What Do You Mean by “AI”?

Pei WANG
Temple University, Philadelphia, USA

Abstract. Many problems in AI study can be traced back to the confusion of different research goals. In this paper, five typical ways to define AI are clarified, analyzed, and compared. It is argued that though they are all legitimate research goals, they lead the research to very different directions, and most of them have trouble to give AI a proper identity. Finally, a working definition of AI is proposed, which has important advantages over the alternatives.

Keywords. Intelligence, working definition, research paradigm

1. The Problem of Defining “Intelligence”

A research project should have a clearly specified research goal; a research field should consist of research projects with related research goals. Though these requirements sound self-evident, Artificial Intelligence (AI) seems to be an exception, where people not only disagree on what is the best solution to the problem (which is usual in any branch of science and engineering), but also on what the problem is (which is unusual, at least given the extent of the disagreement). As evidence of this situation, at the 50th anniversary of the field, the AAAI Presidential Address still asked the question “what really is AI and what is intelligence about?” [1].

It is well known that people have different understandings to what “intelligence”, or “AI”, means. However, this issue has not been explored to the extent it deserves, mainly due to two widely spread opinions:

- There is a *natural* definition of the term “intelligence”, while the different understandings are just different aspects of the same notion, and the various AI schools are exploring different trails to the same summit, or working on different parts of the same whole.
- Like most terms in natural languages, the term “intelligence” cannot be defined, therefore people can keep whatever understanding they like about it, as far as their research produce useful results.

Though these two opinions take opposite positions on whether intelligence can be defined, they lead to the same attitude toward this issue, that is, they see the discussion on the definition of “intelligence” as a waste of time.

The aim of this paper is to show that both above opinions are wrong. In the following, I will clarify various understandings of AI, analyze their relations, and evaluate their potentials. I will then argue for the necessity and possibility of giving “intelligence” a proper “working definition” for the need of AI research. For the field as a whole, multiple working definitions exist, and it will remain to be the case in the near future. Even so, to clearly understand their difference is still very important.

For the emerging field of “Artificial General Intelligence” (AGI), this discussion has special importance. Since AGI treats “intelligence” as a whole [2], a project in this field will be inevitably guided and judged by its working definition of intelligence.

This paper follows my previous discussions on this topic [3, 4], and addresses the topic in a more accurate and comprehensive manner. As discussed in [3], a “working definition” of a term, like “intelligence”, is a definition to be used as the goal of a research project. To carry out the research consistently and efficiently, every researcher needs to select or establish such a working definition. At the early stage of a field, no working definition can be agreed by every researcher, but it does not mean that no one is better than another. A working definition should be *sharp*, *simple*, *faithful* to the original term, and *fruitful* in guiding the research. Since these requirements usually conflict with one another, the final choice is typically a compromise and tradeoff among various considerations. For instance, a working definition often can neither fully agree with the everyday usage of the term (which is fuzzy and vague), nor be fully formalized (which will be too far away from the everyday usage).

Though people have different opinions on how to accurately define AI, on a more general level they do agree on what this field is about. Human beings differ from animals and machines significantly in their *mental* ability, which is commonly called “intelligence”, and AI is the attempt to reproduce this ability in computer systems. This vague consensus sets important constraints on how AI should be defined:

- Since the best example of “intelligence” is the human mind, AI should be defined as *identical to human intelligence* in certain sense. At the early stage of research, this “identical to” (a matter of yes/no) can be relaxed to “similar to” (a matter of degree), and the progress of research can be indicated by the increased degree of similarity.
- Since AI is an attempt to duplicate human intelligence, not to completely duplicate a human being, an AI system is different from a person in certain other aspects. Otherwise the research would be aimed at “artificial person”, rather than intelligent computer. Therefore, it is not enough to say that an AI is similar to human without saying where the similarity is, since it cannot be in every aspect.

To make the analysis and comparison precise, in this paper human beings and computer systems are all specified as “agents” that “receive percepts from the environment and perform actions” [5]. At a given moment t, the full history of an agent can be represented as a triple $<P, S, A>$, where $P = <p_0, …, p_t>$ is the sequence of *percepts*, $A = <a_0, …, a_t>$ is the sequence of *actions*, and $S = <s_0, …, s_t>$ is the sequence of *internal states* the system has gone through. When a typical human mind is represented as $H = <P^H, S^H, A^H>$, and a typical intelligent computer as $C = <P^C, S^C, A^C>$, a working definition of AI corresponds to a definition of *similarity* between C and H, when the two are described at a certain level of abstraction.

Since this discussion is about the qualitative difference among AI working definitions, not about the quantitative difference in intelligence among systems, in the following no attempt will be made to establish a numerical measurement of this similarity. Instead, the focus will be on identifying the *factors* that are relevant to this similarity. To simplify the discussion, it is assumed that two sequences (of percepts, actions, or states) are similar as far as their corresponding components are similar to each other, and that the similarity between two percepts, two actions, and two states can be meaningfully evaluated in certain way.

Limited by length, this paper concentrates on the major types of working definitions of AI, without analyzing every proposed definition in detail. For the same reason, the paper will not address how to build an AI system according to a given working definition.

2. Typical Ways to Define AI

Following the distinction introduced in [3], typical ways to define AI are divided into five types, each of which evaluates the similarity between *C* and *H* by *structure*, *behavior*, *capability*, *function*, and *principle*, respectively. They are discussed in the following, one by one.

(1) By Structure

Since the best known instance of intelligence is produced by the human brain, it is natural to assume that AI can be achieved by building a brain-like *structure*, consisting of massive neuron-like processing units working in parallel.

This idea has been further developed in various forms, such as Connection Machine [6] and Artificial Neural Networks [7]. More recent brain-oriented AGI works include those by Hawkins [8] and de Garis [9].

Due to the complexity of the human brain and its fundamental difference from computer hardware, none of these projects plans to be faithful to the brain structure in all the details. Instead, they only take the brain as the source of inspirations, and the resulting systems approximate to the brain at a certain level and scope of description.

Even so, many people inside and outside the field of AI still believe that accurate "brain modeling" will provide the ultimate solution to AI, when it is allowed by our knowledge of the human brain and the available computer technology. According to this opinion, "the ultimate goals of AI and neuroscience are quite similar" [10].

I will call this type of definition "Structure-AI", since it requires the structural similarity between an AI system and the human brain. In the agent framework, it means that *C* is similar to *H* in the sense that

$$<P^C, S^C, A^C> \approx <P^H, S^H, A^H>$$

that is, the two have similar streams of percepts and actions, as well as similar state transforming sequences, due to their similar internal structure. According to this understanding of AI, even though it is impossible to accurately duplicate the brain structure in the near future, we should try to move to that goal as close as possible, and the distance to it can be used to evaluate the research results.

(2) By Behavior

Since intelligence seems to be more about the human mind than the human brain, many people believe that it is better to concentrate on the system's *behavior* when evaluating its intelligence. The best known idea in this category is the Turing Test [11]. Though Turing proposed his test only as a sufficient condition, not a necessary condition, for intelligence, it nevertheless is taken by many people as the definition of AI [12, 13].

A representative approach towards AGI, following this path, can be found in Newell's discussion of the Soar project [14], which was presented both as an AI system and a model of human psychology. According to this opinion, AI is identified with “cognitive modeling”, where the computer-produced results are evaluated by comparisons with psychological data produced by human subjects. In its later years, Soar has been moving away from this strong psychological orientation, so at the current time a better example for this category is ACT-R [15], though it is not proposed as an AI model, but a psychological model.

Another example of this understanding of AI can be found in the field of “chatbot”, where the intelligence of a system is evaluated according to how much it “talks like a human”, such as in the Loebner Prize Competition [16].

I will call this type of definition “Behavior-AI”, since it requires the behavioral similarity between an AI system and the human mind. In the agent framework, it means that *C* is similar to *H* in the sense that

$$<P^C, A^C> \approx <P^H, A^H>$$

that is, the two should have similar streams of percepts and actions. Here the two systems are treated as “black box”, whose internal structure and state do not matter. Of course, the AI system may be similar to a human mind only after a certain period of training, and that can be accepted in the above representation by setting the starting moment of the percepts and actions at the completion of the training.

(3) By Capability

For people whose interest in AI mainly comes from its potential practical applications, the intelligence of a system should be indicated by its *capability* of solving hard problems [17]. After all, this is how we usually judge the intelligence of a person. Furthermore, the progress of a research field will eventually be evaluated according to the usefulness of its results.

Partly because of such considerations, the earliest practical problems studied by AI were typical intellectual activities like theorem proving and game playing — if a person can solve these problems, we call the person “intelligent”; therefore, if a computer can do the same, then we may have to call the computer “intelligent”, too. Driven by similar motivations, a large number of application-oriented AI projects are “expert systems” in various domains — experts are intelligent, so if a computer can solve a problem that only an expert can, the computer must be intelligent, too.

Especially, a computer is often considered as intelligent if it solves a problem that previously could only be solved by human beings, but no computers. Consequently, AI becomes an expanding frontier of computer application.

The biggest AI achievements so far, according to this understanding, include Deep Blue, the chess-playing system that defeated the world champion, and Stanley, the self-driven vehicle that finished a 132-mile trek in 7 hours.

I will call this type of definition “Capability-AI”, since it requires an AI system to have human capability of practical problem solving. In the agent framework, it means that *C* is similar to *H* in the sense that there are moments *i* and *j* such that

$$< p_i^C, a_i^C> \approx < p_j^H, a_j^H>$$

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 I assume that a single percept can represent the problem, and a single action can represent the solution. Since here what matters is the final solution only, it is irrelevant whether the computer goes through a human-like internal process or produce human-like external behavior beyond this problem-solving process.

In the AGI context, it follows that systems with higher intelligence can solve more and harder problems. A recent form of this idea is Nilsson's "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."[18] Among existing AGI projects, a representative one of this type is Cyc, which encodes vast amounts of commonsense knowledge to achieve human-like problem-solving capability [19].

(4) By Function

Since most AI researchers are computer scientists and engineers, they prefer to represent the ability of an agent as some *function* that maps input (percepts) into output (actions), which is how a computer program is specified.

Typical opinions are like "Intelligence is the computational part of the ability to achieve goals in the world" and "What is important for AI is to have algorithms as capable as people at solving problems", both from McCarthy [20]. A more systematic and influential description came from Marr: "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" [21].

Guided by such opinions, the field of AI is widely seen as consisting of separate cognitive functions, such as searching, reasoning, planning, learning, problem solving, decision making, communicating, perceiving, acting, etc., each having its various computational formulations and algorithmic implementations [5].

I will call this type of definition "Function-AI", since it requires an AI system to have cognitive functions similar to those observed in humans. In the agent framework, it means that C is similar to H in the sense that there are moments i and j such 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 (problem) into an action (solution) in the computer is similar to that of a human. Since here the focus is on the functions, the actual percepts and actions of the two agents do not matter too much.

In the AGI context, such a working definition implies that a system should have many cognitive functions working together. Representative projects moving in this direction include LIDA [22] and Novamente [23].

(5) By Principle

Science always looks for simple and unified explanations of complicated and diverse phenomena. Therefore, it is not a surprise that some AI researchers attempt to identify the fundamental *principle* by which human intelligence can be explained and reproduced in computer at a general level.

Intuitively, “intelligence” is associated with the ability to get the best solution. However, such a definition would be trivial if it asks the agent to exhaustively evaluate all possible solutions and to select the best among them. To be more realistic, Simon proposed the notion of “Bounded Rationality”, which restricts what the agent can know and do [24]. Russell argued that intelligent agents should have “Bounded Optimality”, the ability to generate maximally successful behavior given the available information and computational resources [25].

Among AGI projects, AIXI [26] and NARS [4] can be seen as different attempts to build AI as some type of rational or optimal system, though they specify rationality in different ways, and make very different assumptions on the environment of the system. AIXI aims at the highest expected reward, under the assumption that the system has sufficient resources and the environment is a Turing Machine. On the other hand, in NARS “intelligence” is defined as “adaptation with insufficient knowledge and resources”, which puts no restriction on the environment, while requiring the system to be finite, real-time, and open.

I will call this type of definition “Principle-AI”, since it requires an AI system to follow similar normative principles as the human mind. 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 into the whole stream of actions in the computer is similar to that of a human. Again, here the focus is on the function, not the actual percepts and actions. Here the function is called a “principle”, to stress that it is not just about a single problem and its solution, but about the agent's life-long history in various situations, when dealing with various types of problems.

3. The Necessity of Distinction

The above five types of working definition all set legitimate research goals, but they are different from each other.

- Structure-AI contributes to the study of the human brain. It also helps to explain how the brain carries out various cognitive activities, but if the research goal is in the behavior, capability, function, or principle of the mind, then to duplicate the brain structure is often not the best way (in terms of simplicity and efficiency), because the brain is formed under biological and evolutionary restrictions largely irrelevant to computers.
- Behavior-AI contributes to the study of human psychology. Very often, “the human way” gives us inspirations on how to use a computer, but it is not the best way to solve a practical problem, or to implement a cognitive function or principle. Also, behavior similarity does not necessarily require structural similarity.
- Capability-AI contributes to various application domains, by solving practical problems there. However, due to the lack of generality of the solutions, this

kind of solution usually contributes little to the study of brain or mind outside the scope of the domain problems.

- Function-AI contributes to computer science, by producing new software (sometimes also hardware) that can carry out various type of computation. However, the best way to implement the required computation is usually not exactly the way such a process is carried out in the human mind/brain complex. Since a cognitive function is generalized over many concrete problems, it is not necessarily the best way to solve each of them. If an agent is equipped with multiple cognitive functions, they are not necessarily designed according to the same principle.
- Principle-AI contributes to the study of information processing in various situations, by exploring the implications of different assumptions. Given the generality of a principle, it cannot explain all the details of the human brain or the human mind, nor does it provide the best way to solve every practical problem. Even though a principle-based system usually does carry out various cognitive functions, they are not necessarily separate processes, each with its own computational formulation and algorithmic implementation.

Therefore, these five trails lead to different summits, rather than to the same one.

If these working definitions of AI all originated from the study of the human brain/mind, how can they end up in different places? It is because each of them corresponds to a different level of description. Roughly speaking, the five types of working definitions in the above list are listed in the order of increasing generality and decreasing specificity, with Structure-AI being the most "human-centric" approach, and Principle-AI the least (though it still comes from an abstraction of human thinking). Each level of description comes with its concepts and vocabulary, which make certain patterns and activities more visible, while ignore other patterns and activities visible in the other levels, either above it or below it. In this context, there is no such a thing as the "true" or "correct" level of description, and each of the five can be used as goals of legitimate scientific research.

To distinguish five types of AI definition does not mean that they are unrelated to each other. It is possible to accept a working definition as the primary goal, and also to achieve some secondary goals at the same time, or to benefit from works aimed at a different goal. For example, when implementing a principle, we may find that the "human way" is very simple and efficient, which also provides good solutions to some real-world problems. However, even in such a situation, it is still necessary to distinguish the primary goal of a research from the additional and secondary results it may produce, because whenever there is a design decision to make, it is the primary goal that matters most.

Even though each of the five types of AI definition is valid, to mix them together in one project is not a good idea. Many current AI projects have no clearly specified research goal, and people working on them often swing between different definitions of intelligence. Such a practice causes inconsistency in the criteria of design and evaluation, though it may accidentally produce some interesting results.

A common mistake is to believe that there is a "true" ("real", "natural") meaning of "intelligence" that all AI research projects must obey. Some people think that AI should follow the common usage (i.e., the dictionary definition) of the word "intelligence". This is not going to work. The meaning of "intelligence" in English (or a similar word in another natural language) was largely formed before AI time, and therefore is mainly about human intelligence, where the descriptions at various levels (structure, behavior, capability, function, principle, etc.) are unified. On the contrary, for computer systems these aspects become different goals, as discussed above.

For similar reasons, AI cannot simply borrow the definition of “intelligence” from other disciplines, such as psychology or education, though the notion does have a longer history in those fields. This is not only because there are also controversies in those fields about what intelligence is, but also because the notion “intelligence” is mainly used there to stress *the difference among human beings* in cognitive ability. On the contrary, for AI this difference is almost negligible, and the notion is mainly used to stress *the difference between human beings and computer systems*. Also for this reason, it may not be a good idea to use IQ test to judge the ability of AI systems.

Some people argue that “AI is simply what the AI researchers do”. Though a survey of the field provides a valid *descriptive definition* of AI, it is not a valid *working definition*, which should be precise and coherent to guide a research project. Under the common name “AI”, AI researchers are actually doing quite different things, as described previously. Even if there is a majority point of view, it does not necessary become the “true meaning” of AI that everyone must concur.

It is true that in many science disciplines the basic notions become well defined only after long-term research. However, in those disciplines, at least the *phenomena* to be studied are clearly identified at the beginning, or the disagreements in working definitions of those notions do not make too much difference in the direction of the research. On the contrary, in AI each researcher has to decide *which phenomena* of the human intelligence should be studied and at which level of description. Such a decision is inevitably based on an explicitly or implicitly accepted working definition of intelligence. There is no way to be “definition-neutral”, because otherwise the research would have nowhere to start — a phenomenon is relevant to AI only when the term “AI” has meaning, no matter how vague or poor the meaning is.

Furthermore, the existing working definitions of AI are incompatible with each other, as discussed previously, to the extent that progress toward one may be moving away from another. It is very difficult, if meaningful, to design or evaluate an AI system without considering its research goal first.

The confusion among different definitions is a common root of many controversies in AI. For example, there has been a debate on whether Deep Blue is a success of AI [27, 28]. According to the above analysis, the conclusion should clearly be “yes” if “AI” is taken to mean “Capability-AI”, otherwise the answer should be “not much”, or even “no”. We cannot assume people are talking about the same thing only because they are all using the term “AI”.

4. The Possibility of Comparison

To say there are multiple valid working definitions of intelligence (and therefore, AI) does not mean that they cannot be compared, or that they are equally good.

In [3], four criteria of a good working definition were borrowed from Carnap's work (when he tried to define “probability”) [29]:

- It should have a *sharp* boundary.
- It should be *faithful* to the notion to be clarified.
- It should lead to *fruitful* research.
- It should be as *simple* as possible.

Given their forms as defined previously, the five types of definition are not too different with respect to the requirements of *sharpness* and *simplicity*. Therefore, the following discussion focuses on the other two criteria, *faithfulness* and *fruitfulness*.

As mentioned before, in general it is hard to say which of the five is more faithful to the everyday usage of the word "intelligence", because each of them captures a different aspect of it. Similarly, each of the five produces important results, and which one is "more fruitful" can only be determined after decades or even longer.

Therefore, instead of trying to decide which working definition is the best *in general*, in the following I will focus on one aspect of this topic: which working definition will give the field "AI" a proper *identity,* which should explain how the field differs from the other fields, as well as elicits the common natures of its subfields.

AI has been suffering from a serious identity crisis for years. Many AI researchers have complained that the field has not got the recognition it deserves, which is sometimes called "The AI Effect" — as soon as a problem is solved, it is no longer considered as a problem for AI anymore [30]. Within the field, fragmentation is also a big problem [1] — each subfield has its own research goal and methods, and to collectively call them "AI" seems only to have a historical reason, that is, they all more or less came out of attempts of making computer "intelligent", whatever that means.

For AI to be considered as a field of its own, its definition must satisfy the following conditions:

- AI should not be defined in such a *narrow* way that takes human intelligence as the only possible form of intelligence, otherwise AI research would be impossible, by definition.
- AI should not be defined in such a *broad* way that takes all existing computer systems as already having intelligence, otherwise AI research would be unnecessary, also by definition.

Now let us analyze the responsibility of each type of working definition with respect to the identity problem the field AI faces. Especially, how the definition posits AI with respect to human psychology and computer science.

There is no doubt that the best example of "intelligence" is "human intelligence", and therefore all working definitions attempt to make computer systems "similar" to humans, in various senses, and to various extents. However, Structure-AI and Behavior-AI seem to leave too little space for "non-human intelligence", so they may be sufficient conditions for "intelligence", but unlikely to be necessary conditions. If an intelligent system must have human brain structure or produce human cognitive behaviors, then some other possibilities, such as "animal intelligence", "collective (group) intelligence", and "extraterrestrial intelligence", all become impossible *by definition*. It would be similar to defining "vision" by the structure or behavior of the human visual organ. For AI, such a definition will seriously limit our imagination and innovation of novel forms of intelligence. Human intelligence is developed under certain evolutionary and biological restrictions, which are essential for human, but hardly for intelligence in general. After all, "Artificial Intelligence" should not be taken to mean "Artificial Human Intelligence", since "Intelligence" should be a more general notion than "Human Intelligence".

On the other hand, Capability-AI and Function-AI seem to allow too many systems to be called "intelligent". It is not hard to recognize that works under the former is just like what we usually call "computer application", and the latter, "computer science", except that the problems or tasks are those that "humans can do or try to do" [27]. Do these definitions give enough reason to distinguish AI from Computer Science (CS)?

Marr's computation-algorithm-implementation analysis of AI [21] can be applied to every problem studied in CS, and so does the following textbook definition: “we define AI as the study of agents that receive percepts from the environment and perform actions” [5]. This consequence is made explicit by the claim of Hayes and Ford that AI and CS are the same thing [31].

If the only difference between AI and CS is that the “AI problems” are historically solved by the human mind, then how about problems like sorting or evaluating arithmetic expression? Some people have taken the position that all programs are intelligent, and their difference in intelligence is just a matter of degree. Such a usage of the concept of “intelligence” is coherent, except that the concept has been trivialized too much. If intelligence is really like this, there is no wonder why AI has got little credit and recognition — if everything developed in the field of AI can be done in CS, and the notion “intelligent agent” has no more content than “agent”, or even “system”, what difference does it make if we omit the fancy term “intelligence”?

Furthermore, the widely acceptance of Capability-AI and Function-AI as working definitions of AI is responsible for the current fragmentation of AI. Both of them define AI by a *group* (of capabilities and functions, respectively), without demanding much commonality among its members. As a result, AI practitioners usually assume they can, and should, start to work on a single capability or function, which may be integrated to get a general intelligence in the future. Since the best ways to solve practical problems and to carry out formal computations differ greatly from case to case, there is not too much to be learned from each other, even though all of them are called “AI”. As far as people continue to define their problems in this way, the fragmentation will continue.

The above analysis leaves us only with Principle-AI. Of course, like the other four types discussed above, Principle-AI is not a single working definition, but a group of them. Different members in the group surely lead to different consequences. Obviously, if the “principle” under consideration is too broad, it will include all computer systems (which will be bad); if it is too narrow, it will exclude all non-human systems (which will be bad, too). Therefore we need something *in between*, that is, a principle that

(1) is followed by the human mind,
(2) can be followed by computer systems,
(3) but are not followed by traditional computer systems.

An example of such a working definition of AI is the one accepted in the NARS project. Briefly speaking, it identifies “intelligence” with “adaptation with insufficient knowledge and resources”, which implies that the system is finite, works in real-time, is open to novel tasks, and learns from experience [3, 4]. There are many reasons to believe that the human mind is such a system. The practice of NARS shows that it is possible to develop a computer system following this principle. Finally, traditional computer systems do not follow this principle. Therefore, such a working definition satisfies the previous requirements. Though NARS can be studied in different aspects, the system cannot be divided into independent functions or capabilities, since the components of the system tangle with one another closely, so cannot be treated in isolation. The notion of “intelligence” is not an optional label in this research, since it does introduce ideas not available in computer science or cognitive psychology. Designed in this way, NARS has shown many interesting properties [4], though to discuss them is far beyond the scope of this paper.

To prefer the NARS definition of AI does not mean that it can replace the others for all purposes. As discussed before, each valid working definition of AI has its value. Principle-based definitions are often described as “looking for a silver bullet”, labeled as “physics envy”, and rejected by arguments like “intelligence is too complicated to be explained by a few simple principles”. However, all these criticisms take such a definition (of Principle-AI) as the *means* to achieve other *ends* (Structure-AI, Behavior-AI, Capability-AI, or Function-AI), which is a misconception. The NARS definition may give AI a better identity than the other definitions do, though it does not produce all the values that can be produced by the others.

Obviously, the NARS definition of AI is not a *descriptive definition* of the term “AI”, that is, its common usage in the field, and most of the existing “AI systems” do not satisfy this definition. However, it does not necessarily mean that this definition should be rejected, but may imply that the field should change into a more coherent and fruitful discipline of science.

5. Conclusion

Though intuitively everyone agrees that AI means to build computer systems that are similar to the human mind in some way, they have very different ideas on where this similarity should be. Among the many existing definitions [32], typical opinions define this similarity in terms of structure, behavior, capability, function, or principle [3].

These working definitions of AI are all valid, in the sense that each of them corresponds to a description of the human intelligence at a certain level of abstraction, and sets a precise research goal, which is achievable to various extents. Each of them is also fruitful, in the sense that it has guided the research to produce results with intellectual and practical values.

On the other hand, these working definitions are different, since they set different goals, require different methods, produce different results, and evaluate progress according to different criteria. They cannot replace one another, or be “integrate” into a coherent definition that satisfies all the criteria at the same time.

The common beliefs on this topic, “AI cannot be defined” and “All AI definitions are roughly equivalent” are both wrong. AI can have working definitions that serve as ultimate research goals. Every researcher in the field usually holds such a definition, though often implicitly. To improve the coherence and efficiency of research and communication, it is better to make our working definitions explicit.

This topic matters for AI, since the current AI research suffers from the confusion of various goals and the missing of an identity. Consequently, many debates are caused by different meanings of the term “AI”, and the field as a whole is fragmented within, as well as has trouble to justify its uniqueness and integrity to the outside world.

This topic is crucial for AGI, given its stress on the “big picture” of an intelligent system. Even though at the current time no working definition is perfect or final, to dismiss the issue will damage the consistency of system design and evaluation.

Though there are many valid ways to define AI, they are not equally good. We will not reach a consensus on which one is the best very soon, so in the field the different working definitions will co-exist for a long time. Even so, it is important to understand their difference and relationship.

Different working definition gives the field of AI different identities. To solve the problems of internal fragmentation and external recognition, the most promising way is to define AI by a principle of rationality that is followed by the human mind, but not by traditional computer systems. The NARS project shows that such a solution is possible.

Reference

[1] Ronald J. Brachman, (AA)AI: more than the sum of its parts, AI Magazine, 27(4):19-34, 2006.

[2] Pei Wang and Ben Goertzel, Introduction: Aspects of artificial general intelligence, In Advance of Artificial General Intelligence, B. Goertzel and P. Wang (editors), 1-16, IOS Press, Amsterdam, 2007.

[3] Pei Wang, On the working definition of intelligence, Technical Report No. 94, Center for Research on Concepts and Cognition, Indiana University, Bloomington, Indiana, 1994.

[4] Pei Wang, Rigid Flexibility: The Logic of Intelligence, Springer, Dordrecht, 2006.

[5] Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 2nd edition, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[6] W. Daniel Hillis, The Connection Machine, MIT Press, Cambridge, Massachusetts, 1986.

[7] Paul Smolensky, On the proper treatment of connectionism, Behavioral and Brain Sciences, 11:1-74, 1988.

[8] Jeff Hawkins and Sandra Blakeslee, On Intelligence, Times Books, New York, 2004.

[9] Hugo de Garis, Artificial Brains, In Artificial General Intelligence, Ben Goertzel and Cassio Pennachin (editors), 159-174, Springer, Berlin, 2007.

[10] George N. Reeke and Gerald M. Edelman, Real brains and artificial intelligence, Dædalus, 117:143-173, 1988.

[11] Alan M. Turing, Computing machinery and intelligence, Mind, LIX:433-460, 1950.

[12] Ian F. Brackenbury and Yael Ravin, Machine intelligence and the Turing Test, IBM Systems Journal, 41:524-529, 2002.

[13] Lenhart K. Schubert, Turing's dream and the knowledge challenge, Proceedings of the Twenty-first National Conference on Artificial Intelligence, 1534-1538, Menlo Park, California, 2006.

[14] Allen Newell, Unified Theories of Cognition, Harvard University Press, Cambridge, Massachusetts, 1990.

[15] John R. Anderson and Christian Lebiere, The atomic components of thought. Erlbaum, Mahwah, New Jersey, 1998.

[16] Michael L. Mauldin, ChatterBots, TinyMuds, and the Turing Test: entering the Loebner Prize competition, Proceedings of the Twelfth National Conference on Artificial Intelligence, 16-21, 1994.

[17] Marvin Minsky, The Society of Mind, Simon and Schuster, New York, 1985.

[18] Nils J. Nilsson, Human-level artificial intelligence? Be serious! AI Magazine, 26(4):68-75, 2005.

[19] Douglas B. Lenat, Cyc: a large-scale Investment in Knowledge Infrastructure, Communications of the ACM, 38(11):33-38,1995.

[20] John McCarthy, What is Artificial Intelligence?
On-line paper at www-formal.stanford.edu/jmc/whatisai/whatisai.html, 2004.

[21] David Marr, Artificial intelligence: a personal view, Artificial Intelligence, 9:37-48, 1977.

[22] Stan Franklin, A foundational architecture for artificial general intelligence. In Advance of Artificial General Intelligence, B. Goertzel and P. Wang (editors), 36-54, IOS Press, Amsterdam, 2007.

[23] Moshe Looks, Ben Goertzel, and Cassio Pennachin, Novamente: An Integrative Architecture for General Intelligence, In papers from AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, 54-61, 2004.

[24] Herbert A. Simon, Models of Man: Social and Rational, John Wiley, New York, 1957.

[25] Stuart Russell, Rationality and intelligence, Artificial Intelligence, 94:57-77, 1997.

[26] Marcus Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, Berlin, 2005.

[27] James F. Allen, AI growing up: The changes and opportunities, AI Magazine, 19(4):13-23, 1998.

[28] Drew McDermott, Mind and Mechanism, MIT Press, Cambridge, Massachusetts, 2001.

[29] Rudolf Carnap, Logical Foundations of Probability, University of Chicago Press, Chicago, 1950.

[30] Roger C. Schank, Where is the AI? AI Magazine, 12(4):38-49, 1991.

[31] Patrick Hayes and Kenneth Ford, Turing test considered harmful, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 972-977, 1995.

[32] Shane Legg and Marcus Hutter, A Collection of Definitions of Intelligence, In Advance of Artificial General Intelligence, B. Goertzel and P. Wang (editors), 17-24, IOS Press, Amsterdam, 2007.