Note on the Terminology

Artificial beings (in contrast to human beings) are the subject of this book. As up to now most of their activities have been performed by humans, an extensive vocabulary can be used to describe actions when the performer is human, but few words are available for when it is an artificial being. Therefore, I will speak of artificial beings which think, choose, prefer, are conscious, want, etc. However, when I say that an artificial being "thinks" that only means that it processes some data so that it generates new data, which will help it to determine what it will do. For instance, it gives up its present task because it is too difficult, it creates a method for solving problems, it chooses the next rule that it will execute, it tries to understand the reason for a previous mistake, etc.

There is no anthropomorphism intended in this use of the vocabulary. It is simply due to the fact that it is impossible to create a new word for each activity of an artificial being when it is in the same situation as a human being, and when its results are often as good as ours. This does not mean that I believe that an artificial being operates as a human does. On the contrary, I will try to show that they can process information in a completely different way, particularly when they observe their behavior, which gives them a large advantage over us.

Let us consider an example: when I speak of the artificial being which I built, I say that it is autonomous. When we say that a human being is autonomous, that means that he can make his decisions without asking for help. In this meaning, my creature is certainly autonomous, it works without any interference for more than one month. However, this characteristic is not

enough; a traffic light can work alone for years, it does not need an exterior agent to go from green to orange or from orange to red; an autonomous agent must not be completely predictable. However, this point raises a problem, because artificial beings are usually completely determined, they are following a program and, theoretically, we can predict all of their future decisions when we know their program and their data. Naturally, this comparison is unfair because we do not know how our brain works, even if it is as determined as a computer program. We cannot reject the possibility of an autonomous artificial being only because we do not know how human beings work. Moreover, theoretically predictable does not mean practically predictable. It is impossible to predict the behavior of a complex artificial being, it depends on its history and on a huge number of small operations, we will never have enough time to execute them. At different instants, it can take different decisions although it is in exactly the same situation because they also depend on the events that happened between these instants. The behavior of an artificial being may also be more difficult to predict than the behavior of a human one. For these reasons, I will speak of autonomous artificial beings, although it does not choose its act in the same way as us.

Finally, in Artificial Intelligence (AI) it is important to indicate unambiguously whether we are considering a human or an artificial being, particularly when we use pronouns. Indeed, both kinds of beings have similar activities and the context does not always indicate the nature of the agent. Thus, I will always use “it” to refer to an artificial being, which will often be my system CAIA. When I consider an activity in common with human and artificial beings, the agent is often represented by “one” as in “Let us assume that one has decided to interpret a program”; all we know of this being is that is able to make a decision, and it is not specified whether this agent is human or artificial. The passive form is also convenient to mention a being without specifying its nature of an actor as in “when the program has been written”. In some situations, to insist on the fact that I am simultaneously speaking of human and artificial beings, I will write his/its, he/it, or who/which. I will never use his/her/its, the situation is complicated enough with two kinds of actors: when I introduce a human being, he may be male or female in all this book, even when I use “man” or “he”. “We” will always represent human beings, the reader–author pair as in “We will later see that ...” or all human beings as in “We do not like the idea that artificial beings could become more intelligent than human beings”.

Chapter 1

Presenting the Actors

At the beginning of a book, it is natural to present its actors. Of course, the first actor is the book itself. Then I will examine the qualities and drawbacks of two important families of actors: human beings and artificial beings. I do not forget the computer, an essential support for an artificial being. Next, I will present my reasons for developing my present research, and introduce the subject of this research: my colleague CAIA, which is an artificial scientist. Studying its behavior and its mechanisms will supply most of the examples of this book. Finally, I look at the domains where CAIA is carrying out research.

1.1. The book

The goal of this book is to examine whether an artificial being can have some capacities similar to those that consciousness or conscience give to human beings. Over many centuries, a huge amount of work has been done on human consciousness and conscience, but sometimes I feel that I still do not understand them. In relation with my work, to understand a hypothesis on the working of our brain, I need to see how the hypothesis can be implemented in a computer program. Naturally, scientists who lived before the computer era could not express themselves in such a precise way because they did not know of computers and how we can program them. This is a difficult and unnatural way of thinking, and even now, we have to realize several Artificial Intelligence systems before we can easily feed computer

programs with attractive ideas. When I read these early books, I can see a dozen ways to implement each of their ideas and so I feel confused: among all these possibilities, which one was seen by the author? For instance, I have never understood how the qualia could be exactly represented. The qualia are at the core of many theories of consciousness, but philosophers agree neither on their meaning, nor on their properties. We can roughly say that their goal is to represent in our brain sensations like redness, but when the authors want to be more precise, each one has his own definition. How could we use this concept in a computer program?

Over several years, some outstanding books and papers, such as those of Marvin Minsky, Daniel Dennett, Gérard Sabah or François Anceau, have been written by scientists who know exactly how a computer works, and the reader can understand how their ideas might be implemented. However, these ideas have not always been effectively implemented. An AI scientist knows that it is impossible to foresee in a program all the elements that would be necessary to give excellent results: either some mechanisms are forgotten, or their description is not sufficiently accurate. We have to experiment with the system, and modify it to work better. Thus, a description without an implementation is an indispensable starting point, but it is not sufficient, although it contains many interesting and promising ideas. Moreover, the goal of most of the research has been to improve our understanding of consciousness and of conscience for human beings. Usually, they were not concerned to see if these faculties could be expressed by artificial beings, which have different ways to process information. However, there are some exceptions; in particular, John McCarthy has written a theoretical paper on the possibility of making robots conscious of their mental states. My book complements his work and I do not consider the theoretical point of view, but rather its realization by a practical computer system. His paper must therefore be read by all those who are interested in this approach. Several papers have been published in the last years on the realization of various aspects of consciousness. They contribute new and interesting ideas, but this is not enough: intelligence is a whole, and we cannot examine only how some of its aspects can be implemented, we have also to study what kind of consciousness has been given to artificial beings which have to perform difficult tasks.

Examining existing artificial beings

Another approach is to examine an existing artificial being, and consider whether some of its possibilities could be related to consciousness and conscience as they exist for human beings. Since, in a program, everything has to be defined, we can find which mechanisms generate a behavior similar to our own behavior. Besides, when an artificial being uses these mechanisms, it may have possibilities out of reach for us, whose performance depends on the structure of our brain and the characteristics of its basic element, the neuron. Then, we are no longer interested in the understanding of human cognition, but in the study of an artificial cognition. Its main goal is the realization and the understanding of artificial beings, for them to become as efficient as possible. In that situation, we have a huge advantage over the psychologists: we can examine all the programs that make up an artificial being, thus we can exactly know the reasons for its actions and for its limits. However, there is a practical problem because we may have to examine programs with several hundreds of thousands instructions: we may misinterpret these programs if we are not their author. Papers in journals may give a sketchy description, but thousands of pages would be necessary for an accurate description. Thus, a paper, and even a book, can only give a general idea of the methods used by an artificial being; the details cannot be included, although they are often essential to understanding its performance.

So, the only person who can accurately describe the properties of a system is its author. Several artificial beings have already used interesting mechanisms which give them some capacities related to consciousness or to conscience, one of the most impressive is Lenat's EURISKO. Unfortunately, I cannot describe them as much as I would because I cannot examine their programs in detail, I cannot know how some points have been dealt with, I cannot make experiments in order to evaluate their possibilities accurately. For this reason, I will mainly take my examples from the CAIA (Chercheur Artificiel en Intelligence Artificielle: an Artificial Artificial Intelligence Scientist) system, which I have experimented with for more than 20 years. CAIA has several attributes in common with human scientists, who discover new methods for solving problems, and perform many experiments to do so. However, CAIA does not have to deal with other important roles of current human scientists: CAIA does not publish scientific papers yet, CAIA does not search for scientific grants or funds, CAIA does not interact with the

scientific community or policy makers, CAIA does not manage or advise junior human researchers.

CAIA is a step toward the realization of an artificial AI scientist. For the present, its main research domain is solving problems defined by a set of constraints, which must be satisfied by the solution. This family of problems includes many applications, and we often have to solve such problems. For instance, when we are choosing a meal for guests, we have to take into account a lot of constraints: their likes, the contents of the fridge, our cooking tools, our budget, the food supply at the nearest supermarket, etc. Crosswords and Sudoku problems are also in this category. My goal was not to develop an artificial consciousness or conscience; it was to realize a system able to learn to solve problems without the need to imitate the human behavior. Once CAIA was successful, I looked for and analyzed the mechanisms related to consciousness and conscience. When they were present, they were necessary for the success of the artificial scientist.

Plan of the book

This book comprises ten chapters. In Chapter 2, we describe some possible meanings for consciousness and conscience. Then in Chapter 3 we show that the concept of an individual is different for a human being and an artificial being. Chapters 4 to 6 examine what the consciousness of an artificial being could be: we describe several ways it can observe itself, why it is useful to observe its own behavior and how that can be implemented. Chapters 7 and 8 show that an artificial being can and must have a conscience. In Chapter 9, we examine some problems related to the conscience such as the importance of emotions, the difficulty in modifying its own conscience, and also the consequences of the existence of artificial beings on the human conscience. Chapter 10 discusses what the future has in store for CAIA. Appendix 1 describes the methods used for solving the problems given to CAIA while Appendix 2 gives more details of the implementation of CAIA's consciousness.

1.2. Human and artificial beings

Two of the most amazing features of human behavior are consciousness and conscience, which philosophers and psychologists have tried to explain

for many centuries. How is it possible to observe a part of the processes that occur in our brain? How can we acquire and use the knowledge of what is right and wrong? These problems are difficult because we do not have access to the processes which are executed in our brain: if we could observe our mechanisms, it would be easy to answer these questions. Besides, no one can observe what happens in the brain of another person. Finally, although some experiments might help to understand what happens in the brain of a person if some of its parts were destroyed, they are naturally forbidden for ethical reasons. Therefore, studies on the working of the brain are very limited, and this helps explain why there are so many theories, which are often contradictory. However, everybody agrees that these processes are essential for satisfactory behavior. Socrates' motto was "Know Thyself", which can be realized because we are conscious. Furthermore, societies can survive only because the conscience of their members incites them to perform useful acts, and forbids them from committing actions dangerous for the future of the society.

Realizing artificial beings

We, AI scientists, realize computer systems, which are artificial beings. For the same reasons that consciousness and conscience are useful for human beings, we are led to give them the possibility to behave as if they were conscious and as if they had a conscience. The goal of some scientists is to model how the human brain works so that they will have a better understanding of human behavior. Indeed, if a mechanism generates a behavior similar to ours, it is possible that we are also using it. On the other hand, if the two behaviors are completely different, our brain certainly does not use the same methods as those implemented in the model. This approach is very interesting, but the goal of other scientists, and I am one of them, is not to understand how we work, but to realize systems that are as efficient as possible: we no longer claim to model the human brain. Naturally, a useful heuristic is to start from a hypothesis on how our brain works, but we are not obliged to follow it if we find better ways for performing the tasks given to our system. Naturally, once a system has been completed, it is interesting to analyze its methods, and perform some experiments on human beings to find whether or not we are using these same methods. It is also important to analyze its results for building an artificial cognition, which includes the description of new ways for implementing consciousness and conscience.

Constraints coming from the use of neurons

We must not blindly imitate human behavior because computers are tools which work differently from our brain. An obvious advantage is their speed, the number of operations executed in one second may be approximately the same as the number of operations that a human will execute in its whole life; besides, the computer will not make the thousands of mistakes made by a human. This is true when we are working in a serial mode, that is we cannot make several operations simultaneously, as is the case in arithmetic where we cannot make two divisions at the same time. However, the brain can also work in parallel, where it processes much data at the same time, for instance when we are perceiving a picture; for such application, the advantage of computers is not so large. Yet, the slowness of the neurons is not the only drawback coming from their use: we can neither observe their state, nor create new neurons. Another drawback is that it is not easy to store specific information such that it could be correctly used immediately, we have to use it several times before it is operational. I know that cars drive on the left in England, but I must be careful when I am in Great Britain because this information is not available each time that I cross a street. After a few days, this becomes automatic; unfortunately, when I return to the continent, it will be almost as long to acquire new reflexes. For a computer program, it is sufficient to swap "right" and "left" in a program, and everything is adapted in the new environment immediately; it is also easy to return to the previous behavior by doing the inverse swap.

Constraints coming from the structure of our brain

However, some characteristics of human intelligence do not come from using neurons, but from the structure of our brain. For instance, our working memory can hold around seven elements, which is a serious restriction for many tasks. We are not aware of the consequences of the small size of this memory because every human has this handicap; we avoid putting other people in situations where a larger working memory would be necessary. For instance, we speak in such a way that our interlocutor can process our sentences. We could possibly generate sentences such as:

The mouse that the cat that the dog that the woman that the soldier loves pets hates devours.

However, our interlocutor cannot understand it, while he can easily understand the following equivalent sentence:

The soldier loves the woman who pets the dog which hates the cat which devours the mouse.

We cannot stack in our working memory enough items so that we can wait for the many verbs that are at the end of the first sentence. Our brain evolved so that our ancestors, hunters and gatherers, could survive, breed and raise their children. The natural selection has created a wonderful tool for all of these activities, where a large working memory is not essential. Therefore, evolution has not developed this kind of memory, and it is the same for many other activities such as proving mathematical theorems, or managing large organizations. Our intellectual powers were useful for our ancestors' activities, and it happens that some of these powers are also very useful for other activities. However, there is no reason why they would be optimal for these new applications: the structure of our brain has not been optimized for them. Consciousness and conscience are among those useful aptitudes which evolution has developed for our ancestors, but it is likely that it would be possible to improve them so that we would be more efficient for most of our new activities. In fact, with the restrictions due to the neuron and those due to the structure of our brain, we are handicapped in comparison with artificial beings which are faster and which can simulate a huge variety of mechanisms.

Therefore, when we want to realize an efficient artificial being, we do not restrict its performance by imitating our way of thinking, thus enforcing constraints that are not essential. On the contrary, we have to take advantage of their extraordinary possibilities so that we obtain the best results. The goal of this book is to show that several features of consciousness and of conscience may be improved in that way. However, some of the human capacities have not been given to CAIA because they did not seem useful for the tasks which it is doing at present. Naturally, for future developments, it will certainly be necessary to give CAIA some of these missing capacities.

1.3. The computer

The computer is the extraordinary tool that enables an AI scientist to implement and experiment with his ideas. The methods which will be given

to an artificial being are first defined as programs and data, then they are given to a computer which will execute these programs and give the desired results. We do not need exceptional computers for performing most AI research. I use an ordinary PC, such as those which are sold in their millions each year. It is not ultra-fast (1.8 GHz), and it only has 512 MB of memory.

However, a computer is almost useless on its own, we need a program to manage its working, which is the operating system. These systems did not always exist: the first computers were used without an operating system, but the present computers are so complex that it is almost impossible to use them without an operating system. It manages the computer resources, it decides which programs will be executed (a program being executed is called a process), it allocates them memory chunks, it deals with the inputs made through the mouse and the keyboard, the outputs such as those on the printer or the monitor, the network connections with other computers, and so on. It also checks that several parameters necessary for a satisfactory use of the computer are correct, for instance it will start a fan if it finds that the temperature inside the processing unit is too high. I am using Linux as an operating system.

The behavior of an artificial being is defined by a set of programs and data. To start it, one gives it to the operating system of a computer, which will create a process corresponding to an execution of this program. A process contains data and code expressed in a language understandable by the computer. In this book, I will distinguish the program, which is a sequence of instructions given to a computer with its data, from the corresponding process generated by the operating system controlling the execution of this program.

Thus, the operating system manages many processes, including the process corresponding to the execution of our AI program, and the several processes that our original process may have launched when needed. Even when we have given no task to our computer, 30 or so processes may be managed by the operating system, for its own needs or possibly for other users of the same computer. As many processes are simultaneously controlled by the operating system, it must avoid any interference between them: if our process were to write in an area of the memory used by a process belonging to another user, this last process would give the wrong results. So, the operating system allocates an area of the memory to each process; if an instruction of a process requires use of a memory which is not

in this area, the computer will automatically realize that there is a mistake; it will reject the execution of this instruction, and a message is sent to the guilty process. Everything happens as if our process was the only user of a computer with a memory restricted to the area allocated by the operating system.

1.4. The author

My thinking about consciousness and conscience comes from CAIA, which I have been developing for more than 20 years. I wanted to realize a performing general system and, at the start, my goal was not at all to study the consciousness and conscience of artificial beings. It was only when the system began to have satisfactory results that I wondered whether some of its behavior had some aspects similar to human consciousness and conscience. So, all the characteristics which are relevant to these domains were not introduced to imitate human behavior, but because they were necessary to improve efficiency.

I started developing CAIA because I was struck by the slowness of the development of AI. Although I believe that it is certainly possible that artificial beings could be more intelligent than human beings, I have come to doubt whether human intelligence alone could ever realize such a difficult task, without exterior help.

Bootstrapping AI

Thus, we need help, and who could be intelligent enough to help us? The only possible candidate is AI itself; we have to build AI systems which can help us to develop AI. This means that we must bootstrap AI. The idea of a bootstrap is paradoxical: how can a system be used to implement itself? Actually, there is a succession of artificial beings, and system number N helps us to implement system number N+1. If system N+1 is better than system N, then it will help us to realize system N+2 which will be better still than system N+1. The bootstrap is based on the collaboration of a human and a system in order to produce a better version of this system. In that way, we realize a sequence of artificial beings; as each one is better than its predecessor, either the steps are higher or the human help can be decreased. For CAIA, the value of N is around 1,500 at the present time. So, it is

completely different from its initial state in 1985, with much more than 1,000 intermediary versions, each one being used to generate the following one. I hope that, eventually, artificial beings will create better artificial beings without any human help; in that way, AI could progress by itself but, at present, we are very far from that state. Bootstrapping is often used for solving difficult problems, for instance we could not have realized the present computers if previous computers did not exist. The very first computers (in the 1940s and in the 1950s) were very simple, and could be made by human engineers without the help of a computer. As computers grew more and more powerful, they have been used to design still more powerful computers, and the design of current computer systems (from the microprocessors to the operating system and application software) would be unthinkable without a wide use of computers.

The key step for bootstrapping AI is the realization of an artificial AI scientist. Indeed, once such a scientist would be completed, it could develop AI without any human help. Naturally, it could also develop all of the other domains: mathematics, computer science, physics, management, etc., since AI is useful in all these domains. Besides, it will not only build methods for solving problems in these various domains, but it will also improve itself as an AI scientist; so, it will become a more and more competent AI scientist, and all the possible applications will benefit from these improvements. Thus, I started to develop the first steps of this venture. At the beginning, it is illusory to try to immediately create a scientist as able as a human one in every domain: it can only carry out a part of the necessary tasks, and I help it for the other ones.

In a bootstrap, a human has a crucial role, because a bootstrap is a sequence of steps from one version of the system to the following one. Someone has to decide when a new step will be made; to do that, he analyzes the results, and makes the modifications to the current version of the system in order to define its new improved version. I call this person the leader of the bootstrap, and I am the leader of CAIA's bootstrap.

I did not want to ape human methods, but I had to give my artificial scientist some possibilities related to those given to us by consciousness and conscience. Unlike what happens for human scientists, we can observe the inside of an artificial scientist, and see how these capacities have been implemented; due to this advantage, artificial cognition is easier to develop than human cognition. I take most of the examples from my experiments

with my artificial scientist, which has already lived several lives, each one for more than one month, without any outside intervention; moreover, a bug has never stopped it. It would be possible to have longer lives but, when I examine its behavior and the events that happened in one month, I have many ideas for improving it. It is therefore better to make these modifications, and to launch it in a new life where its initial capabilities have been improved. I will mention other works on artificial beings, but I often prefer to speak of CAIA because Know exactly how its modules work.

1.5. CAIA, an artificial AI scientist

When we say that a person is conscious, it is difficult to explain this clearly because the restrictions of our consciousness forbid us to know what happens when we are conscious. This is why papers on consciousness are sometimes not very precise, which seems almost acceptable because we already have an idea on the subject. However, the situation is completely different when we are considering artificial beings, we can know everything of their knowledge and of the steps executed for achieving a task. This is the case for CAIA, my artificial scientist in AI. We work together to realize better AI systems, each one of us performing the tasks at which he/it is the best. While CAIA improves, there are more and more tasks that it can accomplish at least as well as myself. The goal of the completely autonomous AI scientist is very far away but, even now, CAIA is very helpful for many activities.

CAIA, a useful collaborator

CAIA is permanently evolving, on the one hand by itself, and on the other hand through the modifications that I make. I try to give it some of my activities so that I can spend more time on those that it is not yet able to perform. At the beginning of a bootstrap, we simplify the tasks given to the system so that it can achieve them correctly, although it does not yet have all the capacities expected for the final system. One simple way to help the system is to limit the nature of the problems that it will have to solve. This is why I gave it at the beginning only constraint satisfaction problems; now I am adding the capacity to solve arithmetic problems, and other domains will follow. Another way to help it is to ask it to solve only a part of a problem. For instance, when it has to solve a constraint satisfaction problem, I do not

give it the formulation in a natural language such as English or French, it is instead given in a formalized language, similar to the mathematical language in many points. This removes several difficulties: for instance, it is no longer necessary to clear up ambiguities that are always present in a natural language text. In the distant future, when it will be able to solve problems expressed in natural language, another step of the bootstrap will have been made.

As I have already said, my artificial scientist has already had several lives, which lasted more than 3 million seconds each, that is more than one month, night and day. This is huge for a system that runs on a computer which executes more that a billion instructions each second. During each of its lives, it is completely autonomous, I do not step in, I never even examine what is happening. It creates several methods for solving problems, it uses them to solve problems, it creates new problems, it performs some experiments to see whether it could be possible to find better solutions with other methods, it analyzes these experiments (but it does not interpret these analyses, this is still my role), it tries to understand why one of its methods did not lead to a better solution for a problem, etc.. Of course, during this time several bugs occur, a complex program without bugs is a purely theoretical ideal. In such a situation, it finds which instructions are wrong (but it does not correct them, this is also still my role), and if the bug had put it into a loop, it will break this loop. I generally stop CAIA after about one month because I then have more than enough improvements to make from observation of its behavior during this period. It is more interesting to modify it, and observe what will happen in a new life.

Thus, I am progressively increasing the abilities of the artificial scientist, which already helps me because it now performs some of the tasks that I had to make at the start. Some of its activities are related to consciousness and to conscience. First, it observes what it has done so that it can understand the reasons for its successes and its failures, it examines its own knowledge so that it can foresee some possible uses, it observes what it is doing so that it immediately recognizes that it is going in a wrong direction. To do that, it has some of the possibilities given by the consciousness. However, as it is autonomous, it also has the capacity to judge what it is doing, which is what conscience enables us to do. It can see whether some results are good or poor so that it can try to repeat or to avoid them later. It can decide whether a new problem, which it has just created, is interesting so that it can decide whether to keep it. It evaluates the interest of the waiting tasks so that it can manage

its life as a scientist, beginning with those which are the most important and the easiest.

Should we imitate human behavior?

Although my goal is to realize an artificial scientist, I am not bound to imitate a human scientist. We have already seen that our cognition has some restrictions, which artificial beings do not have; it would be a pity not to use their capacities completely. Naturally, observing a human subject, mainly myself for CAIA, is very useful for finding initial ideas, but AI scientists do not reject methods that we could not execute, for instance because we would need too much time to do them. My initial goal was not at all to create an artificial being which would be conscious, and which would have a conscience. It only happened that it has some of these possibilities because they were necessary for its performance: it needs to be conscious to learn and to adapt itself to new situations, and it needs a conscience to be autonomous.

Most examples will come from these experiments, and I am often slightly envious of the possibilities of the artificial beings, which we will never have. We will see some of these in the study of the consciousness and of the conscience, but they also arise in other situations. For instance, an artificial being can not only replace a module insufficiently successful by a more efficient one, but it can add a new module specially tailored to a new application. Unfortunately, we cannot do that; for instance, our mathematical abilities are not as high as we could wish, evolution did not build us for that: this is only a by-product in the creation of modules enabling us to hunt or to seduce the opposite sex; we cannot add several billions of neurons specially organized to perform the tasks useful for a mathematician. On the contrary, an artificial being can possibly add to itself a large number of modules specialized for a new domain.

The main drawback of this approach is that we have no model that we could imitate, we could make many more things with a computer if we only knew that they would be useful. This permanently happens when I am developing CAIA, it can have activities that no human scientist could have but, as we have no model using them, it is difficult to define them. Moreover, we cannot find how human scientists work, because their most interesting processes are unconscious; the illumination of the discovery is