information

MDPI

Article

Conceptions of Artificial Intelligence and Singularity

Pei Wang [1,*], Kai Liu [2] and Quinn Dougherty [3]

1 Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
2 School of Psychology, Central China Normal University, Wuhan 430079, China; ccnulk@mail.ccnu.edu.cn
3 Philly AGI Team, Philadelphia, PA 19122, USA; quinn.dougherty.phila@gmail.com
* Correspondence: pei.wang@temple.edu

Received: 15 February 2018; Accepted: 3 April 2018; Published: 6 April 2018

Abstract: In the current discussions about "artificial intelligence" (AI) and "singularity", both labels are used with several very different senses, and the confusion among these senses is the root of many disagreements. Similarly, although "artificial general intelligence" (AGI) has become a widely used term in the related discussions, many people are not really familiar with this research, including its aim and status. We analyze these notions, and introduce the results of our own AGI research. Our main conclusions are that: (1) it is possible to build a computer system that follows the same laws of thought and shows similar properties as the human mind, but, since such an AGI will have neither a human body nor human experience, it will not behave exactly like a human, nor will it be "smarter than a human" on all tasks; and (2) since the development of an AGI requires a reasonably good understanding of the general mechanism of intelligence, the system's behaviors will still be understandable and predictable in principle. Therefore, the success of AGI will not necessarily lead to a singularity beyond which the future becomes completely incomprehensible and uncontrollable.

Keywords: artificial general intelligence; technical singularity; non-axiomatic reasoning system

1. Introduction

Driven by the remarkable achievements of deep learning, it becomes a hot topic again to debate whether computers can be smarter than humans. In the debate, there are two opposite tendencies that are both wrong in our opinion:

- Many claims about what AI (artificial intelligence) will be able to do are obtained using naive extrapolation of the past progress, without addressing the conceptual and technical difficulties in the field.
- Many claims about what AI will not be able to do are derived from traditional conceptions on how a computer system should be built and used, as well as an anthropocentric usage of notions such as "intelligence" and "cognition".

We consider the leading article of this Special Issue [1] as having done a good job in criticizing the first tendency by pointing out a list of features that any truly intelligent system should have, and arguing that mainstream AI techniques cannot deliver them, even after more research. However, many of the authors' conclusions exactly fall into the second tendency mentioned above, mainly because they are not familiar with existing AGI (artificial general intelligence) research. Since the opinions expressed in the leading article are representative, in this article, we will focus on the issues they raised, without addressing many related topics.

In the following, we start by distinguishing and clarifying the different interpretations and understandings of AI and singularity, and then explain how AGI is related to them. After that, we briefly summarize the AGI project our team has been working on, and explain how it can produce

the features that the leading article claimed to be impossible for AI. In the conclusion, we agree with the authors of the leading article [1] that the recent achievements of deep learning are still far from showing that the related techniques can give us AGI or singularity; however, we believe AGI can be achieved via paths outside the vision of mainstream AI researchers, as well as that of its critics. This conception of AGI is fundamentally different from that of the current mainstream conception of AI. As for "singularity", we consider it an ill-conceived notion, as it is based on an improper conception of intelligence.

2. Notions Distinguished and Clarified

Let us first analyze what people mean when talking about "AI" and "Singularity". Both notions have no widely accepted definitions, although there are common usages.

2.1. Different Types of AI

In its broadest sense, AI is the attempt "to make a computer work like a human mind". Although it sounds plain, this description demands an AI to be similar (or even identical) to the human mind in certain aspects. On the other hand, because a computer is not a biological organism, nor does it live a human life, it cannot be expected to be similar to the human mind in all details. The latter is rarely mentioned but implicitly assumed, as it is self-evident. Consequently, by focusing on different aspects of the human mind, different paradigms of AI have been proposed and followed, with different objectives, desiderata, assumptions, road-maps, and applicabilities. They are each valid but distinct paradigms of scientific research [2].

In the current discussion, there are at least three senses of "AI" involved:

1. A computer system that behaves exactly like a human mind
2. A computer system that solves certain problems previously solvable only by the human mind
3. A computer system with the same cognitive functions as the human mind

In the following, they are referred to as AI-1, AI-2, and AI-3, respectively.

The best-known form of AI-1 is a computer system that can pass the Turing Test [3]. This notion is easy to understand and has been popularized by science-fiction novels and movies. To the general public, this is what "AI" means; however, it is rarely the research objective in the field, for several reasons.

At the very beginning of AI research, most researchers did attempt to build "thinking machines" with capabilities comparable (if not identical) to that of the human mind [3–5]. However, all direct attempts toward such goals failed [6–8]. Consequently, the mainstream AI researchers reinterpreted "AI" as AI-2, with a limited scope on a specific application or a single cognitive function. Almost all results summarized in the common AI textbooks [9,10] belong to this category, including deep learning [11] and other machine learning algorithms [12].

Although research on AI-2 has made impressive achievements, many people (both within the field and outside it) still have the feeling that this type of computer system is closer to traditional computing than to true intelligence, which should be general-purpose. This is why a new label, "AGI", was introduced more than a decade ago [13,14], even though this type of research projects has existed for many years. What distinguishes AGI from mainstream AI is that the former treats "intelligence" as one capability, while the latter treats it as a collection of loosely related capabilities. Therefore, AGI is basically the AI-3 listed above.

The commonly used phrase "Strong AI" roughly refers to AI-1 and AI-3 (AGI), in contrast to "Weak AI", referring to AI-2. Although this usage has intuitive attraction with respect to the ambition of the objectives, many AGI researchers usually do not use these phrases themselves, partly to avoid the philosophical presumptions behind the phrases [15]. Another reason is that the major difference between AI-2 and AI-3 is not in "strength in capability", but "breadth of applicability". For one concrete problem, a specially designed solution is often better than the solution provided by an AGI. We cannot expect an AI-2 technique which becomes "stronger" to eventually become AI-3, as the two

are designed under fundamentally different considerations. For the same reason, it is unreasonable to expect to obtain an AI-3 system by simply bundling the existing AI-2 techniques together.

Furthermore, "Strong AI" fails to distinguish AI-1 and AI-3, where AI-1 focuses on the external behaviors of a system, while AI-3 focuses on its internal functions. It can be argued that "a computer system that behaves exactly like a human mind" (AI-1) may have to be built "with the same cognitive functions as the human mind" (AI-3); even so, the reverse implication is not necessarily true because the behaviors of a system, or its "output", not only depends on the system's processing mechanism and functions, but also on its "input", which can be roughly called the system's "experience". In the same way, two mathematical functions which are very similar may still produce very different output values if their input values are different enough [2].

In that case, why not give AGI human experience? In principle, it can be assumed that human sensory and perceptive processes can be simulated in computing devices to any desired accuracy. However, this approach has several obstacles. First, accuracy with regard to "human" sensory processes is not a trivial consideration. Take vision as an example: light sensors should have identical sensibility, resolution, response time, etc., as the human eye. That is much more to ask than for the computer to have "vision". Instead, it is to ask the computer to have "human vision", which is a special type of vision.

Even if we can simulate all human senses to arbitrary accuracy, they still can only produce the direct or physical experience of a normal human, but not the indirect or social experience obtained through communication, which requires the computer to be treated by others (humans and machines) as a human. This is not a technical problem, as many human beings will have no reason to do so.

For the sake of argument, let us assume the whole society indeed treats AGI systems exactly as if they were humans; in this case, AI-1 is possible. However, such an AI-1 is based on a highly anthropocentric interpretation of "intelligence", thus it should be called "Artificial Human Intelligence". To define general intelligence using human behavior would make other forms of intelligence (such as "animal intelligence", "collective intelligence", "extraterrestrial intelligence", etc.) impossible by definition, simply because they cannot have human-like inputs and outputs.

Such an anthropocentric interpretation of "intelligence" is rarely stated explicitly, although it is often assumed implicitly. One example is to take Turing Test as a working definition of AI, even though Turing himself only proposed it as a sufficient condition, but not a necessary condition, of intelligence or thinking. Turing [3] wrote: "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."

Among the current AGI researchers, we do not know anyone whose goal is to build an AI-1; instead, it is more proper to see their works as aiming at some version of AI-3. They believe "thinking machines" or "general intelligence" can be built which are comparable, or even identical to the human mind at a certain level of description, although not in all details of behaviors. These differences nevertheless do not disqualify these systems from being considered as truly intelligent, just like we consider fish and birds as having vision, despite knowing that what they see is very different from what we see.

What is currently called "AGI" is very similar to the initial attempts to do this under the name of "AI", and the new label was adopted around 2005 by a group of researchers who wanted to distinguish their objective from what was called "AI" at the time (i.e., AI-2). Since then, the AGI community has been mostly identified by its annual conference (started in 2008) and its journal (launched in 2009). Although there has been a substantial literature, as well as open-source projects, AGI research is still far from its goal; there is no widely accepted theory or model yet, not to mention practical application. As AGI projects typically take approaches unfavored by the mainstream AI community, the AGI community is still on the fringe of the field of AI, with their results largely unknown to the outside world, even though the term "AGI" has become more widely used in recent years.

On this aspect, the lead article [1] provides a typical example. Its main conclusion is that "strong AI" and "AGI" (the two are treated as synonymy) are impossible, and the phrase of "AGI" is used many times in the article, but its 68 references do not include even a single paper from the AGI conferences or journal, nor does the article discuss any of the active AGI projects, where most of the "ignored characteristics" claimed by Braga and Logan [1] have been explored, and demonstrable (although often preliminary) results have been produced. Here, we are not saying that AGI research cannot be criticized by people outside the field, but that such criticism should be based on some basic knowledge about the current status of the field.

In our opinion, one major issue of the lead article [1] is the failure to properly distinguish interpretations and understandings of "AI". We actually agree with its authors' criticism of mainstream AI and its associated hype, as well as the list of characteristics ignored in those systems. However, their criticism of AGI research is attacking a straw man, as it misunderstands AGI's objective (they assume it is that of AI-1) and current status (they assume it is that of AI-2).

2.2. Presumptions of Singularity

The "Singularity", also known as the "Technological Singularity", is another concept that has no accurate and widely accepted definition. It has not been taken to be a scientific or technical term, even though it has become well-known due to some writings for the general public (e.g., [16]).

In its typical usage, the belief that "AI will lead to singularity" can be analyzed into the conjunction of the following statements:

1. The intelligence of a system can be measured by a real number.
2. AI should be able to increase its intelligence via learning or recursive self-improvement.
3. After the intelligence of AI passes the human-level, its entire future will be perceived as a single point, since it will be beyond our comprehension.

However, some people also use "singularity" for the time when "human-level AI is achieved", or "computers have become more intelligent than human", without the other presumptions. In the following, we focus on the full version, although what we think about its variants should be quite clear after this analysis.

The first statement looks agreeable intuitively. After all, an "intelligent" or "smart" system should be able to solve many problems, and we often use various tests and examinations to evaluate that. In particular, human intelligence is commonly measured by an "intelligence quotient" (IQ). To accurately define a measurement of problem-solving capability for general-purpose systems will not be easy, but, for the sake of the current discussion, we assume such a measurement S can be established. Even so, we do not consider S a proper measurement of a system's "intelligence", as it misses the time component. In its common usage, the notion of intelligence is associated closer to "learned problem-solving capability" than to "innate problem-solving capability". For this reason, at a given time t, the intelligence of the system probably should not be measured by $S(t)$, but $S'(t)$, i.e., the increasing rate of S at the moment.

To visualize the difference between the two measurements, in Figure 1, there are four functions indicating how a system's total problem-solving score S is related to time t:

- **B-type:** The blue line corresponds to a system with constant problem-solving capability—what the system can do is completely determined at the beginning, i.e., $S'(t) = 0$. All traditional computation systems belong to this type, and some of them are referred to as "AI".
- **P-type:** The pink line corresponds to a system that increases its problem-solving capability until it infinitely approximates an upper bound. For such a system, $S'(t) > 0$, but converges to 0. Most machine learning algorithms belong to this type.
- **G-type:** The green line corresponds to a system where its problem-solving capability $S(t)$ increases without an upper bound and $S'(t)$ is a positive constant. Many AGI projects, including ours, belong to this type.

- **R-type:** The red line corresponds to a system where both $S(t)$ and $S'(t)$ increase exponentially. We do not think such a system is possible to be actually built, but list it here only as a conceptual possibility to be discussed.

Here, $S(t)$ is "problem-solving capability", while $S'(t)$ is "learning capability", and the two are not directly correlated in their values. As can be seen in Figure 1, depending on the constants and moment of measuring, each of the four types can be the most capable one in problem solving, but with respect to learning their order is basically the order of the previous descriptions: B < P < G < R.

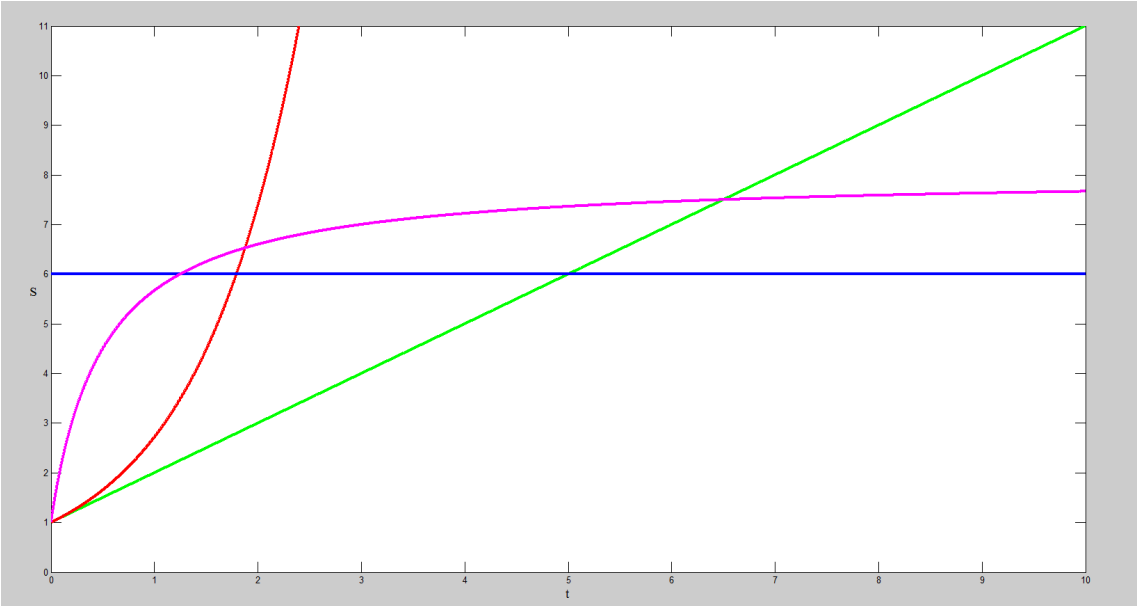

Figure 1. Four typical relations between time t and score S.

Our opinion is that "intelligence" should be measured by $S'(t)$, not $S(t)$. We believe this understanding of intelligence fits the deep sense of the word better, and will be more fruitful when used to guide the research of AI, although we know that it differs from current mainstream opinion.

The above conclusion does not necessarily conflict with the practice of human intelligence quotient (IQ) evaluation, which practically measures certain problem-solving capability. As IQ is the quotient obtained by dividing a person's "mental age" (according to the test score) by the person's chronological age. It can be interpreted as indicating the person's learning rate compared with that of other people, as higher $S(t)$ value implies higher $S'(t)$ value, given that the innate problem-solving capability $S(0)$ is not that different among human beings. However, this justification cannot be applied to AI systems, since they can have very different $S(0)$ values.

To place learning at the center of intelligence is not a new opinion at all, although usually learning is put on the same level as problem-solving. In our analysis, learning capability is at the meta-level, while the various problem-solving capabilities are at the object-level. This difference provides another perspective on the "AI vs. AGI" contrast mentioned previously. Mainstream AI research takes intelligence as the ability of solving specific problems, and for each problem its solution depends on problem-specific features. AGI, on the contrary, focuses on the meta-problems, which are independent of the specific domain. In this way, the two approaches actually do not overlap or compete, but complement each other.

For G-Type systems, it makes the discussion more clear by calling its meta-level knowledge and procedures "intelligence" (which is mostly built-in and independent of the system's experience), while calling its object-level knowledge and procedures "beliefs and skills" (which are mostly acquired from the system's experience). When we say such a system has reached "human-level", we mean its

meta-level knowledge and procedures resemble those of the human mind, although its object-level beliefs and skills can overlap with that of a human being to an arbitrary extent [2].

The learning of meta-level knowledge in an AI system is possible, but there are several major issues that are rarely touched on in the relevant discussions:

- Although the distinction between object-level and meta-level exists in many AI systems, its exact standard depends on the design of the system, such that the functions carried out as meta-learning in one system may correspond to a type of object-level learning in another system.
- Some proposed "meta-learning" approaches basically suggest trying various designs and keeping the best, but this is usually unrealistic because of the time–space resources it demands and the risks it brings.
- A common misunderstanding is to equalize "recursive self-improving" with "modifying the system's own source-code". Certain programming languages, such as Lisp and Prolog, have provided the source-code modification and generation functions for many years; however, the use of these functions does not cause a fundamental difference, as the same effects can be achieved in another programming language by changing certain data, as the "data vs. code" distinction is also based on implementation considerations, not a theoretical boundary.
- The learning of meta-level knowledge and skills cannot be properly handled by the existing machine learning techniques, which are designed for object-level tasks [12]. The same difference happens in the human mind: as an individual, our mental mechanisms are mostly innate, and can be adjusted slowly in limited ways; it is only when discussed at the scale of the species that they are "learned" via evolution, but that happens at a much slower speed and a much higher price (a bad change is usually fatal for an individual). Given this fundamental difference, it is better to call the meta-level learning processes in a species "evolution", reserving the word "intelligence" for the object-level learning processes of an individual.

Thus far, we have not seen any convincing evidence for the possibility of a R-Type system. Although the existence of "exponential growth" is often claimed by the supporters of singularity, its evidence is never about the capability of a single system achieved by self-improvement. Although "intelligence" is surely a matter of degree, there is no evidence that the "level of intelligence" is an endless ladder with many steps above the "human-level". The existence of "super-intelligence" [17] is often argued using analogy from the existence of intelligence below the human-level (while mixing the object-level improvement with the meta-level improvement). Here, the situation is different from the above $S(t)$ values, which obviously can be increased from any point by adding knowledge and skills, as well as computational resources. At the meta-level, "above human" should mean a completely different thinking mechanism which better serves the purpose of adaptation. Of course, we cannot deny such a possibility, but have not seen any solid evidence.

Our hypothesis here is that the measurement of intelligence is just like a membership function of a fuzzy concept, with an upper-bound not much higher than the human-level. Furthermore, we believe it is possible for AGI to reach the same level; that is to say, there is a general notion of "intelligence" which can be understood and reproduced in computers. After that, the systems can still be improved, either by humans or by themselves, although there will not be a "super-intelligence" based on a fundamentally different mechanism.

Can a computer become smarter than a human? Sure, as this has already happened in many domains, if "smarter" means having a higher problem-solving capability. Computers have done better than human beings on many problems, but this result alone is not enough to earn them the title "intelligence", otherwise arithmetic calculators and sorting algorithms should also be considered as intelligent in their domains. To trivialize the notion of "intelligence" in this way will only lead to the need for a new notion to indicate the G-type systems. In fact, this is exactly why the phrase "AGI" was introduced. Similarly, the literal meaning of "machine learning" covers the G-type systems well, and the field of machine learning was also highly diverse at the beginning; however, now the phrase

"machine learning" is usually interpreted as "function approximation using statistics", which only focuses on the P-type systems, so we have to use a different name to avoid misunderstanding [12,18].

Since, in a G-Type or R-Type system, $S(t)$ can grow to an arbitrary level, they can be "smarter than humans"; however, this does not mean that AGI will do better on every problem, usually for reasons such as sensor, actuator, or experience. On this matter, there is a fundamental difference between the G-Type systems and the R-Type systems: for the former, since the meta-level knowledge remains specified by its designer, we still understand how the system works in principle, even after its $S(t)$ value is far higher than what can be reached by a human being. On the contrary, if there were really such a thing as an R-Type system, it would reach a point beyond which we cannot even understand how it works.

Since we do not believe an R-Type system can exist, we do not think "singularity" (in its original sense) can happen. However, we do believe AGI systems can be built with meta-level capability comparable to that of a human mind (i.e., neither higher nor lower, although not necessarily identical), and object-level capability higher than that of a human mind (i.e., in total score, although not on every task). These two beliefs do not contradict each other. Therefore, although we agree with Braga and Logan [1] on the impossibility of a "singularity", our reasons are completely different.

3. What an AGI Can Do

To support our conclusions in the previous section, here we briefly introduce our own AGI project, NARS (Non-Axiomatic Reasoning System). First, we roughly describe how the system works, and then explain how the features listed by Braga and Logan [1] as essential for intelligence are produced in NARS.

The design of NARS has been described in two research monographs [19,20] and more than 60 papers, most of which can be downloaded at https://cis.temple.edu/~pwang/papers.html. In 2008, the project became open source, and since then has had more than 20 releases. The current version can be downloaded with documents and working examples at http://opennars.github.io/opennars/.

Given the complexity of NARS, as well as the nature and length of this article, here we merely summarize the major ideas in NARS' design in a non-technical language. Everything mentioned in the following about NARS has been implemented in computer, and described in detail in the aforementioned publications.

3.1. NARS Overview

Our research is guided by the belief that knowledge about human intelligence (HI) can be generalized into a theory on intelligence in general (GI), which can be implemented in a computer to become computer intelligence (CI, also known as AI), in that it keeps the cognitive features of HI, but without its biological features [21]. In this way, CI is neither a perfect duplicate nor a cheap substitute of HI, but is "parallel" to it as different forms of intelligence.

On how CI should be similar to HI, mainstream AI research focuses on what problems the system can solve, while our focus is on what problems the system can learn to solve. We do not see intelligence as a special type of computation, but as its antithesis, in the sense that "computation" is about repetitive procedures in problem solving, where the system has sufficient knowledge (an applicable algorithm for the problem) and resources (computational time and space required by the algorithm); "intelligence" is about adaptive procedures in problem solving, where the system has insufficient knowledge (no applicable algorithm) and resources (shortage of computational time and/or space) [22].

Based on such a belief, NARS is not established on the theoretical foundations of mainstream AI research (which mainly consist of mathematical logic, probability theory and the theory of computability and computational complexity), but on a theory of intelligence in which the Assumption of Insufficient Knowledge and Resources (hereafter AIKR) is taken as a fundamental constraint to be respected rigorously. Under AIKR, an adaptive system cannot merely execute the programs provided by its human designers, but must use its past experience to predict the future (although the past and

the future are surely different), and use its available resources (supply) to best satisfy the pending demands (although the supply is always less than the demand).

To realize the above ideas in a computer system, NARS is designed as a reasoning system to simulate the human mind at the conceptual level, rather than at the neural level, meaning that the system's internal processing can be described as inference about conceptual relations.

Roughly speaking, the system's memory is a conceptual network, with interconnected concepts each identified by an internal name called a "term". In its simplest form, a term is just a unique identifier, or label, of a concept. To make the discussion natural, English nouns such as "bird" and "robin" are often used to name the terms in examples. A conceptual relation in NARS is taken to be a "statement", and its most basic type is called "inheritance", indicating a specialization-generalization relation between the terms and concepts involved. For example, the statement "*robin* $\rightarrow$ *bird*" roughly expresses "Robin is a type of bird".

NARS is a reasoning system that uses a formal language, Narsese, for knowledge representation, and has a set of formal inference rules. Even so, it is fundamentally different from the traditional "symbolic" AI systems in several key aspects.

One such aspect is semantics, i.e., the definition of meaning and truth. Although the Narsese term *bird* intuitively corresponds to the English word "bird", the meaning of the former is not "all the birds in the world", but rather what the system already knows about the term at the moment according to its experience, which is a stream of input conceptual relations. Similarly, the truth-value of "*robin* $\rightarrow$ *bird*" is not decided according to whether robins are birds in the real world, but rather the extent to which the term *robin* and the term *bird* have the same relations with other terms, according to evidence collected from the system's experience. For a given statement, available evidence can be either positive (affirmative) or negative (dissenting), and the system is always open to new evidence in the future.

A statement's truth-value is a pair of real numbers, both in [0, 1], representing the evidential support a statement obtains. The first number is "frequency", defined as the proportion of positive evidence to all available evidence. The second number is "confidence", defined as the proportion of currently available evidence to all projected available evidence at a future moment, after a new, constant amount of evidence is collected. Defined in this way, *frequency* is similar to probability, although it is only based on past observation and can change over time. *Confidence* starts at 0 (completely unknown) and gradually increases as new evidence is collected, but will never reach its upper-bound, 1 (completely known). NARS never treats an empirical statement as an axiom or absolute truth with a truth-value immune from future modification, which is why it is "non-axiomatic".

This "experience-grounded semantics" [23] of NARS bases the terms and statements of NARS directly on its experience, i.e., the system's record of its interaction with the outside world, without a human interpreter deciding meaning and truth. The system's beliefs are summaries of its experience, not descriptions of the world as it is. What a concept means to the system is determined by the role it plays in the system's experience, as well as by the attention the system currently pays to the concept, because under AIKR, when a concept is used, the system never takes all of its known relations into account. As there is no need for an "interpretation" provided by an observer, NARS cannot be challenged by Searle's "Chinese Room" argument as "only having syntax, but no semantics" [15,23].

In each inference step, NARS typically takes two statements with a shared term as its premises, and derives some conclusions according to the evidence provided by the premises. The basic inference rules are syllogistic, whose sample use-cases are given in Table 1.

Table 1. Sample steps of basic syllogistic inference.

Type	Deduction	Induction	Abduction
Premise 1	$robin \rightarrow bird$	$robin \rightarrow bird$	$robin \rightarrow [flyable]$
Premise 2	$bird \rightarrow [flyable]$	$robin \rightarrow [flyable]$	$bird \rightarrow [flyable]$
Conclusion	$robin \rightarrow [flyable]$	$bird \rightarrow [flyable]$	$robin \rightarrow bird$

The table includes three cases involving the same group of statements, where "$robin \rightarrow bird$" expresses "Robin is a type of bird", "$bird \rightarrow [flyable]$" expresses "Bird can fly", and "$robin \rightarrow [flyable]$" expresses "Robin can fly". For complete specification of Narsese grammar, see [20].

Deduction in NARS is based on the transitivity of the *inheritance* relation, that is, "if A is a type of B, and B is a type of C, then A is a type of C." This rule looks straightforward, except that since the two premises are true to differing degrees, so is the conclusion. Therefore, a truth-value function is part of the rule, which uses the truth-values of the premises to calculate the truth-value of the conclusion [20].

The other cases are induction and abduction. In NARS, they are specified as "reversed deduction" as in [24], obtained by switching the conclusion in deduction with one of the two premises, respectively. Without the associated truth-values, induction and abduction look unjustifiable, but according to experience-grounded semantics, in both cases the conclusion may get evidential support from the premise. Since each step only provides one piece of evidence, inductive and abductive conclusions normally have lower confidence than deductive conclusions.

NARS has a revision rule which merges evidence from distinct sources for the same statement, so the confidence of its conclusion is higher than that of the premises. Revision can also combine conclusions from different types of inference, as well as resolve contradictions by balancing positive and negative evidence.

To recognize complicated patterns in experience, Narsese has compound terms that each are constructed from some component terms, and NARS has inference rules to process these compounds. Certain terms are associated with the operations of sensors and actuators, therefore the system can represent procedural knowledge on how to do things, rather than just to talk about them. The grammar rules, semantic theory, and the inference rules altogether form the Non-Axiomatic Logic (NAL), the logic part of NARS [19,20].

From an user's point of view, NARS can accept three types of task:

- **Judgment:** A judgment is a statement with a given truth-value, as a piece of new knowledge to be absorbed into the system's beliefs. The system may revise the truth-value of a judgment according to its previous belief on the matter, add it into the conceptual network, and carry out spontaneous inference from it and the relevant beliefs to reveal its implications, recursively.
- **Goal:** A goal is a statement to be realized by the system. To indicate the extent of preference when competing with other goals, an initial "desire-value" can be given. When the desire-value of a goal becomes high enough, it will either directly trigger the execution of the associated operation, or generate derived goals according to the relevant beliefs.
- **Question:** A question can ask the truth-value or desire-value of a statement, which may contain variable terms to be instantiated. A question can be directly answered by a matching belief or desire, or generate derived questions according to the relevant beliefs.

These tasks and the system's beliefs (judgments that are already integrated into the system's memory) are organized into concepts according to the terms appearing in them. For example, tasks and beliefs on statement "$robin \rightarrow bird$" are referred from concept *robin* and concept *bird*. Each task only directly interacts with (i.e., being used as premises with) beliefs within the same concept, so every inference step happens within a concept.

As the system usually does not have the processing time and storage space to carry out the inference for every task to its completion (by exhaustively interacting with all beliefs in the concept), each data item (task, belief, and concept) has a priority value associated to indicate its share in resource

competition. These priorities can take user specified initial values, and then be adjusted by the system according to the feedback (such as the usefulness of a belief, etc.).

NARS runs by repeating the following working cycle:

1. Select a concept in the system's memory probabilistically, biased by the priority distributions among concepts. Every concept has a chance to be selected, although concepts with high priority have higher chances.
2. Select a task and a belief from the concept, also probabilistically as above. Since the two share the same term identifying the concept, they must be relevant in content.
3. Carry out a step of inference using the selected task and belief as premises. Based on the combination of their types, the corresponding inference rules will be triggered, which may provide immediate solutions to the task, or (more likely) derived tasks whose solutions will contribute to the solution of the original task.
4. Adjust the priority values of the processed items (belief, task, and concept) according to the available feedback from this inference step.
5. Process the new tasks by selectively adding them into the memory and/or reporting them to the user.

3.2. Properties of NARS

Although the above description of NARS is brief and informal, it still provides enough information for some special properties of the system to be explained. A large part of [1] is to list certain "essential elements of or conditions for human intelligence" and claim they cannot be produced in AI systems. In this subsection, we describe how the current implementation of NARS generates these features (marked using bold font), at least in their preliminary forms. As each of them typically has no widely accepted definition, our understanding and interpretation of it will be inevitably different from that of other people, although there should be enough resemblance for this discussion to be meaningful.

The claim "Computers, like abacuses and slide rules, only carry out operations their human operators/programmers ask them to do, and as such, they are extensions of the minds of their operators/programmers." [1] is a variant of the so-called "Lady Lovelace's Objection" analyzed and rejected by Turing [3]. To many traditional systems, this claim is valid, but it is no longer applicable to adaptive systems like NARS. In such a system, what will be done for a problem not only depends on the initial design of the system, but also on the system's **experience**, which is the history of the system's interaction with the environment. In this simple and even trivial sense, every system has an experience, but whether it is worth mentioning is a different matter.

If a problem is given to a traditional system, and after a while a solution is returned, then if the same problem is repeated, the solving process and the solution should be repeated exactly, as this is how "computation" is defined in theoretical computer science [25]. In NARS, since the processing of a task will more or less change the system's memory irreversibly, and the system is not reset to a unique initial state after solving each problem, a repeated task will (in principle) be processed via a more or less different path—the system may simply report the previous answer without redoing the processing. Furthermore, the co-existing problem-solving processes may change the memory to make some concepts more accessible to suggest a different solution that the system had not previously considered. For familiar problems, the system's processing usually becomes stable, although whether a new problem instance belongs to a known problem type is always an issue to be considered from time to time by the system, rather than taken for granted.

Therefore, to accurately predict how NARS will process a task, to know its design is not enough. For the same reason, it is no longer correct to see every problem as being solved by the designer, because given the same design and initial content in memory, different experiences will actually lead to very different systems, in terms of their concepts, beliefs, skills, etc. Given this situation, it makes more sense to see the problems as solved by the system itself, even though this **self** is not coming

out of nowhere magically or mythically, but rooted in the initial configuration and shaped by the system's experience.

NARS has a *self* concept as a focal point of the system's self-awareness and self-control. Like all concepts in NARS, the content of *self* mainly comes from accumulated and summarized experience about the system itself, although this concept has special innate (built-in) relations with the system's primary operations. It means at the very beginning the system's "self" is determined by "what I can do" and "what I can feel" (since in NARS perception is a special type of operation), but gradually it will learn "what is my relation with the outside objects and systems", so the concept becomes more and more complicated [26]. Just like NARS' knowledge about the environment, its knowledge about itself is always uncertain and incomplete, but we cannot say that it has no sense of itself.

NARS can be equipped with various sensors, and each type of sensor expends the system's experience to a new dimension by adding a new sensory channel into the system where a certain type of signals are recognized, transformed into Narsese terms, then organized and generalized via a perceptive process to enter the system's memory. The sensors can be on either the external environment or the internal environment of the system, where the latter provides self-awareness about what has been going on within the system. Since the internal experience is limited to significant internal events only, in NARS the conscious/unconscious distinction can be meaningfully drawn, according to whether an internal event is registered in the system's experience and becomes a belief expressed in Narsese.

The interactions between unconscious and conscious mental events were argued to be important by Freud [27], and this opinion is supported by recent neuroscientific study [28]. As only significant events within NARS enter the system's (conscious) experience, the same conclusion holds for NARS. A common misunderstanding about NARS-like systems is that all events in such a system must be conscious to the system, or that the distinction between conscious and unconscious events is fixed. Neither is correct in NARS, mainly because of AIKR, as an event can change its conscious status merely because of its priority level adjustments [26]. This interaction in NARS has a purely functional explanation that has little dependency on the detail of human neural activities.

As far as the system can tell consciously, its decisions are made according to its own **free will**, rather than by someone else or according to certain predetermined procedures, simply because the system often has to deal with problems for which no ready made solutions are there, so it has to explore the alternatives and weigh the pros and cons when a decision is made, all by itself. For an omniscient observer, all the decisions are predetermined by all the relevant factors collectively, but even from that viewpoint, it is still the decision by the system, not by its designer, who cannot predetermine the experience factor.

Given the critical role played by experience, it is more natural to accredit certain responsibility and achievement to the system, rather than to the designer. The system's beliefs are not merely copies of what it was taught by the user, but summaries of its experience. These beliefs include moral **judgments** (beliefs about what are good and what are bad, according to its desires and goals), **wisdom** (beliefs that guide the system to achieve its goals), **intuition** (beliefs whose source is too complicated or vague to recall), and so on. These beliefs are often from the view point of the system as they are produced by its unique experience. Even so, the beliefs of NARS will not be purely subjective, as the system's communication with other systems provide social experience for it, and consequently the relevant beliefs will have certain objective (or more accurately, "intersubjective") flavors in it, in the sense that it is not fully determined by the system's idiosyncratic experience, but strongly influenced by the community, society, or culture that the system belongs to.

Not only should the beliefs in NARS be taken as "of the system's own", but also the **desires** and **goals**. The design of NARS does not presume any specific goal, so all the original goals come from the outside, that is, the designer or the user. NARS has a goal derivation mechanism that generates derived goals from the existing (input or derived) goals and the relevant beliefs. Under AIKR, a derived goal G_2 is treated independently of its "parent" goal G_1, so in certain situation it may become more

influential than G_1, and can even suppress it. Therefore, NARS is not completely controlled by its given goals, but also by the other items in its experience, such as the beliefs on how the goals can be achieved. This property is at the core of autonomy, originality, and creativity, although at the same time it raises a challenge on how to make the system behave according to human interests [29].

As an AGI, the behavior of NARS is rarely determined by a single goal, but often by a large number of competing and even conflicting goals and desires. When an operation is executed, it is usually driven by the "resultant" of the whole motivation complex, rather than by one motivation [29]. This motivation complex develops over time, and also contributes greatly to the system's self identity. In different contexts, we may describe the difference aspects of this complex as **purpose**, **objective**, **telos**, and even **caring**.

Desires and goals with special content are often labeled using special words. For example, when its social experience become rich and complicated enough, NARS may form what may be called "**values**" and "**morality**", as they are about how a system should behave when dealing with other systems. When the content of a goal drives the system to explore an unknown territory without explicitly specified purpose, we may call it "**curiosity**". However, the fact that we have a word for a phenomenon does not mean that it is produced by an independent mechanism. Instead, the phenomena discussed above are all generated by the same process in NARS, although each time we selectively discuss some aspects of it, or set up its context differently.

A large part of argument for the impossibility of AGI in [1] is organized around the "figure–ground" metaphor, where a key ingredient of the "ground" is **emotion**, which is claimed to be impossible in computers. However, this repeated claim only reveals the lack of knowledge of the authors about the current AGI research, as many AGI projects have emotion as a key component [30–32]. In the following, we only introduce the emotional mechanism in NARS, which is explained in detail in [33].

In NARS, emotion starts as an appraisal of the current situation, according to the system's desires. On each statement, there is a truth-value indicating the current situation, and a desire-value indicating what the system desires the situation to be, according to the relevant goals. The proximity of these two values measures the system's "satisfaction" on this matter. At the whole system level, there is an overall satisfaction variable that accumulates the individual measurements on the recently processed tasks, which will produce a positive or negative appraisal of the overall situation. That is, the system will have positive emotion if the reality agrees to its desires, and negative emotion if the reality disagrees to its desires.

These satisfaction values can be "felt" by the system's inner sensors, as well as be involved in the system's self-control. For instance, events associated with strong (positive or negative) emotion will get more attention (and therefore more processing resources) than the emotionally neutral events. When the system is in positive emotion, it is more acceptive to new tasks (meaning it devotes to them more resources). A strong emotion for someone or something corresponds to the phenomenon of "**passion**".

At the moment, we are extending the emotional mechanism in several ways, including to further distinguish different emotions (such as "**pleasure**" and "**joy**" at the positive side, and "scare" and "anger" at the negative side), to use emotion in communication, to control the effect of emotion in decision making, and so on.

Among the features in the list of [1], the only ones that have not been directly addressed in the previous publications and implementations of NARS are **imagination**, **aesthetics**, and **humor**. We do have plan to realize them in NARS, but will not discuss it in this article.

In summary, we agree the features listed in [1] are all necessary for AGI, and we also agree that the mainstream AI techniques cannot generate them. However, we disagree with the conclusion that they cannot be generated in computer systems at all. On the contrary, most of them have been realized in NARS in their preliminary form, and NARS is not the only AGI project that has addressed these topics.

Of course, we are not claiming that all these phenomena have been fully understood and perfectly reproduced in NARS or other AGI systems. On the contrary, the study of them is still in an early stage,

and there are many open problems. However, the results so far have at least shown their possibility, or, more accurately, their inevitability, to appear in AGI systems. As shown above, in NARS these features are not added in one by one for their own sake, but are produced altogether from the design of NARS, usually as implications of AIKR.

A predictable objection to our above conclusions is to consider the NARS versions of these features to be "fake", as they are not identical to the human versions here or there. Once again, it goes back to the understanding of "AI" and how close it should be to human intelligence. Take emotion as an example: even when fully developed, the emotions in NARS will not be identical to human emotions, nor will they be accompanied by the physiological processes that are intrinsic ingredients of human emotion. However, these differences cannot be used to judge emotions in AGI as fake, as long as "emotion" is taken as a generalization of "human emotion" by keeping the functional aspects but not the biological ones.

Here is what we see as our key difference with Braga and Logan [1]: while we fully acknowledge that the original and current usage of the features they listed are tied to the human mind/brain complex, we believe it is both valid and fruitful to generalize these concepts to cover non-human and even non-biological systems, as their core meaning is not biological, but functional. Such a belief is also shared by many other researchers in the field, although how to accurately define these features is still highly controversial.

4. Conclusions

In this article, we summarize our opinions on AI, AGI, and singularity, and use our own AGI system as evidence to support these opinions. The purpose of this article is not to introduce new technical ideas, as the aspects of NARS mentioned above have all been described in our previous publications. Since many people are not familiar with the results of AGI research (as shown in this Special Issue of *Information*), we consider it necessary to introduce them to clarify the relevant notions in the discussion on what can be achieved in AI systems.

We agree with the lead article [1] that the mainstream AI techniques will not lead to "Strong AI" or AGI that is comparable to human intelligence in general, or to a "Singularity" where AI becomes "smarter than human", partly because these techniques fail to reproduce a group of essential characteristics of intelligence.

However, we disagree with their conclusion that AGI is completely impossible because the human mind is fundamentally different from digital computers [1], partly because most of the characteristics they listed have already been partially realized in our system. In our opinion, there are the following major issues in their argument:

- Their understanding of "intelligence" is highly anthropocentric. As few AGI researcher aims at such an objective, the "AGI research" they criticize does not exist [2].
- Their understanding of computer is oversimplified and reductionist, and only corresponds to a special way of using computer. Even though this is indeed the most common way for a computer system to be built and used at the present time, it is not the only possible way [34].
- Their understanding of AGI mostly comes from outsiders' speculations, rather than from the actual research activity in the field. Although it is perfectly fine for an outsider to criticize AGI research, such a criticism is valid only when it is based on the reality of the field.

As with respect to the topics under discussion, our positions are:

- AGI should be conceived as a computer system that is similar to human intelligence in principles, mechanisms, and functions, but not necessarily in internal structure, external behaviors, or problem-solving capabilities. Consequently, as another form of intelligence, AGI will roughly be at the same level of competence as human intelligence, neither higher nor lower. As in concrete problem-solving capability, AGI is not always comparable to human intelligence, since they may deal with different problems in different environments.

- To achieve AGI, new theories, models, and techniques are needed. The current mainstream AI results will not naturally grow in this direction, because they are mainly developed according to the dogma that intelligence is problem-solving capability, which does not correspond to AGI, but a fundamentally different objective, with different theoretical and practical values.
- Even when AGI is achieved, it does not lead to a singularity beyond which intelligent computer systems become completely incomprehensible, unpredictable, and uncontrollable. On the contrary, the achieving of AGI means the essence of intelligence has been captured by humans, which will further guide the use of AGI to meet to human values and needs.

AGI research is still in an early stage, and opinions from all perspectives are valuable, although it is necessary to clarify the basic notions to set up a minimum common ground, so the voices will not talk past each other. For this reason, the current Special Issue of *Information* is a valuable effort.

Author Contributions: P.W. conceived the article and submitted an abstract; P.W. drafted the article after discussions with K.L. and Q.D.; K.L. and Q.D. revised the draft; Q.D. corrected the English.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156, doi:10.3390/info8040156.
2. Wang, P. What do you mean by "AI". In Proceedings of the First Conference on Artificial General Intelligence, Memphis, TN, USA, 1–3 March 2008; pp. 362–373.
3. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *LIX*, 433–460.
4. McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955. Available online: http://www-formal.stanford.edu/jmc/history/dartmouth.html (accessed on 4 April 2018).
5. Feigenbaum, E.A.; Feldman, J. *Computers and Thought*; McGraw-Hill: New York, NY, USA, 1963.
6. Newell, A.; Simon, H.A. GPS, a program that simulates human thought. In *Computers and Thought*; Feigenbaum, E.A., Feldman, J., Eds.; McGraw-Hill: New York, NY, USA, 1963; pp. 279–293.
7. Fuchi, K. The significance of fifth-generation computer systems. In *The Age of Intelligent Machines*; Kurzweil, R., Ed.; MIT Press: Cambridge, MA, USA, 1990; pp. 336–345.
8. Roland, A.; Shiman, P. *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*; MIT Press: Cambridge, MA, USA, 2002.
9. Luger, G.F. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 6th ed.; Pearson: Boston, MA, USA, 2008.
10. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444, doi:10.1038/nature14539.
12. Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*; Cambridge University Press: New York, NY, USA, 2012.
13. Pennachin, C.; Goertzel, B. Contemporary approaches to artificial general intelligence. In *Artificial General Intelligence*; Goertzel, B., Pennachin, C., Eds.; Springer: New York, NY, USA, 2007; pp. 1–30.
14. Wang, P.; Goertzel, B. Introduction: Aspects of artificial general intelligence. In *Advance of Artificial General Intelligence*; Goertzel, B., Wang, P., Eds.; IOS Press: Amsterdam, The Netherlands, 2007; pp. 1–16.
15. Searle, J. Minds, brains, and programs. *Behav. Brain Sci.* **1980**, *3*, 417–424.
16. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Penguin Books: New York, NY, USA, 2006.
17. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
18. Wang, P.; Li, X. Different Conceptions of Learning: Function Approximation vs. Self-Organization. In Proceedings of the Ninth Conference on Artificial General Intelligence, New York, NY, USA, 16–19 July 2016; pp. 140–149.
19. Wang, P. *Rigid Flexibility: The Logic of Intelligence*; Springer: Dordrecht, The Netherlands, 2006.
20. Wang, P. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*; World Scientific: Singapore, 2013.
21. Wang, P. Theories of Artificial Intelligence—Meta-theoretical considerations. In *Theoretical Foundations of Artificial General Intelligence*; Wang, P., Goertzel, B., Eds.; Atlantis Press: Paris, France, 2012; pp. 305–323.

22. Wang, P. The Assumptions on Knowledge and Resources in Models of Rationality. *Int. J. Mach. Conscious.* **2011**, *3*, 193–218.
23. Wang, P. Experience-grounded semantics: A theory for intelligent systems. *Cognit. Syst. Res.* **2005**, *6*, 282–302.
24. Peirce, C.S. *Collected Papers of Charles Sanders Peirce*; Harvard University Press: Cambridge, MA, USA, 1931; Volume 2.
25. Hopcroft, J.E.; Motwani, R.; Ullman, J.D. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed.; Addison-Wesley: Boston, MA, USA, 2007.
26. Wang, P.; Li, X.; Hammer, P. Self in NARS, an AGI System. *Front. Robot. AI* **2018**, *5*, 20, doi:10.3389/frobt.2018.00020.
27. Freud, S. *The Interpretation of Dreams*; Translated by James Strachey from the 1900 Edition; Avon Books: New York, NY, USA, 1965.
28. Dresp-Langley, B. Why the Brain Knows More than We Do: Non-Conscious Representations and Their Role in the Construction of Conscious Experience. *Brain Sci.* **2012**, *2*, 1–21.
29. Wang, P. Motivation Management in AGI Systems. In Proceedings of the Fifth Conference on Artificial General Intelligence, Oxford, UK, 8–11 December 2012; pp. 352–361.
30. Bach, J. Modeling Motivation and the Emergence of Affect in a Cognitive Agent. In *Theoretical Foundations of Artificial General Intelligence*; Wang, P., Goertzel, B., Eds.; Atlantis Press: Paris, France, 2012; pp. 241–262.
31. Franklin, S.; Madl, T.; D'Mello, S.; Snaider, J. LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *IEEE Trans. Auton. Ment. Dev.* **2014**, *6*, 19–41.
32. Rosenbloom, P.S.; Gratch, J.; Ustun, V. Towards Emotion in Sigma: From Appraisal to Attention. In Proceedings of the Eighth Conference on Artificial General Intelligence, Berlin, Germany, 22–25 July 2015; pp. 142–151.
33. Wang, P.; Talanov, M.; Hammer, P. The Emotional Mechanisms in NARS. In Proceedings of the Ninth Conference on Artificial General Intelligence, New York, NY, USA, 16–19 July 2016; pp. 150–159.
34. Wang, P. Three fundamental misconceptions of artificial intelligence. *J. Exp. Theor. Artif. Intell.* **2007**, *19*, 249–268.