Chapter 1

Theories of Artificial Intelligence — Meta-theoretical considerations

Pei Wang
Temple University, Philadelphia, USA
http://www.cis.temple.edu/∼pwang/

This chapter addresses several central meta-theoretical issues of AI and AGI. After analyzing the nature of the field, three criteria for desired theories are proposed: *correctness*, *concreteness*, and *compactness*. The criteria are clarified in the AI context, and using them, the current situation in the field is evaluated.

1.1. The problem of AI theory

Though it is a common practice for a field of science or engineering to be guided and identified by the corresponding theories, the field of Artificial Intelligence (AI) seems to be an exception. After more than half of a century since its formation, AI still has no widely accepted theory, and in the related discussions the following opinions are often heard:

- "*The best model of intelligence is the human brain itself (and all theories are merely poor approximations ...)*"
- "*There is no need for any new theory, since AI can be built according to* X *(depending on who said it, the* X *can be mathematical logic, probability theory, theory of computation, ...)*"
- "*A theory of AI has to be established piece by piece, and we are starting from* Y *(depending on who said it, the* Y *can be search, reasoning, learning, perception, actions, ...)*"
- "*There cannot be any good theory of intelligence (since intelligence is so complicated, though our work is obviously central to it ...)* "
- "*Theoretical debating is a waste of time (and we should focus on practical applications. For example, an intelligent system should be able to ...)*"
- "*A good theory only comes at the end of the research (so don't worry about it now, and it will come as long as we continue the current research on ...)*"

There is a historical reason for this situation. Though the idea of "thinking machine" can be traced further back in history, the field of AI was started from the realization that computers, though initially designed to do numerical calculations, can be made to carry out other mental activities, such as theorem proving and game playing, which are hard

intellectual problems that are usually considered as demanding "intelligence" [McCarthy *et al.* (1955); Feigenbaum and Feldman (1963)]. This "problem-oriented" attitude toward AI focuses on the problem-solving capability of a computer system, while does not care much for the underlying theory. Consequently, the early works in AI often showed the "Look, ma, no hands" syndrome — "A paper reports that a computer has been programmed to do what no computer program has previously done, and that constitutes the report. How science has been advanced by this work or other people are aided in their work may be unapparent." [McCarthy (1984)]. For such a work, "the question, Where's the AI? is a tough one to answer" [Schank (1991)].

To many AI researchers, the lack of a common theory is not an issue at all. As said by Minsky (1985), "Our minds contains processes that enable us to solve problems we consider difficult. 'Intelligence' is our name for whichever of those processes we don't yet understand." According to this opinion, a "theory of AI" is impossible *by definition*, since we cannot have a theory for "those processes we don't yet understand" — as soon as we have a good theory for such a process, it is no longer considered as AI anymore [Minsky (1985); Schank (1991)].

To get out of this annoying situation, in mainstream AI "intelligence" is treated as the collaboration of a group of loosely coupled functions, each of them can be separately specified in computational and algorithmic terms, implemented in computers, and use to solve certain practical problems [Marr (1982); Russell and Norvig (2010)]. In an influential AI textbook by Russell and Norvig (2010), it is written that in the late 1980s "AI adopts the scientific method", since "It is now more common to build on existing theories than to propose brand new ones ...". However, it is not mentioned that none of those "existing theories" were proposed with intelligence as the subject matter, nor has shown the potential of solving the problem of intelligence as a whole.

Though in this way the field has produced valuable results in the past decades, it still suffers from internal fragmentation [Brachman (2006)] and "paradigmatic mess" [Chandrasekaran (1990)], largely due to the lack of a common theoretical foundation. There have been many debates on the nature or objective of the field, or on what type of theory it should or can have [Wilks (1990); Bundy and Ohlsson (1990); Simon (1990); Kirsh (1991)].

Though the pursuit of unified theories of AI is widely considered as futile in the field, there is still a small number of AI researchers who believe that such a theory is possible, and worthwhile to be investigated. The best known work in this direction is the "Unified Theories of Cognition" by Newell (1990), in which he argued for the necessity for AI and cognitive science to have unified theories, and proposed his theory, which attempts to cover both AI and human intelligence. Similar attempts include the works of Albus (1991) and Pollock (2006).

In recent years, the term "Artificial General Intelligence" (AGI) is adopted by a group of AI researchers to emphasize the general-purpose and holistic nature of the "intelligence" they are after [Goertzel and Pennachin (2007); Wang and Goertzel (2007)]. Since AGI treats intelligence as a whole, there are more efforts to establish unified theories [Bach

(2009); Baum (2004); Bringsjord (2008); Cassimatis (2006); Franklin (2007); Goertzel (2009); Hutter (2005); Schmidhuber (2007); Wang (2006)], though none of them is mature or convincing enough to obtain wide acceptance in the field at the current moment [Bringsjord and Sundar G (2009)].

Even though the AGI community is more "AI-theory-oriented" than mainstream AI, not every AGI project is based on some theory about intelligence. As in mainstream AI, a project is often guided by one, or more than one, of the following considerations:

Practical problem-solving demands: Since intelligence is displayed in the problem-solving capability of a system, many projects target problems that currently can be solved by humans only. Such a system is usually designed and analyzed according to the theory of computation [Marr (1982); Hayes and Ford (1995)].

Knowledge about human intelligence: Since the human mind has the best-known form of intelligence, many projects aim at duplicating certain aspects of the human mind or brain. Such a system is usually designed and analyzed according to the theories in psychology or neuroscience [Newell (1990); Rumelhart and McClelland (1986)].

Available normative models: Since intelligence intuitively means "to do the right thing", many projects are designed and analyzed as models of rationality or optimization, according to mathematical theories like classical logic and probability theory, though usually with extensions and/or revisions [McCarthy (1988); Pearl (1988)].

Even the AGI projects that are based on certain theories on AI are moving in very different directions, mainly because of the difference in their theoretical foundations, as well as the influence of the above considerations. This collection provide a representative example of the diversity in the theoretical study of AGI.

Consequently, currently in the field of AI/AGI there are very different opinions on research goal [Legg and Hutter (2007); Wang (2008)], roadmap [McCarthy (2007); Goertzel *et al.* (2009)], evaluation criteria [Alvarado *et al.* (2002); Laird *et al.* (2009)], etc. Though each researcher can and should make decisions on the above issues for his/her own project, for the field as a whole this paradigmatic mess makes comparison and cooperation difficult, if not impossible.

In this chapter, I will not promote my own theory of AI (which is described in my previous publications [Wang (1995, 2006, 2010)]), nor to evaluate the other theories one by one, but to address the major *meta-level* issues about AI theories, such as

- What is the nature of an AI theory?
- How to evaluate an AI theory?
- Why do we lack a good theory?

This chapter attempts to clarify the related issues, so as to pave the way to a solid theoretical foundation for AGI, which is also the original and ultimate form of AI. For this reason, in the following "AI" is mainly used to mean "AGI", rather than the current mainstream practice.

1.2. Nature and content of AI theories

In a field of science or engineering, a "theory" usually refers to a system of concepts and statements on the subject matter of the field. Generally speaking, there are two types of theory:

Descriptive theory: Such a theory starts with certain observations in the field. The theory provides a generalization and explanation of the observations, as well as predictions for future events, so as to guide people's behaviors. The theories in natural science are the best examples of this type.

Normative theory: Such a theory starts with certain assumptions, then derives conclusions from them. When the assumptions are accepted as applicable in a field, all the conclusions should also be accepted as true. Mathematics and engineering theories are the best examples of this type.[a]

Though it is possible for these two types of theory to interweave (in the sense that parts of a theory may belong to the other type), for a theory as a whole its type is still usually clear. For example, modern physics uses a lot of mathematics in it, but it does not change the overall descriptive nature of the theories in physics. On the contrary, computer science is mainly based on normative theories on how to build and use computer systems, even though empirical methods are widely used to test the systems.[b]

What makes a "Theory of AI" special on this aspect is that it needs to be *both* descriptive and normative, in a certain sense.

AI studies the similarity and the difference between "The Computer and the Brain", as suggested by the title of von Neumann (1958). This research is directly driven by the observation that though the computer systems can take over human's mental labor in many situations (and often do a better job), there are nevertheless still many features of the human mental activities that have not been reproduced by computers. An AI theory should provide a bridge over this gap between "the Brain" and "the Computer", so as to guide the designing and building of computer systems that are similar to the human mind in its "mental power". "Intelligence" is simply the word whose intuitive meaning is the closest to the capability or property to be duplicated from the brain to the computer, though some people may prefer to use other words like "cognition", "mind", or "thinking". The choice of word here does not change the nature of this problem too much.

Given this objective, an AI theory must identify the (known or potential) similarities between two entities, "the Brain" and "the Computer", which are very different on many aspects. Furthermore, human intelligence is an existing phenomenon, while computer intelligence is something to be built, for which an accurate description does not exist at this

[a]In fields like economics and law, a "normative" theory or model specifies what people *should* do, often for *ethical* reasons. It is not what the word means here. Instead, in this chapter a "normative" theory specifies what people should do for *rational* reasons. This usage is common in the study of human reasoning and decision making, for example see Gabbay and Woods (2003).

[b]On this topic, I disagree with Newell and Simon's opinion on "Computer science as empirical inquiry" [Newell and Simon (1976)].

moment. Consequently, an AI theory should be *descriptive* with respect to human intelligence (not in all details, but in basic principles, functions and mechanisms), and at the same time, be *normative* to computer intelligence. That is, on one hand, the theory should summarize and explain how the human mind works, at a proper level and scope of description; on the other hand, it should guide the design and development of computer systems, so as to make them "just like the human mind", at the same level and scope of description.

A theory for this field is surely centered at the concept of "intelligence". Accurately speaking, there are three concepts involved here:

Human Intelligence (HI), the intelligence as displayed by human beings;
Computer Intelligence (CI), the intelligence as to be displayed by computer systems;
General Intelligence (GI), the general and common description of both HI and CI.

For the current discussion, HI can also be referred to as "natural intelligence", CI as "artificial intelligence", and GI simply as "intelligence", which also covers other concepts like "animal intelligence", "collective intelligence", "alien intelligence", etc., as special cases [Wang (2010)].

Roughly speaking, the content of the theory must cover certain mechanisms in the human mind (as the HI), then generalize and abstract them (to be the GI), and finally specify them in a computational form (to become the CI). No matter what names are used, the distinction and relationship among the three concepts are necessary for an AI theory, because the theory needs to identify the common properties between human beings and computer systems, while still to acknowledge their differences in other aspects.[c]

Now it is easy to see that in an AI theory, the part about HI is mostly descriptive, that about CI is mostly normative, and that about GI is both.

The human mind is a phenomenon that has been studied by many branches of science from different perspectives and with different focuses. There have been many theories about it in psychology, neuroscience, biology, philosophy, linguistics, anthropology, sociology, etc. When talking about HI, what AI researchers usually do is to selectively acquire concepts and conclusions from the other fields, and to reorganize them in a systematic way. As a result, we get a theory that summarizes certain observed phenomenon of the human mind. Such a theory is fundamentally *synthetic* and *empirical*, in that its conclusions are summaries of common knowledge on how the human mind works, so it is verified by comparing its conclusions to actual human (mental) activities. Here the procedure is basically the same as in natural science. The only special thing is the *selectivity* coming from the (different) understandings of the concept "intelligence": different researchers may include different phenomena within the scope of HI, which has no "natural" boundary.

[c]Some people may argue that AI researchers are only responsible for the CI part of the picture, because the HI part should be provided by psychologists, and the GI part should be covered by a "theory of general intelligence", contributed by philosophers, logicians, mathematicians, and other researchers working on general and abstract systems. Though there is some truth in this argument, at the current time there is no established theory of GI that we AI researchers can accept as guidance, so we have to work on the whole picture, even though part of it is beyond our career training.

On the contrary, a theory about CI has to be normative, since this phenomenon does not exist naturally, and the main function of the theory is to tell the practitioners how to produce it. As a normative theory, its basic assumptions come from two major sources: knowledge of intelligence that describes what *should* be done, and knowledge of computer that describes what *can* be done. Combined together, this knowledge can guide the whole design and development process, by specifying the design objective, selecting some theoretical and technical tools, drawing a blueprint of the system's architecture, planning a development roadmap, evaluating the progress, and verifying the results. Here the procedure is basically the same as in engineering. The only special thing is the *selectivity* coming from the (different) understandings of the concept "intelligence": different researchers may define the concept differently, which will change everything in the following development.

As the common generalization of HI and CI, a theory of GI is both descriptive and normative. On one hand, the theory should explain how human intelligence works as a special case, and on the other hand, it should describe how intelligence works in general, so as to guide how an intelligent computer system should be designed. Therefore, this theory should be presented in a "medium-neutral" language that does not assume the special details of either the human brain or the computer hardware. At the same time, since it is less restricted by the "low-level" constraints, this part of the theory gives the researchers the largest freedom, compared to the HI and the CI part. Consequently, this is also where the existing theories differ most from each other — the differences among the theories are not much on *how* the brain, mind, or computer works, but on *where* the brain and the machine should be similar to each other [Wang (2008)].

In the textbook by Russell and Norvig (2010), different approaches toward AI are categorized according to whether they are designed to be *thinking* or *acting* "humanly" or "rationally". It seems that the former is mainly guided by descriptive theories, while the latter by normative theories. Though such a difference indeed exists, it is more subtle than what these two words suggest. Since the basic assumptions and principles of all models of rationality come from abstraction and idealization of the human thinking process, "rationally" thinking/acting is actually a special type of "humanly" thinking/acting. For example, though the "Universal AI" model AIXI by Hutter (2005) is presented in a highly abstract and mathematical form, its understanding of "intelligence" is still inspired and justified according to certain opinions about the notion in psychology [Legg and Hutter (2007)]. On the other extreme, though Hawkins' HTM model of intelligence is based on certain neuroscientific findings, it is not an attempt to model the human brain in all aspects and in all details, but to *selectively* emulate certain mechanisms that are believed to be "the crux of intelligence" [Hawkins and Blakeslee (2004)]. Therefore, the difference between AIXI and HTM, as well as among the other AGI models, is not on whether to learn from the human brain/mind (the answer is always "yes", since it is the best-known form of intelligence), or whether to idealize and simplify the knowledge obtained from the human brain/mind (the answer is also always "yes", since a computer cannot become identical to the brain in all aspects), but on *where* to focus and *how much* to abstract and generalize.

From the same knowledge about the human mind, there are many meaningful ways to establish a notion of HI, by focusing on different aspects of the phenomena; from the same notion of HI, there are many meaningful ways to establish a notion of GI, by describing intelligence on different levels, with different granularities and scopes; from the same notion of GI, there are many meaningful ways to establish a notion of CI, by assuming different hardware/software platforms and working environments. The systems developed according to different notions will surely have different properties and practical applications, and are "similar to the human mind" in different senses. Unless one commits to a particular definition of intelligence, there is no absolute standard to decide which of these ways is "the correct way" to establish a theory of AI.

The current collection to which this chapter belongs provides a concrete example for this situation: though the chapter authors all use the notion of "intelligence", and are explaining related phenomena, the theories they proposed are very different. It is not necessarily the case that at most one of the theory is "correct" or really captures intelligence "as it is", while all the others are "wrong", since each of them represents a certain perspective; nor can the issue be resolved by pooling the perspectives altogether, because they are often incommensurable, due to the usage of different concepts. This diversity is a major source of difficulty in theoretical discussions of AI.

1.3. Desired properties of a theory

Though there are reasons for different AI theories to be proposed, it does not mean that all of them are equally good. The following three desired properties of a scientific theory are proposed and discussed in my own theory of intelligence [Wang (2010)] (Section 6.2):

- *Correctness*: A theory should be supported by available evidence.
- *Concreteness*: A theory should be instructive in problem solving.
- *Compactness*: A theory should be simple.

Though these properties are proposed for scientific theories in general, in this chapter they will be discussed in the context of AI. Especially, let us see what they mean for an AI theory that is both descriptive (for human minds) and normative (for computer systems).

Correctness

Since the best-known form of *intelligence* is human intelligence, an AI theory is *correct* if it is supported by the available knowledge about the human mind. In this aspect, AI is not that different from any natural science, in that the correctness of a theory is verified empirically, rather than proved according to some priori postulates. Since the study of the human mind has been going on in many disciplines for a long time, AI researchers often do not need to carry out their own research on human subjects, but to inherit the conclusions from the related disciplines, including (though not limited to) psychology, linguistics, philosophy, neuroscience, and anthropology.

This task is not as simple as it sounds, since an AI theory cannot simply copy the concepts and statements from the related disciplines — to form the HI part of the theory, *selection* is needed; to form the GI part of the theory, *generalization* is needed.

"Intelligence" is not related to every aspect of a human being, and AI is not an attempt to clone a human. Even though the concept of intelligence has many different understandings, it is mainly about the *mental* properties of human beings, rather than their *physical* or *biological* properties (though those properties have impacts in the *content* of human thought). Furthermore, even only for lexical considerations, the notion of "Intelligence" should be more *general* than the notion of "Human Intelligence", so as to cover the non-human forms of intelligence. Therefore, an AI theory needs to decide the boundary of its empirical evidence, by indicating which processes and mechanisms in the human mind/brain/body is directly relevant to AI, and which of them are not. In other words, an AI theory must specify the scope and extent to which a computer is (or will be) similar to a human.

The following two extreme positions on this issue are obviously improper — if HI is specified in such a "tight" way that is bounded to all aspects of a human being, non-human intelligence would be impossible by definition; if HI is specified in such a "loose" way that the current computer systems are already intelligent by definition, AI will be trivialized and deserves no attention.

This task uniquely belongs to AI theories, because even though there are many studies on the human mind in the past in the related disciplines, little effort is made to separate the conclusions about "intelligence" (or "cognition", "mind") in general from those about "*human* intelligence" (or "*human* cognition", "*human* mind") in specific.

For example, though there is a huge literature on the psychological study of human intelligence, which is obviously related to AI, an AI theory cannot use the conclusions indiscriminately. This is because the notion of "intelligence" is used in psychology as an attribute where the difference *among human beings* is studied, while in AI it is an attribute where the difference *between humans and computers* is much more important. Many common properties among human beings are taken for granted in psychology, so they are rarely addressed in psychological theories. On the contrary, these properties are exactly what AI tries to reproduce, so they cannot be omitted in AI theories. For this reason, it is not helpful to directly use human IQ tests to evaluate the intelligence of a computer system. Similarly, the correctness of an AI theory cannot be judged in the same way as a theory in a related empirical discipline, such as psychology.

On the other hand, the human–computer difference cannot be used as an excuse for an AI theory to contain conclusions that are clearly inconsistent with the existing knowledge of the human mind. In the current context, even a theorem proved in a formal theory is not necessarily "correct" as a conclusion about intelligence, unless the axioms of the theory can be justified as acceptable in AI. If a normal human being is not "intelligent" according to an AI theory, then the theory is not really about intelligence as we know it, but about something else. This is especially the case for the GI part of the theory — even though generalization and simplification are necessary and inevitable, overgeneralization and oversimplification can cause serious distortion in a theory, to the extent that it is no

longer relevant to the original problem.

For an AI theory to be correct, it does not need to explain every phenomenon of the human mind, but only those that are considered as essential for HI by the theory. Though each theory may select different phenomena, there are some obvious features that should be satisfied by every theory of intelligence. Suggested features are exemplified by the requirement of being *general* [Bringsjord and Sundar G (2009)], or being *adaptive* and can work with *insufficient knowledge and resources* [Wang (2010)].

At the current time, the correctness of an AI theory is usually a matter of degree. The existence of certain counterevidence rarely falsifies a theory completely (as suggested by Popper (1959)), though it does decrease its correctness, and therefore its competitiveness when compared with other theories. We will return to this topic later.

Concreteness

The practical value of a scientific theory shows in the guidance it provides to human activities. In the current context, this requirement focuses on the relation between an AI theory (especially the CI part) and the computer systems developed according to it.

Since the objective of AI is to build "thinking machines", the content of an AI theory need to be concrete enough to be applicable into system design and development, even though it does not have to specify all the technical details.

This requirement means that a pure descriptive theory about how human intelligence works will not qualify as a good AI theory. In the theoretical discussions of AI and Cognitive Science, there are some theories that sound quite correct. However, they are very general, and use fuzzy and ambiguous concepts, so seem to be able to explain everything. What is missing in these theories, however, is the ability of making *concrete*, *accurate*, and *constructive* suggestions on how to build AI systems.

Similarly, it is a serious problem if a theory of AI proposes a design of AI systems, but some key steps in it cannot be implemented — for example, the AIXI model is uncomputable [Hutter (2005)]. Such a result cannot be treated as an unfortunate reality about intelligence, because the involved notion of "intelligence" is a construct in the theory, rather than a naturally existing phenomenon objectively described by the theory. The human mind has provided an existing proof for the possibility of intelligence, so there is no reason to generalize it into a notion that cannot be realized in a physical system.

In summary, a good AI theory should include a description of intelligence using the terminology provided by the existing computer science and technology. That is, the theory not only needs to tell people *what should be done*, but also *how to do it*.

"To guide the building of AI systems" does not necessarily mean these systems come with practical problem-solving capability. It again depends on the working definition of intelligence. According some opinion, "intelligence" means to be able to solve human-solvable problems [Nilsson (2005)], so an AI theory should cover the solutions to various practical problems. However, there are also theories that do not take "intelligence" as problem-solving capability, but learning capability [Wang (2006)]. According to such a theory, when an AI system is just built, it may have little problem-solving ability, like a

human baby. What it has is the *potential* to acquire problem-solving ability via its interaction with the environment. The requirement of concreteness allows both of the previous understandings of intelligence — no matter how this concept is interpreted, it needs to be realized in computer systems.

To insist that the CI part of an AI theory must be presented using concrete (even computational) concepts does not mean that the theory of AI can be replaced by the existing theories of computer science. Not all computer systems are intelligent, and AI is a special type of computer systems that is designed according to a special theory. It is just like that a theory of architecture cannot be replaced by a theory of physics, even though every building is constructed from physical components with physical relations among them. The claim that AI needs no theory beyond computer science [Hayes and Ford (1995)] cannot explain the obvious difference between the human mind and the conventional computers.

Compactness

While the previous two properties (correctness and concreteness) are about the *external* relation between an AI theory and outside systems (human and computer, respectively), compactness is a property of the *internal* structure of the theory. Here the word "compactness" is used to mean the conceptual simplicity of a theory's content, not merely on its "size" measured literally.

Since scientific theories are used to guide human behaviors, simple theories are preferred, because they are easier to use and to maintain (to verify, to revise, to extend, etc.). This opinion is well known in various forms, such as "Occam's Razor" or "Mach's Economy of Thought", and is accepted as a cornerstone in several AGI theories [Baum (2004); Hutter (2005); Schmidhuber (2007)].

To establish a compact theory in a complicated domain, two common techniques are *axiomatization* and *formalization*.

Axiomatization works by compressing the core of the theory into a small number of fundamental concepts and statements, to be taken as the basic notions and axioms of the theory. The other notions are defined recursively from the basic ones, and the other statements are proved from the axioms as theorems. Consequently, in principle the theory can be reduced to its axioms. Besides efficiency in usage, axiomatization also simplifies the verification of the theory's consistency and applicability.

Formalization works by representing the notions in a theory by symbols in an artificially formed language, rather than by words in a naturally formed language. Consequently, the notions have relatively clear and unambiguous meaning. The same theory can also be applied to different situations, by giving its symbols different interpretations. Even though it looks unintuitive to outsiders, a formal theory is actually easier to use for various purposes.

Axiomatization and formalization are typically used in mathematics, as well as in logic, computer science, and other normative theories. The same idea can also be applied to empirical science to various degrees, though because the very nature of those theories, they cannot be *fully* axiomatized (because they must open to new evidence) or *fully* formalized

(because their key concepts already have concrete meaning associated, and cannot be taken as symbols waiting to be interpreted).

Since a theory of AI has empirical content, it cannot be fully axiomatized or formalized, neither. Even so, it is still highly desired for it to move in that direction as far as possible, by condensing its empirical content into a small set of assumptions and postulations, then deriving the other part of the theory from it using justifiable inference rules. To a large extent, it is what a "Serious Computational Science" demands, with the requirements of being "cohesive" and "theorem-guided" [Bringsjord and Sundar G (2009)].

1.4. Relations among the properties

To summarize the previous discussions, a good AI theory should provide a *correct* description about how intelligence works (using evidence from human intelligence), give *concrete* instructions on how to produce intelligence in computer systems (using feasible techniques), and have a *compact* internal structure (using partial axiomatization and formalization).

These three requirements are *independent*, in the sense that in general there is no (positive or negative) correlation among them. All the three C's are desired in a theory, for different reasons, and one cannot be reduced into, or replaced by, the others.

For example, a simpler theory is not necessarily more correct or less correct, when compared with other theories. On this topic, one usual misconception is about Occam's Razor, which is often phrased as "Simpler theories are preferred, because they are more likely to be correct". This is not proper, since the original form of this idea was just something like "Simpler theories are preferred", and it is not hard to find examples where simpler theories are actually less correct. A common source of this misconception is the assumption that the only desired feature of a scientific theory is its correctness (or "truth") — in that case, if simpler theories are preferred, the preference must come from its correctness. However, generally speaking, compactness (or simplicity) is a feature that is preferred *for its own sake*, rather than as an indicator of correctness. It is like when we compare several products, we prefer cheaper ones when everything else is about the same, though it does not mean that we prefer cheaper products because they usually have higher quality. Here "price" and "quality" are two separate factors influencing our overall preference, and additional information is needed to specify their relationship.[d]

In certain special situations, it is possible for the requirements to be taken as correlated. One such treatment is Solomonoff's "universal prior", which assumes that without domain knowledge, the simpler hypotheses have higher probability to be correct [Solomonoff (1964)]. Though Solomonoff's model of induction has its theoretical and practical values, the soundness of its application to a specific domain depends on whether the assumptions

[d]Some people may argue that a simpler theory is more correct because it is less likely to be wrong, but if a theory becomes simpler by saying less, such a simplification will make the theory covers less territory, so it will also have less supporting evidence. To simply remove some conclusions from a theory does not make it more correct, unless "correctness" is defined according to Popper's falsification theory about science [Popper (1959)], so the existence of supporting evidence does not contribute to the correctness of a theory. Such a definition is not accepted here.

of the model, including the above one, can be satisfied (exactly or approximately) in the domain. For the related AGI models (such as AIXI [Hutter (2005)]), such justifications should be provided, rather than taken for granted. After all, we often meet simple explanations of complex phenomena that turn out to be wrong, and Occam's Razor cannot be used as an argument for the *correctness* of a theory (though it can be an argument for why a theory is preferred). For the same reason, a formal theory is not necessarily more correct than an informal one, though the former is indeed preferred when the other features of the two theories are similar.

These three C's are arguably *complete*, because altogether they fully cover the subject matter: the descriptive ingredients of the theory need to be correct, the normative ingredients need to be concrete, and the whole theory needs to be compact. Of course, each of the three can be further specified with more details, while all of them must be possessed by a theory that is about intelligence, rather than only about one part or one aspect of it.

All three C's can only be *relatively* satisfied. As a result, though probably there will not be a *perfect* theory of AI, there are surely *better* theories and *not-so-good* ones. When a theory is superior to another one in all three dimensions, it is "generally better". If it is superior in one aspect, but inferior in another, then whether it is better for the current purpose depends on how big the differences are, as well as on the focus of the comparison. Intuitively speaking, we can think the overall "score" on the competitiveness of an AI theory as a multiplication of the three "scores" it obtains on the three C's, though we do not have numerical measurements for the scores yet. In this way, an acceptable theory must be acceptable in all the three dimensions. Even if a theory is excellent in two aspects, it still can be useless for AI if it is terrible in the third.

1.5. Issues on the properties

In the current theoretical explorations in AGI, a common problem is to focus on some desired property of a theory, while ignoring the others.

Issues on *correctness* typically happen in formal or computational models of intelligence. Sometimes people think as long as they make it clear that a model is based on "idealized assumptions", they can assume whatever they want (usually the assumptions required by their available theoretical or technical tools). For example, Schmidhuber thought that for AI systems, the assumption of Markovian environments is too strong, so "We will concentrate on a much weaker and therefore much more general assumption, namely, that the environment's responses are sampled from a computable probability distribution. If even this weak assumption were not true then we could not even formally specify the environment, leave alone writing reasonable scientific papers about it." [Schmidhuber (2007)] It is true that every formal and computational model is based on some idealized assumptions, which are usually never fully satisfied in realistic situations. However, this should not be taken as an excuse to base the model on highly unrealistic assumptions or assumptions that can only be satisfied in special situations. Since the conclusions of the model are largely determined by its assumptions, an improper assumption may completely change the nature

of the problem, and consequently the model will not be about "intelligence" (as we know it) at all, but about something else. One cannot force people to accept a new definition of "intelligence" simply because there is a formal or computational model for it — it reminds us the famous remark of Abraham Maslow: "If you only have a hammer, you tend to see every problem as a nail". We do want AI to become a serious science, but to change the problem into a more "manageable" one is not the way to get there.

On the other hand, to overemphasize correctness at the cost of the other requirements also leads to problems. The "Model Fit Imperative" analyzed by Cassimatis (Chapter 2 of this book) is a typical example. A theory of AI is not responsible for explaining or reproducing all the details of human intelligence. The most biologically (or psychologically) accurate model of the human brain (or mind) is not necessarily the best model for AI.

Issues on *concreteness* typically happen in theories that have rich philosophical content. Though philosophical discussions are inevitable in AI theories, to *only* present a theory at that level of description is often useless, and such a discussion quickly degenerates into word games, which is why among AI researchers "this is a philosophical problem" is often a way to say "It doesn't matter" or "You can say whatever you want about it". Similarly, if some theory contains descriptions that nobody knows how to implement or even to approximate, such a theory will not be very useful for AI. Just to solve the AI problem "in principle" is not enough, unless those principles clearly lead to technical decisions in design and development, even if not in all details.

Issues on *compactness* widely exist in AGI projects that are mainly guided by psychological/biological inspirations or problem-solving capabilities. Such a project is usually based on a theory that basically treats intelligence as a collection of "cognitive functions" that are organized into a "cognitive architecture" (see Chapter 7 and 8 of this book).

One problem about this approach is that the functions recognized in the human mind are not necessarily carried out by separate processes or mechanisms. In a psychological theory, sometimes it is reasonable to concentrate on one aspect of intelligence, but such a practice is not always acceptable in an engineering plan to realize intelligence, since to reproduce a single mechanism of intelligence may be impossible, given its dependency on the other mechanisms. For example, "reasoning" and "learning" may be two aspects of the same process [Michalski (1993); Wang (2006)]; "perceiving" may be better considered as a way of "acting" [Noë (2004)]; "analogy" may be inseparable from "high-level perception" [Chalmers *et al.* (1992)].

Though from an engineering point of view, a modular architecture may be used in an AI system, the identification and specification of the modules must follow an AI theory — the functions and modules should be the "theorems" of a theory that are derived from a small number of "axioms" or "principles", so as to guarantee the coherence and integrity of the system as a whole. Without such an internal structure, a theory of AI looks like a grab bag of ideas — even when each idea in it looks correct and concrete, there is still no guarantee that the ideas are indeed consistent, nor guidance on how to decide if on a design issue different ideas point to different directions. Such an architecture often looks arbitrary, without convincing reason for the partition of the overall function into the modules. Consequently,

the engineering practice will probably be full of trial-and-error, which should not happen if the theory is well-organized.

1.6. Conclusion

A major obstacle of progress in AI research is "theoretical nihilism" — facing the well-known difficulty in establishing a theory of AI, the research community as a whole has not made enough effort in this task, but instead either follows some other theories developed for certain related, though very different, problems, or carries out the research based on intuitions or practical considerations, with the hope that the theoretical problems can be eventually solved or avoided using technical tricks.

Though AI is indeed a very hard problem, and it is unlikely to get a perfect (or even satisfactory) theory very soon, to give up on the effort or to depend on an improper substitute is not a good alternative. Even though the research can go ahead without the guidance of a theory, it may run in a wrong direction, or into dead alleys. Even an imperfect theory is still better than no theory at all, and a theory developed in another domain does not necessarily keep its authority in AI, no matter how successful it is in its original domain.

Given the special situation in the field, an AI theory must be descriptive with respect to the human mind, and be normative with respect to computer systems. To achieve this objective, it should construct a notion of general intelligence, which does not depend in the details of either the biological brain or the electrical computer. The desired properties of such a theory can be summarized by the Three C's: *Correctness*, *Concreteness*, and *Compactness*, and the overall quality of the theory depends on all the three aspects. Among the existing theoretical works, many issues are caused by focusing only on one (or two) of the properties, while largely ignoring the other(s).

To a large extent, the above issues come from the science–engineering duality of AI. A theory of AI is similar to a theory of natural science in certain aspects, while that of engineering in other aspects. We cannot work in this field like typical natural scientists, because "intelligent computers" are not existing phenomena for us to study, but something to be created; on the other hand, we cannot work like typical engineers, because we are not sure what we want to build, but have to find that out by studying the human mind. The theoretical challenge is to find a minimum description of the human mind at a certain level, then, with it as specification, to build computer systems in which people will find most of the features of "intelligence", in the everyday sense of the word.

Though the task is hard, there is no convincing argument for its impossibility. What the field needs is to spend more energy in theoretical exploration, while keeping a clear idea about what kind of theory we are looking for, which is what this chapter attempts to clarify.

Acknowledgements

Thanks to Joscha Bach for the helpful comments.

References

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

Alvarado, N., Adams, S. S., Burbeck, S. and Latta, C. (2002). Beyond the Turing Test: Performance metrics for evaluating a computer simulation of the human mind, in *Proceedings of the 2nd International Conference on Development and Learning*, pp. 147–152.

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

Baum, E. B. (2004). *What is Thought?* (MIT Press, Cambridge, Massachusetts).

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

Bringsjord, S. (2008). The logicist manifesto: At long last let logic-based artificial intelligence become a field unto itself, *Journal of Applied Logic* **6**, 4, pp. 502–525.

Bringsjord, S. and Sundar G, N. (2009). Toward a serious computational science of intelligence, Call for Papers for an AGI 2010 Workshop.

Bundy, A. and Ohlsson, S. (1990). The nature of AI principles, in *The foundation of artificial intelligence—a sourcebook* (Cambridge University Press, New York), pp. 135–154.

Cassimatis, N. L. (2006). Artificial intelligence and cognitive science have the same problem, in *Papers from the AAAI Spring Symposium on Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, pp. 27–32.

Chalmers, D. J., French, R. M. and Hofstadter, D. R. (1992). High-level perception, representation, and analogy: a critique of artificial intelligence methodology, *Journal of Experimental & Theoretical Artificial Intelligence* **4**, pp. 185–211.

Chandrasekaran, B. (1990). What kind of information processing is intelligence? in *The foundation of artificial intelligence—a sourcebook* (Cambridge University Press, New York), pp. 14–46.

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

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

Gabbay, D. M. and Woods, J. (2003). Normative models of rational agency: The theoretical disutility of certain approaches, *Logic Journal of the IGPL* **11**, 6, pp. 597–613.

Goertzel, B. (2009). Toward a general theory of general intelligence, *Dynamical Psychology*, URL: *http://goertzel.org/dynapsyc/dynacon.html#2009*.

Goertzel, B., Arel, I. and Scheutz, M. (2009). Toward a roadmap for human-level artificial general intelligence: Embedding HLAI systems in broad, approachable, physical or virtual contexts, *Artificial General Intelligence Roadmap Initiative*, URL: *http://www.agi-roadmap.org/images/HLAIR.pdf*.

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

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

Hayes, P. and Ford, K. (1995). Turing Test considered harmful, in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 972–977.

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

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

Laird, J. E., Wray, R. E., Marinier, R. P. and Langley, P. (2009). Claims and challenges in evaluating human-level intelligent systems, in *Proceedings of the Second Conference on Artificial General Intelligence*, pp. 91–96.

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

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Process-*

ing of Visual Information (W. H. Freeman & Co., San Francisco).
McCarthy, J. (1984). We need better standards for AI research, *AI Magazine* **5**, 3, pp. 7–8.
McCarthy, J. (1988). Mathematical logic in artificial intelligence, *Dædalus* **117**, 1, pp. 297–311.
McCarthy, J. (2007). From here to human-level AI, *Artificial Intelligence* **171**, pp. 1174–1182.
McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, URL: *http://www-formal.stanford.edu/jmc/history/dartmouth.html*.
Michalski, R. S. (1993). Inference theory of learning as a conceptual basis for multistrategy learning, *Machine Learning* **11**, pp. 111–151.
Minsky, M. (1985). *The Society of Mind* (Simon and Schuster, New York).
Newell, A. (1990). *Unified Theories of Cognition* (Harvard University Press, Cambridge, Massachusetts).
Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search, *Communications of the ACM* **19**, 3, pp. 113–126.
Nilsson, N. J. (2005). Human-level artificial intelligence? Be serious! *AI Magazine* **26**, 4, pp. 68–75.
Noë, A. (2004). *Action in Perception* (MIT Press, Cambridge, Massachusetts).
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann Publishers, San Mateo, California).
Pollock, J. L. (2006). *Thinking about Acting: Logical Foundations for Rational Decision Making* (Oxford University Press, USA, New York).
Popper, K. R. (1959). *The Logic of Scientific Discovery* (Basic Books, New York).
Rumelhart, D. E. and McClelland, J. L. (eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Foundations* (MIT Press, Cambridge, Massachusetts).
Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, 3rd edn. (Prentice Hall, Upper Saddle River, New Jersey).
Schank, R. C. (1991). Where is the AI? *AI Magazine* **12**, 4, pp. 38–49.
Schmidhuber, J. (2007). The new AI: General & sound & relevant for physics, in B. Goertzel and C. Pennachin (eds.), *Artificial General Intelligence* (Springer, Berlin), pp. 175–198.
Simon, T. W. (1990). Artificial methodology meets philosophy, in *The foundation of artificial intelligence—a sourcebook* (Cambridge University Press, New York), pp. 155–164.
Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I and II, *Information and Control* **7**, 1-2, pp. 1–22,224–254.
von Neumann, J. (1958). *The Computer and the Brain* (Yale University Press, New Haven, CT).
Wang, P. (1995). *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*, Ph.D. thesis, Indiana University.
Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence* (Springer, Dordrecht).
Wang, P. (2008). What do you mean by 'AI', in *Proceedings of the First Conference on Artificial General Intelligence*, pp. 362–373.
Wang, P. (2010). A General Theory of Intelligence, An on-line book under development. URL: *http://sites.google.com/site/narswang/EBook*.
Wang, P. and Goertzel, B. (2007). Introduction: Aspects of artificial general intelligence, in B. Goertzel and P. Wang (eds.), *Advance of Artificial General Intelligence* (IOS Press, Amsterdam), pp. 1–16.
Wilks, Y. (1990). One small head: models and theories, in *The foundation of artificial intelligence—a sourcebook* (Cambridge University Press, New York), pp. 121–134.