the traditional computers are already included as intelligent (though maybe at a lower level). Besides being counter-intuitive, such a definition would be trivial or redundant, as it introduces no new insight into research. Despite their problem-solving capabilities, traditional computer systems should be taken as unintelligent, as they are designed according to principles that are fundamentally different from what we call intelligence. From a theoretical point of view, AI should not be considered as the same as computer science, or a part of it. Though AI will eventually be implemented in computer systems, AI systems should show fundamental differences when compared with the traditional systems, rather than merely being able to solve more problems. Intelligence should demand a different way to design and to use computers, compared to the traditional way, which is already captured by the definition of computation.

In computer science, "computation" does not mean whatever a computer does, but is accurately defined as a finite and repeatable process that carries out a predetermined algorithm to realize a function that maps input data to output data (Hopcroft, Motwani, and Ullman, 2007). As summarized by Marr (1982), to solve a problem "by computation" means:

1. To define the problem as a mapping from a domain of valid input values to a range of possible output values;

2. To find an algorithm that carries out this mapping step by step, starting from the given input and ending with the corresponding output;

3. To implement the algorithm in a computer system so as to use it to solve each instance of the problem.

However, this approach cannot handle a problem if any of the following is the case:

- The problem is not well-defined as a mapping or function;
- The function is well-defined, but the system has no algorithm to solve it, due to either its ignorance, or the impossibility of such an algorithm;
- There are implemented algorithms for the problem, but the system cannot afford the resources (mainly computational time and space) to use any of them on the problem instances to be processed.

It is not hard to recognize that many problems handled by the human mind have these issues. We cannot solve them perfectly, though we still survived and live reasonably well. Isn't this ability what intelligence is about? Why can we not make computers to do the same?

It is right to say that the intelligence of a system is eventually displayed in its problem-solving capabilities. However, to me intelligence is more like the flexible, versatile, and unified "hands" that can use the efficient-but-rigid "tools" provided by the various hardware and software, rather than a "toolbox" that contains certain problem-specific capabilities (Wang, 2004c).

The above deliberation suggests that intelligence corresponds to a *working condition* and a *coping strategy* that are both different from those of computation. This difference is what my working definition of intelligence stresses.

3.2 Description

Here is the working definition proposed in Wang (1995):

> "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

Since a working definition of a new concept must depend on some existing concepts (otherwise it would cause an infinite regression), in this definition "information-processing system" is used in a broad sense of the phrase to include all computer systems and robotic devices, as well as many animals, though it will not include everything, such as rocks and rivers. Such a system can be described as being driven by some *tasks* or *problems* (though the system does not need to know or understand them "consciously"), and can carry out or solve them (at least some of them to certain degrees) by taking some *actions*. The internal relations between the tasks and actions can be called the system's *knowledge* or *beliefs*. The system's information-processing activities are basically to choose and carry out proper actions to accomplish the existing tasks, and these activities cost computational *resources*, mainly the time for computation and action and the storage space for tasks and knowledge.

Among all information-processing systems, the above definition draws a line between the *intelligent* ones and the *unintelligent* ones, according to two features that specify the system's working condition and its coping strategy.

Here "with insufficient knowledge and resources" specifies the normal working condition of the system, and is with respect to the concrete problems the system must deal with. In this context, "having sufficient knowledge" means the system knows the proper algorithm of solving the problem; "having sufficient resources" means the system can afford the resources (chiefly time) required by this algorithm when applying it on the instances of the problem. On the contrary, "with insufficient knowledge and resources" can be further divided into three restrictions or requirements, that is, the system must be *finite* and *open*, and operate in *real-time*.

Being *finite* means the system's information processing capability (such as how fast it can run and how much it can remember) is roughly a constant at any period of time. This requirement seems trivial, as any concrete system is surely finite. However, to acknowledge the finite nature means the system should manage its own resources, rather than merely spending them. This requirement is mainly for the theoretical models, since most traditional ones, like the Turing Machine, completely ignore resource restrictions (Hopcroft, Motwani, and Ullman, 2007).

Being *open* to new tasks means to make no restriction on the *content* of a task, as long as it is expressed in an acceptable *form*. Every system surely has limitations in the signals its sensorimotor mechanism can recognize or the languages its linguistic competence can handle, but there cannot be any restriction on what it can be told or asked to do. Even if a new experience conflicts with its current beliefs, or a new task is beyond its current capability, the system still should handle them properly, rather than simply ignore it or even be crushed by it.

For the system to live and work in *real-time* means that new tasks of various types may show up at any moment, rather than come only when the system is idly waiting for them. It also means that every task has a response time restriction, which may be in the form of an absolute deadline, or as a relatively expressed time pressure, such as "as soon as possible." In general, we can consider the utility of a solution to be a decreasing function of time, so even a correct solution may become worthless when produced too late. A real-time system must respond according to these restrictions, even when they cannot have been completely satisfied.

The above three assumptions are collectively called the *Assumption of Insufficient Knowledge and Resources* (AIKR), which identifies the normal working condition of an intelligent system. Of course, this assumption is about the overall situation, not on every task, as there are surely simple tasks for which the system's knowledge and resources are relatively sufficient, at least roughly speaking. However, such tasks are not where intelligence is really demanded.

People may argue that as far as a task is processed to the system's satisfaction, it must already have the knowledge and resources from which the solution is produced. This is not really an argument against AIKR, because the insufficiency occurs at the moment when the system starts to process a task, rather than after it has been processed; moreover, the assumption is more about the overall mechanism than about the processing of an individual task. It means that the system cannot depend on predetermined algorithms and procedures to process its tasks, as there is always missing or uncertain knowledge, and it does not have the time to consider every possibility when processing a task.

Furthermore, the environment may change constantly, so every belief the system has at a moment may be challenged by new experience. AIKR does not allow the future to be taken as repeatable (like in games such as Chess and Go), or as *statistically stationary* where every event has a fixed probability (which may be unknown). Consequently, many decision-making strategies loss their validity when applied in this situation.

A direct implication of AIKR is that there cannot be absolutely correct or optimal solutions. As the system is open to an unrestricted future, no prediction can have guaranteed confirmation from future observations, no matter how well it has been supported by past experiences; as the system works under a constant time pressure, it has to omit certain possibilities (even though they are known to be relevant) when processing a task, so there is always a risk of missing a better solution related to a neglected possibility.

However, it does not follow that under AIKR all strategies are equally good (or equally bad). This is where the second point in the working definition, *adaptation*, comes into play. In this context, this term refers to the mechanism for a system to summarize its *past* experience to predict the *future* situations accordingly, and to allocate its *bounded* resources to meet the *unbounded* demands. As new experience becomes available, it will be absorbed into the system's knowledge and beliefs, so as to put its solutions and decisions on a more stable foundation.

Though adaptation is a well-known concept and is often associated with intelligence, its usage here still contains certain subtle points:

- I consider intelligence as an advanced form of adaptation, which happens within the lifetime of a single system, and the changes it produces depend on the system's past experience. Therefore, it is different from the adaptation realized via evolution in a species, where the changes happen in an experience-independent manner, then selectively kept according to the future experience, as in evolutionary computation (Holland, 1992). There are surely important similarities between these two forms of adaptation, though their differences are also important.

- Though this individual-level experience-driven adaptation is often referred to as "learning," what my definition proposes is very different from mainstream machine learning algorithms at the present time (Flach, 2012; LeCun, Bengio, and Hinton, 2015; Domingos, 2018), where a learning process is defined as approximating an input-output mapping by generalizing training samples. The adaptation process in my definition is lifelong, cumulative, open-ended, multi-objective, and does not necessarily converge (Wang and Li, 2016; Thórisson et al., 2019).

- Adaptation means not only changing the system itself to meet the restrictions of the environment, but also changing the environment to meet the system's desires. This is very different from some other models, such as AIXI (Hutter, 2005), where the environment is assumed to be an unknown Turing Machine that cannot be changed by the system's actions.

- In my definition, adaptation refers to the attempt or effort, not the consequence. The system adjusts its behavior according to its past experience, but that will improve the system's performance only when the future is similar to the past in the relevant aspects. Under AIKR, such a similarity can be *hypothesized* to guide the system's decisions, but cannot be *guaranteed* to be infallible, as argued by Hume (1748). When the future turns out to be very different from the past, the system's adjustments will fail to meet its anticipation, and may even make things worse. However, even in this situation, the adjustments are still considered as adaptive. In this context, whether an adjustment is *adaptive* is judged according to the system's past experience, rather than its future experience. Therefore, according to my definition, being intelligent does not mean to be successful all the time, not even with a promised success rate. Instead, it means the system is using its available knowledge and resources to the maximum allowed by the restrictions at the moment, in light of AIKR.

AIKR and adaptation are closely related to each other. A system has the needs to adapt only when it has insufficiency in knowledge or resources; on the other hand, acknowledging the insufficiency but making no attempt to improve the situation is effectively equivalent to denying the insufficiency. Together this working condition and coping strategy consist of a *relative rationality* (Wang, 2011), meaning to make the best effort in light of the available knowledge and resources.

According to this definition, the opposite of intelligence is not "cannot solve any problem," but "having a constant and invariant ability," which corresponds to the notion of "computation" in computer science. In this way, *intelligence* is defined as a strategy of problem-solving that is fundamentally different from *computation*, as suggested in the previous section. However, this does not inhibit this working definition from being used to guide the design of an AI system that is implemented in a computer.

3.3 Implications

The above working definition has many implications. Here I briefly introduce the guidance it provides in the design of NARS (Non-Axiomatic Reasoning System) (Wang, 1995, 2006b, 2013).

As explained above, an intelligent system defined in this way cannot always solve problems by following problem-specific algorithms, since according to AIKR, such algorithms are not always available or affordable. On the other hand, a computer system eventually runs according to algorithms. The solution of this dilemma is to combine algorithm-specified steps to handle each problem-instance in a *case-by-case* manner (Wang, 2009). As the system is adaptive, its internal state changes without repeating any previous state, and so does the environment. Consequently, the problem-solving processes are no longer accurately repeatable, and there may not be an algorithm for the solving of a (type of) problem (which is a set of problem-instances). The actual processing of a problem-instance can still be recorded and considered as an "algorithm" afterward, but since the same process may not occur again, such a conceptualization makes no contribution to the system, nor to its designers.

Therefore, the design of such a system cannot focus on the algorithms for specific problems anymore as suggested by Marr (1982). Instead, it should focus on the design of the algorithmic steps as the building blocks of problem-solving processes, as well as on the mechanism to combine these steps at run time for each individual problem-instance. Both above tasks are independent of the application domain and the specific features of the problems in the domain.

This situation naturally suggests the system to be designed in the framework of a *reasoning system*, interpreted broadly. Such a system is equipped with a set of *inference rules*, each of them is specified and justified in a domain-independent manner, and these rules can be combined into *inference processes* in a flexible manner to handle various tasks. Such a system is often considered as implementing a *logic*, which specifies a knowledge representation format (often using a formal grammar), as well as the valid operations on the representation (often using formal rules). Beside the logic part, the system also needs a control part to manage the memory and to select the rule and the premises for each inference step.

Logic-based AI has been proposed and pursued for a long time and by many researchers (Hayes, 1977; McCarthy, 1988; Nilsson, 1991), though it has also been widely criticized (Hofstadter, 1985; McDermott, 1987; Birnbaum, 1991) and has become much less popular in recent years. To me, the notions of *logic* and *reasoning* are still productive in AI, although the concrete logic or reasoning systems built in the past for AI are not based on the proper assumptions. According to AIKR, an intelligent system cannot derive new truth (theorems) from given truth (axioms) anymore, even if "true" is relaxed into "probably true" (Nilsson, 1986; Adams, 1998). Instead, the validity of reasoning has to be justified as a form of adaptation, which leads to defining "truth-value" as the degree of evidential support (Wang, 2005), which is based on the past, though used for the future.

Though the truth-value of NARS is intuitively similar to probability, in principle it is a different measurement, as it does not follow the axioms of probability theory, by which the probability of an event (or statement) is a single number. Since the truth-value of NARS is experience-grounded, it may change as new experience comes in a way that cannot be properly handled by Bayesian conditioning, Jeffrey's rule, or other methods from probability theory (Wang, 2004a). This is the case partly because under AIKR, the consistency among beliefs cannot be guaranteed, though the system makes efforts to reduce the inconsistency by revising its beliefs.

This experience-grounded truth-value does not fit predicate logic well, so NARS comes with a new logic, Non-Axiomatic Logic (NAL), that is designed as a term logic, in which multiple types of inference are unified (both in format and in semantics), including *deduction, induction, abduction, revision, choice, comparison, analogy*, etc., in the tradition of Aristotle (1882) and Peirce (1931), though the technical details are very different from these classic models. NAL also shares common features with set theory, propositional logic, predicate logic, non-monotonic logic, and fuzzy logic, but still differs from them fundamentally, as none of them is designed for adaptation under AIKR. For comparisons between NAL and those formal models, see Wang (2006b, 2013).

To cover various cognitive functions, in NARS the reasoning framework is extended to cover "practical reasoning," which is the traditional name for reasoning on actions and goals, or reasoning about "what to do " (which is different from the reasoning on "what is true" or "what to believe"). The approach taken by NAL is inspired by logic programming (Kowalski, 1979), where goals and actions are expressed by statements with special interpretations. Consequently, various cognitive functions, including *learning, planning, searching, categorizing, observing, acting, communicating*, etc., become different aspects of the same underlying process in NARS (Wang, 2006b, 2013). These processes are all formulated according to the *adaptation under AIKR* principle, and only try to produce the optimal solution with respect to the currently available knowledge and resources, so quite different from how each of them is specified and carried out in other AI techniques.

As NARS usually processes many tasks in parallel, and new tasks come constantly both from outside (observation and communication) and inside (reasoning and learning), under AIKR it is impossible to process each of them completely and exhaustively. Instead, the system distributes its resources among the tasks, biased by their priority values evaluated according to the system's beliefs.

For each task, its processing path depends on the related beliefs that are selected at the moment, also according to the experience of the system. Consequently, even when a task is repeated, its processing path and results may be different, as the context has changed. That is why it has been claimed previously that at the level of problem solving, the system's input-output relation is not a fixed mapping, and the process does not follow a predetermined algorithm (not even a randomized algorithm). For this reason, NARS is a "Constructivist AI" (Thórisson, 2012). Many phenomena come from this dynamic resource allocation mechanism altogether, without being simulated one-by-one: *attention, forgetting, association, activation spreading*, etc.

This article is not intended to serve as an introduction to NARS[5], and the above descriptions are used only to show that the working definition of intelligence given previously does serve as the cornerstone for the design of an A(G)I system by supporting and restricting its major design decisions. Technical differences between NARS and the other AI projects can mostly be traced back to their different design principles, which in turn come from their different understandings, or working definitions, of intelligence.

[5] Many publications on NARS can be accessed at `https://cis.temple.edu/~pwang/papers.html`, and its open-source implementation is at `http://opennars.org/`.

4. Comparison with Other Definitions

4.1 With other rational principles

According to the classification used in Section 2.2, my working definition of intelligence belongs to Principle-AI. Therefore, it is compared to the other opinions under that category first.

The definition proposed above and the associated *relative rationality* (Wang, 2011) are clearly influenced by the *bounded rationality* of Simon (1957). "Within the behavioral model of bounded rationality, one doesn't have to make choices that are infinitely deep in time, that encompass the whole range of human values, and in which each problem is interconnected with all the other problems in the world" (Simon, 1983).

In spite of this fundamental similarity, there are still major differences between my position and Simon's on this matter.

First, "insufficient" is more restrictive than "bounded" or "limited." Even bounded knowledge and resources may still be sufficient to solve certain problems, so a trivial strategy to work with bounded rationality is to only accept tasks that fall within the range of the system's capability. On the contrary, AIKR will not allow such a strategy. Moreover, bounded rationality roughly corresponds to the *finite* requirement within AIKR, while not requiring the system to be open to novel tasks (though it assumes incomplete knowledge), or to work in real time (though it assumes limited time). Therefore, AIKR is not implied by bounded rationality.

Actually, Simon did not explicitly use bounded rationality to define *AI*, but applied it mainly in the explanation of *human* behaviors. In the AI projects he participated, GPS (Newell and Simon, 1963) and BACON (Simon, Langley, and Bradshaw, 1981), the restriction of knowledge and resources was taken into consideration in certain aspects (such as using heuristics to get satisficing solutions), but not in the sense of AIKR. For instance, none of these systems works in real time (as specified in Section 3.2), and in heuristic search the learning and revising of heuristic functions were not considered.

Russell and Wefald's *limited rationality* also moved in the same direction by stating that "Intelligence was intimately linked to the ability to succeed as far as possible given one's limited computational and informational resources" (Russell and Wefald, 1991). In Russell (1997), several types of rationality are compared, and it is argued that the closest to the needs of AI is *Bounded Optimality*, the capacity to generate maximally successful behavior given the available information and computational resources.

The work of Russell and his colleagues had gone beyond Simon's, as it provided formal specifications of the concepts proposed in the AI context. However, since its formal specification is based on decision theory and computational complexity theory, it is still not under AIKR. For example, a system that accept AIKR cannot solve problems merely by selecting one program from a given set of programs, but has to create programs from existing components under a variable and potentially unpredictable time pressure. Furthermore, as the problem-solving processes do not accurately repeat, they cannot be analyzed using computational complexity theory.

A more recent definition from Legg and Hutter (2007) states that "Intelligence measures an agents ability to achieve goals in a wide range of environments." This definition is formalized in the reinforcement learning framework, where "All tasks that require intelligence to be solved can naturally be formulated as a maximization of some expected utility in the framework of agents" as shown in the AIXI model of "universal intelligence" (Hutter, 2005). The model is based on the assumption that the true environment can be described by a computable probability distribution over

Turing Machines that is unknown to the agent, while the intelligence of the agent is indicated by its ability to maximize the expected reward according to the observation so far. Among the predictions consistent with the observation, the simplest one is favored as more likely to be true.

Though AIXI shares certain intuition with NARS, the assumptions behind these two projects about the environment and the agent are fundamentally different.

To take the environment as a computable probability distribution over Turing Machines means that whatever the agent does, the actions can only change the rewards it gets, but cannot change the environment. It is a strong postulation about the relation between an agent and its environment, which is only justified as "in standard physics there is no law of the universe that is not computable in the above sense" (Legg and Hutter, 2007). The problem about this justification is that the descriptions about the world provided in classical physics should not be equated with the world itself, and this reductionist position denies the need for generalization and abstraction that describe the environment at multiple levels, where the descriptions are not exactly the same.

For example, even though it makes sense to see human perception as compression, it is surely not lossless compression. Though the definition of intelligence in Legg and Hutter (2007) is presented as a generalization of many psychological definitions of intelligence, at least on this point it is not supported by common psychological theories. Even from a normative point of view, to assume that AI should eliminate hypotheses that are not perfectly consistent with the observations would decline many meaningful generalizations.

Another major issue is that AIXI assumes infinite computational resources, and Legg and Hutter (2007) explicitly stated that "We consider the addition of resource limitations to the definition of intelligence to be either superfluous, or wrong. . . . Normally we do not judge the intelligence of something relative to the resources it uses." As far as I know, all tests of intelligence have explicit or implicit time limits, and people usually do not take "to exhaustively evaluate all alternatives and pick the best" as an intelligent way of solving a problem, but that is basically what AIXI does.

To be clear, I am not claiming the Legg-Hutter definition of intelligence to be *wrong*, but at least *different* from mine and many others. NARS and AIXI target fundamentally different problems, though both are associated with the notion of intelligence. The *exactness* and *simplicity* of Legg and Hutter's definition are admirable, but it does not mean that the issues in *similarity* (to the common usage) and *fruitfulness* (in guiding the research) can be ignored —After all, an AI theory inevitably contains empirical contents and engineering implications, so it should not be evaluated as a mathematical theory. Even though AIXI has been revised and approximated in various ways to become computable and implementable, which has produced some interesting results (Beattie et al., 2016), the above fundamental issues in the model are still there, which limit what can be obtained from this work.

4.2 With other perspectives of AI

4.2.1 WITH STRUCTURE-AI

In the design of NARS, no explicit attempt has been made to simulate the brain structure, either at the whole brain scale or as a neural network. This decision does not come from considerations on *usefulness* (brain models will contribute greatly to neuroscience), *possibility* (the model will surely be increasingly accurate), and *difficulty* (a scientific exploration should not be abandoned just because it is hard!), but on *generality* and *necessity*.

As argued previously, as far as we agree that "brain" and "mind" are both meaningful concepts, and human intelligence is a form of intelligence, there is no necessary reason to insist that the general notion of intelligence has to be reduced to the specific form of human intelligence, or that a (functional) mind can only be produced by a (biological) brain, especially when the notion is used to describe or construct non-human systems. This is especially the case in computer systems, where the underlying physical processes are very different from biological systems, not to mention the motivational and environmental factors.

That said, the design of NARS does get inspiration from the human brain, and there are similarities between NARS and brain models, both in mechanism and in behavior. For example, it is shown in (Wang and Hammer, 2015) that the induction rule and comparison rule of NARS are similar to the Hebbian rule in that repeated occurrences lead to substitutability between concepts (including percepts and actions), and when temporal information is added, the reasoning process can model classical conditioning and causal inference, where causality is taken to be "a property of mind, not matter", as argued by Freeman (1999). NARS also utilizes a forgetting mechanism to deal with the insufficiency of time and space, which is similar to the human memory in spirit, though not in details (Wang, 2004b).

There are reasons to believe that all types of intelligence share certain structural features, no matter whether they are biological, electronic, or something else. Even so, *to be structured as faithfully to the human brain as possible* is not the objective for a system like NARS, because this requirement contains irrelevant factors with respect to what such a system aims at. Even though the human brain indeed produces the best-known form of intelligence, we should not let it limit our imagination when designing an electronic form of intelligence, because intelligence is not fundamentally biological.

4.2.2 WITH BEHAVIOR-AI

For a similar reason, NARS is not designed to replicate human behaviors to the extent that will allow it to pass the Turing Test, because passing such a test is not a *necessary condition* of intelligence, though it may be a *sufficient condition*. Turing was not wrong, literally speaking, but a little misleading by stressing the behavioral indistinguishability between a thinking machine and a human being, and his proposal has been misunderstood by many people.

According to my working definition of intelligence, the indistinguishability between a human mind and a thinking machine should be in the *relationship between behaviors and experience*, rather than on specific behaviors. An AI that does not have human experience would not behave like a human, because its behaviors should depend on the experience of itself, not that of others. It may be able to successfully pretend to be a human, but that should not be the only way to show its intelligence. If a robot has sensors and actuators that are completely different from that of a human being, there is no way for it to behave as a human. However, it does not mean that it cannot be fully intelligent, even judged by a human being.

Once again, this position does not prevent NARS from showing human-like behaviors here or there, because it is designed according to similar restrictions as imposed on the human mind when it evolved, and it should not be a surprise that the strategies they acquired are similar, though not identical in details. For instance, several "human biases and fallacies" are reproduced in NARS, and justified as inevitable consequences of adaptation under AIKR. These phenomena are often classified improperly as *irrational* (Wason and Johnson-Laird, 1972; Tversky and Kahneman, 1974).

In these studies, human behaviors are judged according to classical logic and probability theory as normative models, but these models are too idealized as they ignore the knowledge and resources restrictions in reasoning and decision making in everyday situations (Wang, 1996, 2001, 2011).[6] Therefore, though NARS is designed as a normative model of intelligence and cognition, it is closer to a descriptive model of human intelligence and cognition when compared with other normative models.

4.2.3 WITH CAPABILITY-AI

NARS is not designed to solve any specific practical problem. Instead, it aims at a theoretical (meta-level) problem: how can a system learn to solve problems beyond its current capability? My definition effectively takes "intelligence" as a meta-level solution, and accordingly, an intelligent system like NARS has little innate problem-solving capabilities, or skills, though is equipped with the potential to acquire such skills from its experience, as far as it is not living in an environment too chaotic or adversarial to be adapted to — there are inhospitable environments where the intelligence of a system cannot guarantee its success or even its survival.

There is a relatively sharp level-separation in NARS. The meta-level knowledge (including the grammar rules, inference rules, control routines, executable operations, etc.) is mostly built-in and independent of the system's experience, though there are some adjustable parameters; the object-level knowledge (including the system's concepts, beliefs, desires, goals, and skills composed recursively from the operations, etc.) is stored in the system's memory, mostly acquired from experience, and remains revisable all the time.

Therefore, what domain problems NARS can solve is mostly determined by its experience (its *nurture*), not by its design (its *nature*). For a specific application, NARS should not be the choice if the designer already has an efficient design. NARS provides a better solution only when AIKR has to be acknowledged, as the other techniques are inapplicable under this condition. In this way, NARS is not designed to compete with the problem-specific systems such as Deep Blue (Hsu, Campbell, and Hoane, 1995) and AlphaGo (Silver et al., 2016), but more like a "Child Machine" suggested by Turing (1950), that can be educated to become a practical problem-solver.

Using the hands-tools metaphor introduced previously, the best solution to any specific problem is usually provided by a specially designed tool, though it is not the reason to conclude that the hands are inferior to the tools, as they should not be evaluated according to their performance on any specific job, but the role they play in the system's lifelong history when facing various problems, many of which are unanticipated and unprecedented. Since computer science and information technology have handled the tool-building well, AI should focus on the hands.

4.2.4 WITH FUNCTION-AI

As mentioned previously, NARS uniformly realizes a large number of cognitive functions studied in AI, though not as separate computational processes, but as different aspects and facets of a single process (as the "range" and "cliff" in Su Shi's poem, being different perspectives of Mount Lu). Consequently, the exact form of each function is quite different in NARS compared to its conventional definition in the current AI community.

[6] I am not claiming that there is no irrationality at all, but that any model of rationality has its preconditions, and would lose its normative status when applied to scenarios where these preconditions are not satisfied.

For example, "learning" is usually specified as carried out by a learning algorithm, which takes some training data as input, and produces a model learned from the data. The model then is used as an algorithm to solve the domain problem (Flach, 2012). On the contrary, in NARS learning is achieved via self-organization, which happens in all aspects at the object-level, as mentioned previously and explained in (Wang and Li, 2016). As a result, NARS is not designed to compete with techniques like deep learning (LeCun, Bengio, and Hinton, 2015), but is more flexible for situations where AIKR has to be acknowledged, since deep learning and other learning algorithms are not easily applicable there—they have trouble to learn incrementally in real-time when data comes piece-by-piece, and the objective of learning is not a single input-output mapping.

The same applies to many other cognitive functions that are often included in the definitions of intelligence. For example, *creativity* is definitely expected from any truly intelligent system, but since it logically follows from being *open* (as the ability to handle a novel problem it never saw before), it is not explicitly mentioned in the definition.

Though it sounds natural to define intelligence as a collection of cognitive functions, such a definition encourages a divide-and-conquer methodology, which is partially responsible for the current fragmentation in AI. Though the techniques developed in this way have great theoretical and practical values, they are not easy to be combined together to form a thinking machine that is comparable to the human mind in general, unless there is a "master algorithm" behind them (Domingos, 2018). However, in that situation the intelligence is arguably in the master algorithm itself, like a general-purpose "hands" that can use the problem-specific "tools", using the hands-tools metaphor again. However, an AGI-aspiring system like NARS is not an "algorithm" even in the broad sense of the term, and the tools are not necessarily produced from NARS. Instead, NARS could be thought of as an intelligent operating system that decides when to use which hardware or software for the tasks the whole system is facing.

4.3 With other types of intelligence

Though the working definition of intelligence proposed in this article mainly comes from AI considerations, it nevertheless also covers other types of intelligence.

The systematic study of intelligence started in psychology, and there has been a huge literature on this topic (Gottfredson, 1997; Goldstein, Princiotta, and Naglieri, 2015). In general, my definition is compatible with many psychological definitions. For example, Piaget sees intelligence as "the most highly developed form of mental adaptation" (Piaget, 1960), and further stated that "Intelligence in action is, in effect, irreducible to everything that is not itself and, moreover, it appears as a total system of which one cannot conceive one part without bringing in all of it" (Piaget, 1963). Medin and Ross (1992) have even made the statement that "Much of intelligent behavior can be understood in terms of strategies for coping with too little information and too many possibilities." These citations show that the two major factors in my definition, adaptation and AIKR, are considered as central to intelligence by many psychologists.

Even so, on the surface my definition does look different from common definitions given by psychologists. For example, a widely accepted definition of intelligence is proposed in Gottfredson (1997):

> "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a

> narrow academic skill, or test-taking smarts. Rather it reflects a broader and deeper capability for comprehending our surroundings 'catching on,' 'making sense' of things, or 'figuring out' what to do."

I completely agree with it as a *description* of intelligence, though do not think it provides a better *definition* than mine. One reason is that the functions listed in it (reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience) are all derivable from my definition as necessary functions of an intelligent system[7], and list them in this way in a definition will suggest a divide-and-conquer strategy, which has the risk of neglecting the strong coupling among these functions. This is already an issue in psychology, and an even bigger one in AI, as discussed previously when Function-AI is analyzed.

In general, a more comprehensive description is not necessary a better working definition, as it may fail to pinpoint the essence of the concept, and distinguish it from the derived or secondary features. Consequently, such a definition cannot effectively guide the research, even though its content can be agreeable, if analyzed sentence by sentence.

A working definition is often introduced into a scientific theory to draw a line and to stress a difference, so it only mentions the most important factors in the distinction it makes. In different theories, even the same notion may be defined differently to serve different purposes. In psychology, the concept of "intelligence" was introduced and is used mainly to study the difference among the intellectual capabilities of human beings, so the attributes shared by all human beings are taken for granted, and do not need to be mentioned. AIKR is just such an attribute, as the human mind obviously works under these restrictions.

However, this is not necessarily the case anymore when the concept is extended to include non-humans, where the line to be drawn is between intelligent systems and unintelligent systems, and the differences and similarities to be addressed are between the human mind and the traditional computers, while interpersonal differences are almost negligible. This is why it is not a good idea for AI to directly use a psychological definition of intelligence, though it does not mean that the psychological definition is wrong or irrelevant. At least the working definition of intelligence accepted in AI should be compatible with the psychological definitions, as part of the "similarity to the explicandum" requirement introduced at the beginning of the article.

For this reason, "human intelligence," "artificial/computer/machine intelligence," and "intelligence" should be taken as three different concepts, with the last one to provide a proper generalization for the first two.

My definition of intelligence also covers "animal intelligence," "collective intelligence," and "extraterrestrial intelligence" as special types, though those types will not be discussed in this article. Here I just want to state that for the first two types, their similarity with human intelligence is mainly in their adaptive nature, while their concrete structure, behavior, capability, and function can be more or less different from that of human beings. Like the case of humans, every animal or group is restricted by AIKR, so it does not need to be stressed. As for extraterrestrial intelligence, my definition suggests recognizing such an entity by checking whether it can adapt to its environment, rather than by its similarity with human on other aspects.

[7] Every function in the list has been realized in NARS, at least in its preliminary form, though it cannot be discussed in this article.

4.4 Evaluation of intelligence

According to my definition, intelligence is still a matter of degree. No system can be "perfectly intelligent", though one system can be more intelligent than another by being able to acquire knowledge in more forms (e.g., additional sensorimotor channels), to reorganize its beliefs and skills in more complicated ways (e.g., more recognizable patterns), or to adapt more efficiently (e.g., faster responses). Of course, this comparability does not define a total order among the intelligence of systems, as one can be more intelligent than the other in a certain aspect, but less intelligent in another one. One implication is that it is not always meaningful to ask whether an AGI is more or less intelligent than a human, as they may follow the same principles while having distinct sensors and actuators, so cannot be compared directly in concrete capabilities.

To accurately define measurements of intelligence according to my working definition remains a research problem to be solved. Such measurements are theoretically possible, though, just like in the situation of human IQ tests, each measurement can only provide a certain perspective about the system's intelligence, rather than considering all relevant factors.

Though many measurements of intelligence have been studied (Hernández-Orallo, 2017), none of them is suitable for this job. In particular, it is unjustified to use human IQ tests to evaluate the intelligence level of computer systems, as the tests usually assume human experience to various extents, and most of them do not measure learning ability properly.

A proper measurement should not be about the system's problem-solving capability at a moment, but how the capability changes. That is, if the problem-solving capability of a system can be represented as a function of time, its level of intelligence is not indicated by the value of the function at a moment, but by its *derivative value* of the function, i.e., how fast it increases (Wang, Liu, and Dougherty, 2018). What makes the problem complicated is that the nature of the environment-system interaction (such as the complexity of the system's sensorimotor mechanism and linguistic competence) and the resources restriction should be considered, too.

5. Conclusion

Though it is unrealistic and unnecessary to require people to define every word they use, "intelligence" in the AI context does demand a more careful treatment than it has been given to date. Its working definition matters, since different choices lead the research in different directions, rather than merely use a term differently. The current field of AI is actually a mixture of multiple research fields, each with its own goal, methods, applicable situations, etc., and they are all called "AI" mainly for historical, rather than theoretical, reasons.

These fields are surely related, but currently the main danger is to overlook their fundamental differences and to indiscriminately refer to all of them as "AI." This practice not only causes a lot of confusion in theoretical discussions and design processes, but also has practical consequences even for the people who do not care about theory. This is the case because to answer any non-trivial question about AI, such as "Is AI possible?" "How to build an AI?" and "Will AI be beneficial?" the "AI" in the question must be defined or at least specified first, since different types of AI correspond to very different answers. For example, in the discussion on AI safety, at least we need to clearly separate the systems whose behaviors are completely determined in its design and development phrase (its *nature*) from those whose knowledge, including moral and ethical knowledge, mainly comes from its own experience after it starts to run (its *nurture*). These two types of AI cannot and should not be regulated in the same way.

According to this analysis, there is no *correct* working definition of AI, as each of them has theoretical and practical values, so is not *wrong*. However, all working definitions are not *equally good* when judged according to the criteria introduced at the beginning of this article. Though there is no such a thing as a perfect working definition, and I do not expect a consensus to form soon on which one is the best, at least the ultimate incompatibility among the perspectives should be recognized.

It is still up to each researcher to choose how to use the term "AI," though it should be clarified when the result is discussed, with its implications understood well. Currently many researchers in the field produce valuable results, but not what they desired or claimed in advance. It is well known that many important ideas and techniques initiated in AI study, though they ended up contributing to the solution of other problems outside the field. Part of the reason for this to happen is the lack of a clear understanding of the assumptions of various AI projects.

Maybe at a future time we can find proper names for each research field involved, so as to resolve this confusion. That will probably happen when one of the working definitions of AI has shown undeniable success. Before that time, at least we can be clearer about what we mean by AI, and have a relatively accurate understanding about the potential and limitations of the concepts involved.

The story from Lewis Carroll quoted at the beginning of the article provides a perfect metaphor: even for people who "don't much care" where they are going, they will still get somewhere no matter which way they take, only if they walk long enough. However, that is not necessary where they initially wanted to get to, nor will they end up in the same place.

References

Adams, E. W. 1998. *A Primer of Probability Logic*. Stanford, California: CSLI Publications.

Albus, J. S. 1991. Outline for a Theory of Intelligence. *IEEE Transactions on Systems, Man, and Cybernetics* 21(3):473–509.

Allen, J. F. 1998. AI growing up: the changes and opportunities. *AI Magazine* 19(4):13–23.

Aristotle. 1882. *The Organon, or, Logical treatises of Aristotle*. London: George Bell. Translated by O. F. Owen.

Bach, J. 2009. *Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition*. Oxford: Oxford University Press.

Beattie, C.; Leibo, J. Z.; Teplyashin, D.; Ward, T.; Wainwright, M.; Küttler, H.; Lefrancq, A.; Green, S.; Valdés, V.; Sadik, A.; Schrittwieser, J.; Anderson, K.; York, S.; Cant, M.; Cain, A.; Bolton, A.; Gaffney, S.; King, H.; Hassabis, D.; Legg, S.; and Petersen, S. 2016. DeepMind Lab. *CoRR* abs/1612.03801.

Bhatnagar, S.; Alexandrova, A.; Avin, S.; Cave, S.; Cheke, L.; Crosby, M.; Feyereisl, J.; Halina, M.; Loe, B. S.; hÉigeartaigh, S. O.; Martnez-Plumed, F.; Price, H.; Shevlin, H.; Weller, A.; Winfield, A.; and Hernández-Orallo, J. 2018. Mapping Intelligence: Requirements and Possibilities. In Müller, V. C., ed., *Philosophy and Theory of Artificial Intelligence 2017*. Berlin: Springer. 117–135.

Birnbaum, L. 1991. Rigor mortis: a response to Nilsson's "Logic and artificial intelligence". *Artificial Intelligence* 47:57–77.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press, 1st edition.

Brachman, R. J. 2006. (AA)AI — more than the sum of its parts, 2005 AAAI Presidential Address. *AI Magazine* 27(4):19–34.

Cabrol, N. A. 2016. Alien Mindscapes – A Perspective on the Search for Extraterrestrial Intelligence. *Astrobiology* 16:661–676.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Cohen, P. R. 2005. If Not Turing's Test, then what? *AI Magazine* 26:61–67.

Davis, R. 1998. What are intelligence? and why? 1996 AAAI Presidential Address. *AI Magazine* 19(1):91–111.

Domingos, P. 2018. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York, NY, USA: Basic Books, Inc.

Executive Office of the President, USA. 2016. The National Artificial Intelligence Research and Development Strategic Plan.

Feigenbaum, E. A., and Feldman, J., eds. 1963. *Computers and Thought*. New York: McGraw-Hill.

Feigenbaum, E. A., and McCorduck, P. 1983. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the world*. Reading, Massachusetts: Addison-Wesley Publishing Company.

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

Franklin, S. 2007. A foundational architecture for artificial general intelligence. In Goertzel, B., and Wang, P., eds., *Advance of Artificial General Intelligence*. Amsterdam: IOS Press. 36–54.

Freeman, W. J. 1999. Consciousness, intentionality and causality. *Journal of Consciousness Studies* 6(11-12):143–172.

Goertzel, B., and Pennachin, C., eds. 2007. *Artificial General Intelligence*. New York: Springer.

Goertzel, B. 2009. Cognitive Synergy: A Universal Principle for Feasible General Intelligence? *Dynamical Psychology*.

Goertzel, B. 2014. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence* 5(1):1–46.

Goldman Sachs. 2016. Profiles in Innovation: Artificial Intelligence - AI, Machine Learning and Data Fuel the Future of Productivity.

Goldstein, S.; Princiotta, D.; and Naglieri, J. 2015. *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts*. New York: Springer.

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

Hájek, A. 2012. Interpretations of Probability. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition. URL: `http://plato.stanford.edu/archives/win2012/entries/probability-interpret/`, accessed in May 20, 2019.

Hawkins, J., and Blakeslee, S. 2004. *On Intelligence*. New York: Times Books.

Hayes, P., and Ford, K. 1995. Turing Test Considered Harmful. In Mellish, C. S., ed., *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, 972–977.

Hayes, P. J. 1977. In Defense of Logic. In Reddy, R., ed., *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 559–565.

Hearst, M. A., and Hirsh, H. 2000. AI's Greatest Trends and Controversies. *IEEE Intelligent Systems* 8–17.

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

Hofstadter, D. R., and FARG. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

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

Hofstadter, D. R. 1985. Waking up from the Boolean dream, or, subcognition as computation. In *Metamagical Themas: Questing for the Essence of Mind and Pattern*. New York: Basic Books. chapter 26.

Holland, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Massachusetts: MIT Press.

Hopcroft, J. E.; Motwani, R.; and Ullman, J. D. 2007. *Introduction to Automata Theory, Languages, and Computation*. Boston: Addison-Wesley, 3rd edition.

Hsu, F.-h.; Campbell, M. S.; and Hoane, Jr., A. J. 1995. Deep Blue System Overview. In *Proceedings of the 9th International Conference on Supercomputing*, 240–244. New York, NY, USA: ACM.

Hume, D. 1748. *An Enquiry Concerning Human Understanding*. London.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer.

Kirsh, D. 1991. Foundations of AI: the big issues. *Artificial Intelligence* 47:3–30.

Koene, R., and Deca, D. 2013. Editorial: Whole Brain Emulation seeks to Implement a Mind and its General Intelligence through System Identification. *Journal of Artificial General Intelligence* 4:1–9.

Kowalski, R. 1979. *Logic for Problem Solving*. New York: North Holland.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions*. Chicago University Press, 2nd edition.

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

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., eds., *Proceedings of the Second Conference on Artificial General Intelligence*, 91–96.

Laird, J. E. 2012. *The Soar Cognitive Architecture*. Cambridge, Massachusetts: MIT Press.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences* 40:E253.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep Learning. *Nature* 521:436–444.

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

Leimeister, J. M. 2010. Collective Intelligence. *Business & Information Systems Engineering* 2(4):245–248.

Luger, G. F. 2008. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Boston: Pearson, 6th edition.

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

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

Marr, D. 1977. Artificial intelligence: a personal view. *Artificial Intelligence* 9:37–48.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman & Co.

McCarthy, J.; Minsky, M.; Rochester, N.; and Shannon, C. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. URL: `http://www-formal.stanford.edu/jmc/history/dartmouth.html`, accessed in May 20, 2019.

McCarthy, J. 1988. Mathematical logic in artificial intelligence. *Dædalus* 117(1):297–311.

McCarthy, J. 2007. From here to human-level AI. *Artificial Intelligence* 171:1174–1182.

McCorduck, P. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natick, MA: A. K. Peters, Ltd., 2nd edition.

McCulloch, W. S., and Pitts, W. H. 1943. A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics* 5:115–133.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* 3:151–160.

McDermott, D. 2007. Level-headed. *Artificial Intelligence* 171:1183–1186.

Medin, D. L., and Ross, B. H. 1992. *Cognitive Psychology*. Fort Worth: Harcourt Brace Jovanovich.

Michalski, R.; Carbonell, J.; and Mitchell, T., eds. 1984. *Machine Learning: An Artificial Intelligence Approach*. Springer-Verlag.

Minsky, M.; Singh, P.; and Sloman, A. 2004. The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine* 25(2):113–124.

Minsky, M. 1983. Introduction to the COMTEX Microfiche Edition of the Early MIT Artificial Intelligence Memos. *AI Magazine* 4(1):19–22.

Minsky, M. 1985a. *The Society of Mind*. New York: Simon and Schuster.

Minsky, M. 1985b. Why intelligent aliens will be intelligible. In Regis, E., ed., *Extraterrestrials: Science and Alien Intelligence*. Cambridge: Cambridge University Press. 117–128.

Minsky, M. 1990. Logical vs. analogical or symbolic vs. connectionist or neat vs. scruffy. In Winston, P. H., and Shellard, S. A., eds., *Artificial Intelligence at MIT, Vol. 1: Expanding Frontiers*. Cambridge, Massachusetts: MIT Press. 218–243.

Minsky, M. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

Monett, D., and Lewis, C. W. P. 2018. Getting clarity by defining Artificial Intelligence - A Survey. In Müller, V. C., ed., *Philosophy and Theory of Artificial Intelligence 2017*. Berlin: Springer. 212–214.

Newell, A., and Simon, H. A. 1963. GPS, a program that simulates human thought. In Feigenbaum, E. A., and Feldman, J., eds., *Computers and Thought*. McGraw-Hill, New York. 279–293.

Newell, A., and Simon, H. A. 1976. Computer science as empirical inquiry: symbols and search. *Communications of the ACM* 19(3):113–126.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, Massachusetts: Harvard University Press.

Nilsson, N. J. 1986. Probabilistic logic. *Artificial Intelligence* 28:71–87.

Nilsson, N. J. 1991. Logic and artificial intelligence. *Artificial Intelligence* 47:31–56.

Nilsson, N. J. 1998. *Artificial Intelligence: A New Synthesis*. San Francisco: Morgan Kaufmann.

Nilsson, N. J. 2005. Human-level artificial intelligence? Be serious! *AI Magazine* 26(4):68–75.

Nilsson, N. J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge: Cambridge University Press.

Peirce, C. S. 1931. *Collected Papers of Charles Sanders Peirce*, volume 2. Cambridge, Massachusetts: Harvard University Press.

Piaget, J. 1960. *The Psychology of Intelligence*. Paterson, New Jersey: Littlefield, Adams & Co.

Piaget, J. 1963. *The Origins of Intelligence in Children*. New York: W.W. Norton & Company, Inc. Translated by M. Cook.

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

Reeke, G. N., and Edelman, G. M. 1988. Real brains and artificial intelligence. *Dædalus* 117(1):143–173.

Regis, E., ed. 1985. *Extraterrestrials: Science and alien intelligence*.

Roland, A., and Shiman, P. 2002. *Strategic computing : DARPA and the quest for machine intelligence, 1983-1993*. Cambridge, Massachusetts: MIT Press.

Rumelhart, D. E., and McClelland, J. L. 1986. PDP models and general issues in cognitive science. In Rumelhart, D. E., and McClelland, J. L., eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Foundations*. Cambridge, Massachusetts: MIT Press. 110–146.

Russell, S., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice Hall, 2nd edition.

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

Russell, S., and Wefald, E. H. 1991. *Do the Right Thing: Studies in Limited Rationality*. Cambridge, Massachusetts: MIT Press.

Russell, S. 1997. Rationality and intelligence. *Artificial Intelligence* 94:57–77.

Schank, R. C. 1991. Where is the AI? *AI Magazine* 12(4):38–49.

Searle, J. 1980. Minds, brains, and programs. *The Behavioral and Brain Sciences* 3:417–424.

Shannon, C. E., and Weaver, W. 1949. *The mathematical theory of communication*. Urbana, IL: The University of Illinois Press.

Shapiro, S. C. 1992. Artificial Intelligence. In Shapiro, S. C., ed., *Encyclopedia of Artificial Intelligence*. New York: John Wiley, 2 edition. 54–57.

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.

Simon, H. A.; Langley, P. W.; and Bradshaw, G. L. 1981. Scientific Discovery as Problem Solving. *Synthese* 47:1–27.

Simon, H. A. 1957. *Models of Man: Social and Rational*. New York: John Wiley.

Simon, H. A. 1983. *Reason in Human Affairs*. Stanford, California: Stanford University Press.

Solomonoff, R. J. 1964. A formal theory of inductive inference. Part I and II. *Information and Control* 7(1-2):1–22,224–254.

Thórisson, K. R.; Bieger, J.; Li, X.; and Wang, P. 2019. Cumulative Learning. In *Proceedings of the Twelfth Conference on Artificial General Intelligence*. To appear.

Thórisson, K. R. 2012. A New Constructivist AI: From Manual Methods to Self-Constructive Systems. In Wang, P., and Goertzel, B., eds., *Theoretical Foundations of Artificial General Intelligence*. Paris: Atlantis Press. 145–171.

Thórisson, K. R. 2013. *Reductio ad Absurdum*: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. AGI-13 workshop on Formalizing Mechanisms for Artificial General Intelligence and Cognition, Beijing, China, July 31st. Retrieved from `http://alumni.media.mit.edu/~kris/ftp/Thorisson-ReductioAdAbsurdum-AGI2013.pdf` in May 20, 2019.

Tomasello, M. 2000. Primate Cognition: Introduction to the Issue. *Cognitive Science* 24(3):351–361.

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

Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185:1124–1131.

von Neumann, J. 1958. *The Computer and the Brain*. New Haven, CT: Yale University Press.

Wang, P., and Goertzel, B. 2007. Introduction: Aspects of artificial general intelligence. In Goertzel, B., and Wang, P., eds., *Advance of Artificial General Intelligence*. Amsterdam: IOS Press. 1–16.

Wang, P., and Hammer, P. 2015. Issues in Temporal and Causal Inference. In Bieger, J.; Goertzel, B.; and Potapov, A., eds., *Proceedings of the Eighth Conference on Artificial General Intelligence*, 208–217.

Wang, P., and Li, X. 2016. Different Conceptions of Learning: Function Approximation vs. Self-Organization. In Steunebrink, B.; Wang, P.; and Goertzel, B., eds., *Proceedings of the Ninth Conference on Artificial General Intelligence*, 140–149.

Wang, P.; Liu, K.; and Dougherty, Q. 2018. Conceptions of Artificial Intelligence and Singularity. *Information* 9(4).

Wang, P. 1994. On the Working Definition of Intelligence. Technical Report 94, Center for Research on Concepts and Cognition, Indiana University, Bloomington, Indiana.

Wang, P. 1995. *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*. Ph.D. Dissertation, Indiana University.

Wang, P. 1996. Heuristics and normative models of judgment under uncertainty. *International Journal of Approximate Reasoning* 14(4):221–235.

Wang, P. 2001. Wason's cards: what is wrong. In Chen, L., and Zhuo, Y., eds., *Proceedings of the Third International Conference on Cognitive Science*, 371–375.

Wang, P. 2004a. The limitation of Bayesianism. *Artificial Intelligence* 158(1):97–106.

Wang, P. 2004b. Problem solving with insufficient resources. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 12(5):673–700.

Wang, P. 2004c. Toward a unified artificial intelligence. In *Papers from the 2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Research and Systems*, 83–90.

Wang, P. 2005. Experience-grounded semantics: a theory for intelligent systems. *Cognitive Systems Research* 6(4):282–302.

Wang, P. 2006a. Artificial Intelligence: What it is, and what it should be. In Lebiere, C., and Wray, R., eds., *Papers from the AAAI Spring Symposium on Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, 97–102.

Wang, P. 2006b. *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.

Wang, P. 2008. What do you mean by "AI". In Wang, P.; Goertzel, B.; and Franklin, S., eds., *Proceedings of the First Conference on Artificial General Intelligence*, 362–373.

Wang, P. 2009. Case-by-case problem solving. In Goertzel, B.; Hitzler, P.; and Hutter, M., eds., *Proceedings of the Second Conference on Artificial General Intelligence*, 180–185.

Wang, P. 2011. The Assumptions on Knowledge and Resources in Models of Rationality. *International Journal of Machine Consciousness* 3(1):193–218.

Wang, P. 2012. Theories of Artificial Intelligence – Meta-theoretical considerations. In Wang, P., and Goertzel, B., eds., *Theoretical Foundations of Artificial General Intelligence*. Paris: Atlantis Press. 305–323.

Wang, P. 2013. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific.

Wason, P. C., and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, Massachusetts: Harvard University Press.

Wiener, N. 1948. *Cybernetics, or control and communication in the animal and the machine*. New York: John Wiley & Sons, Inc.