A Framework for the Foundation of the Philosophy of Artificial Intelligence*

VIOLA SCHIAFFONATI
Artificial Intelligence and Robotics Laboratory, Dipartimento di Elettronica e Informazione,
Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy

Abstract. The peculiarity of the relationship between philosophy and Artificial Intelligence (AI) has been evidenced since the advent of AI. This paper aims to put the basis of an extended and well founded philosophy of AI: it delineates a multi-layered general framework to which different contributions in the field may be traced back. The core point is to underline how in the same scenario both the role of philosophy on AI and role of AI on philosophy must be considered. Moreover, this framework is revised and extended in the light of the consideration of a type of multiagent system devoted to afford the issue of scientific discovery both from a conceptual and from a practical point of view.

Key words: agency, artificial intelligence, circular evolution, multi-agent system, philosophy of artificial intelligence, scientific agency

1. Introduction

The closeness and complexity of the relationship between philosophy and AI, and the consequent demand for approaching and systematizing it, has been constantly observed starting from the early days of AI. Claims as "philosophy has been [. . .] a close and dependable ally of researchers working on the foundations of AI" (Akman, 2000) or "much of AI already builds on works by philosophers" (Sloman, 1995) can be easily encountered both in the philosophical literature and in the one of AI. The core point of these approaches is represented by the central role philosophy covers toward AI: philosophy exerts an influence on AI both from an historical and a methodological point of view.

Besides the acknowledgement of the fundamental role of philosophy for AI, the role of AI for philosophy, as well, covers an important position in this debate, even if it has not been yet completely included as a prominent part of the philosophy of AI. Indeed, the conceptual and practical tools developed within AI offer a stimulus for an innovative approach to some philosophical topics.

Nowadays, the interaction of the two disciplines progressively appears as a significant cross-fertilization in the direction of the disappearance of their respective boundaries. Philosophy plays a relevant role for AI in clarifying its goals and methods, AI offers powerful tools to philosophy in answering several different questions. However, a systematic framework still lacks and several contributions

[*] This paper is dedicated to the memory of my teacher Marco Somalvico.

Minds and Machines **13:** 537–552, 2003.

are exclusively centered just on partial aspects of the problem, despite the various attempts of investigating in an exhaustively way the complex nature of this interplay.

This paper is a preliminary attempt to systematically analyze the nature and the features of the relation between philosophy and AI within the context of the Philosophy of Information (PI). The goal is to propose a stable, comprehensive, and coherent approach to the foundation of the philosophy of AI. The resulting system, called PAI (Philosophy of AI) framework, is articulated both at historical level and at methodological level to respond to different lines of interest.

In order to be tested, the PAI framework is utilized for the analysis of the application of a particular type of multiagent system, called agency, to scientific discovery scenarios. That represents a stimulating topic in philosophy of AI, since it allows to observe in the same context the influence of the philosophy on AI and that of AI on philosophy. According to that, a partial revision of the framework emerges in order to integrate the top-down approach adopted at the beginning of the paper with the bottom-up approach adopted at the end of it.

The paper is organized as follows: in the next section I will delineate the state of the art of the so-called philosophy of AI, letting some problems emerge. In Section 3 I will present the PAI framework articulated at different levels, with the specification of the basis of this approach. In Section 4, after a short presentation of the concepts of multiagent system and agency, I will concentrate on the application of agencies to the context of scientific discovery in order to evaluate how this concrete case of philosophy of AI may be useful as a suggestion to revise the general PAI framework.

2. Toward the Foundation of the Philosophy of AI

This section presents some of the problems related to the characterization of the philosophy of AI as a well-defined and mature discipline and a first simplified state of the art.

2.1. SOME PROBLEMS EMERGING FROM AN HISTORICAL ACCOUNT

Although philosophy and AI have divergences in terms of origins, issues, and methodologies, the two disciplines have been presenting various and different forms of interaction (Ringle, 1979), starting from the Fifties with the advent of AI as a discipline. The name philosophy of AI has been used as the label for including and referring to a very heterogeneous and articulated mass of contributions. The emergence, at the end of the Seventies, of the philosophy of AI as a premature paradigm for the whole information revolution, as it has been made clear by Floridi (2002), has further contributed to make the foundational problem more complicated. From the one side, AI has been considered to play a too emphatic role in the so-called "computer revolution" (Sloman, 1978), stressing the attention only on the revolutionary impact of AI as a new technology. From the other side, the entire situation has obscured the insertion of the philosophy of AI in the wider framework constituted by the PI (Floridi, 2002), with the centrality of the concept of information.

Under the name philosophy of AI it may be collected basically any area of interest in which forms of interaction between the two disciplines are observed. Moreover, in the last years, the interest for the philosophical issues of AI has been partnered with the extensively adoption of AI paradigms and models for addressing philosophical problems, thus promoting a new flurry of research, in particular in philosophy of science (Thagard, 1989) and philosophy of language (Dahl, 1989).

With in this scenario it is possible to observe the following situation: the extensive adoption of the label philosophy of AI without a preliminary clarification of the concepts and the relations lying beneath. In other words, it lacks an analysis of the two terms of the debate — namely philosophy and AI — and their relationship. The lacking of a foundation for this discipline may be perhaps traced back to the lacking of a definite and stable definition of AI. This, however, is an essential part of the discipline which, by nature, is not defined once for all, but is constantly stretched and eventually extended in line with the last technological results (Ringle, 1979; Simon, 1995).

A possible motivation of this unilateral approach to the matter may be individuated in the following causes:

- the tendency to consider the role of philosophy on AI just from an historical perspective, on the basis of the fact that some issues and methodologies adopted by AI derive from philosophy;
- the habit to concentrate just on the influence of philosophy on AI and not, viceversa, on the potential influence of AI on philosophy.

However, one point of this whole tendency must be kept in mind and may be further stressed, namely the acknowledgement of the peculiarity of the relation between philosophy and AI with respect to other disciplines. As John McCarthy states:

> Artificial Intelligence and philosophy have more in common than a science usually has with the philosophy of that science. This is because human level artificial intelligence requires equipping a computer program with some philosophical attitudes, especially epistemological. (McCarthy, 1995)

In my opinion the reason for the peculiarity of this relationship does not lie just in the philosophical attitudes a computer program should have to be considered intelligent. There exist several other reasons for which the evidence of the strong ties between AI and philosophy represents only a starting point. By enlarging the whole scenario, this paper represents a preliminary attempt to include in the same context, with a taxonomical purpose, the different contributions to the philosophy of AI. Of course significative attempts of delineating the connections between philosophy and AI have been already proposed, even if they are mainly in the form of collections of papers (see Boden, 1990; and Cummins and Pollock, 1991, as significative examples). On the contrary, the focus of this paper is on the emergence of some organizational criteria in order to move ahead in the direction of a critical evaluation of the state of the art, with the aim of having a general point of view on the discipline and its particular articulations.

In this foundational effort I am aware of the need of paying attention to a potential criticism. The accusation that a foundational stance entails a reductional stance, as stated in general for all the sciences in Longino (1999). In the current case, a foundational stance for the philosophy of AI is required in order to base it on a systematic and stable framework which can be progressively updated. Moreover, if the framework is well articulated it should maintain the richness and the heterogeneity of the field, while promoting interesting issues from a conceptual point of view.

2.2. STATE OF THE ART

In order to attempt a first systematization for the philosophy of AI, the different contributions to the discipline may be subdivided according to three main sets. The philosophical problems that have to be preliminary afforded at the beginning of AI, the issues deriving from the current interaction between philosophy and AI, and the present and future consequences of that interaction. Let us consider them in more details:

- **Foundational remarks:** they individuate the preliminary questions that had called for a solution at the beginning of AI and that promoted a strong debate between philosophers and AI scholars. Some examples are: the debate about the nature of intelligence (Turing, 1950) and the issues regarding the possibility of achieving intelligent behavior for artificial agents (Searle, 1980).
- **Central issues:** they individuate the topics involving forms of interconnections between AI and philosophy. Some examples are represented by the new trends in the current theories of rationality and, in particular, by computational philosophy of mind (Dennett, 1991) and model-based philosophy of science (Thagard, 1988).
- **Remarkable consequences:** they individuate the issues and problems deriving from the adoption and the use of AI paradigms. Some examples are: human-machine interaction (Winograd and Flores, 1987) and computer information ethics (Bynum, 1985).

Despite its orientation to generality and its purpose of including a wide range of different contributions, this approach presents an essential problem. This state of the art in not able to acknowledge the reciprocal cross-fertilization existing between AI and philosophy. It depicts philosophy of AI as a three-sided discipline, but it is not able to put under the light its main peculiarities, namely the reciprocal influence between philosophy and AI. The problem is represented by the fact that, in accordance with the other philosophies of the different scientific disciplines, it gives evidence just to the influence of philosophy on AI, namely the guiding role performed by philosophy toward AI. However, if what today is called philosophy of AI is evaluated without prejudice, it is worth noting that several contributions regard the applications of paradigms derived from AI to genuine philosophical problems. Therefore, even if this state of the art may represent a promising starting point, it is necessary to design a more articulated framework like that presented in the next section.

3. The PAI Framework

The purpose of this section is to establish the philosophy of AI as the result of the various forms of interaction existing between philosophy and AI. The resulting framework, called PAI framework, shows a double advantage: first, it offers a wide classification of the field which, as said, has received vast attention, but scarce systematization; secondly, it collects different contributions featuring the philosophy of AI both as a new field and as a new methodology.

3.1. THE BASIS OF THIS APPROACH

Before presenting the framework it is important to illustrate the ideas at the basis of it. More than a list of topics and areas of interest this framework is able to offer some criteria to evaluate the relationship between philosophy and AI, by articulating the interaction at different theoretical levels, instead of privileging just the historical dimension.

The basis of this approach can be individuated along two dimensions:

- the parallel evaluation of two different lines of interest, namely the role of philosophy on AI and the role of AI on philosophy;
- the distinction between a foundational level and a methodological level in the philosophy of AI.

Let us consider more in detail the first dimensions: if the analysis of the influence of philosophy of AI corresponds to the traditional role philosophy plays as the philosophy of a discipline, the second one needs some further explanation. As said, usually the philosophy of a given discipline, such as the philosophy of science, concerns the philosophical issues of that discipline, but does not afford conversely the influence of the discipline itself on philosophy. A first motivation of considering two directions of influence within philosophy of AI is to reflect the current literature and practice, evidencing the fact that AI and philosophy are considered to have a great deal in common (see Subsection 2.1). As it will be clearer in the following, the two directions of influence promote an original approach in the direction both of a revitalization of philosophy and of a clarification of the basic principles and methodologies of AI. Moreover, the influence of AI on philosophy is usually afforded in general as part of the philosophy, whereas in this context its belonging to the philosophy of AI and its derivation from it are emphasized.

As regarding to the second dimension, namely the distinction between the foundational level and the methodological level, it represents the core distinction on which the classification proposed in my framework is based. It reflects the conception of the philosophy of AI both as a new field and as a new methodology as outlined in Floridi (2002) for the PI. This states also the connection of the philosophy of AI with the PI, the former one seen as a specific area of the broader scenario of the latter one. The concept of information, or better the integrated theories devoted to the processing, managing, and using information, are the reference frame also for the philosophy of AI. However, within the philosophy of AI these theories are in general given for granted and the attention is devoted to the "intelligent" manipulation and management of information and the philosophical problems connected to it. In the PI one way of constituting and modelling the information environments is the analysis of the concept of information machines. *Information machines* (Amigoni et al., 1999a) are machines whose architecture embeds the particular class of models called information, which regards the activity of processing models (thus, it is a meta-activity). According to that, AI as a subarea of informatics, is devoted to the design of systems which not only process information, but that in processing information realize an intelligent performance, where these results may be obtained by means of a variety of different methods. As a consequence, a consistent part of the philosophy of AI is related to these matters and this debate conversely offers an insight for the investigation of the PI: philosophy of AI is restricted to the fundamental subarea of the analysis of the nature and principles of information, with regards to intelligent activities and tasks.

As a new field, philosophy of AI concerns the critical investigation of the basic principles of AI and its epistemological questions. As a new methodology, it concerns the elaboration and the application of the concepts and tools derived from the theory to different problems. This distinction allows, in addition, to state a difference between the philosophy of AI as an object of study and the philosophy of AI as a method of study, as will be clearer in the following.

In conclusion, the framework presented in the next subsection should be accepted and adopted for both an extrinsic and intrinsic reason. The extrinsic merit is to be one of the first attempts of including in the same frame of reference several different positions and approaches. The intrinsic merit is to include within it perspectives that usually do not have space in the other systematizations. In particular, the parallel consideration of the influence of philosophy on AI and of AI on philosophy and the presence of a criterion in the taxonomical activity based on the idea of considering the philosophy of AI as a new field and as a new methodology.

3.2. THE GENERAL FRAMEWORK

According to the aim of considering both the role of philosophy on AI and the role of AI on philosophy, let us start by analyzing the role of philosophy on AI in order to illustrate the PAI framework. This may be articulated in a foundational level and a methodological one, which represent respectively the philosophy of AI as a new field and the philosophy of AI as a new methodology. Moreover, each of these levels is further articulated in an historical level and a conceptual one in order to clearly distinguish these two different lines of interest.

- **Foundational Level:** individuates the philosophy of AI as a new field regarding the contributions of philosophy in the founding process of AI and can be further articulated in:
  - Historical level: AI affords some of the problems traditionally afforded in the history of philosophy, thus promoting a *communality of problems* between the two disciplines. Significative examples can be considered theories of reasoning and learning (Bratman et al., 1991) and connections between knowledge and action (Pollock, 1991).
  - Conceptual level: philosophy specifies some of the ideas of AI, thus stating the *basic concepts* of the discipline. An example is represented by the research on the features that an artifact must possess in order to judge it as "intelligent" (McCarthy, 1999).
- **Methodological Level:** individuates the philosophy of AI as a new methodology regarding the contributions of philosophy in offering conceptual tools to AI and can be further articulated in:
  - Historical level: AI utilizes methodologies developed by philosophers, thus individuating a *communality of methodologies and tools* between the two disciplines. The use of BDI theories to design communication languages for artificial agents represents an interesting example (Bratman, 1987).
  - Conceptual level: philosophy evaluates the structural notions involved in AI, accounting for its *critical role*. The enrichment of the theories of rationality adopted in AI as the basis of philosophical concepts represents an example (Elgot-Drapkin et al., 1991).

Let us consider now the role of AI on philosophy: this is relative to the use of methodologies and tools of AI to address traditional and novel philosophical problems from a new perspective. Also in this case the interface can be organized in a foundational level and a methodological one, which represent respectively a revision of some classical theories and an implementation of AI tools within philosophical issues.

- **Foundational Level:** individuates the philosophy of AI as a new field, promoting a *theory revision*. Philosophical theories are expressed in computational terms or as programs meeting the requirement of rigor that promotes a new and more formal approach to the philosophy. A classical example of this turn in philosophy is represented by the passage form the old philosophical

Table 1. The PAI framework

	Philosophy→AI		*AI→Philosophy*
	Historical level	Conceptual level	
Foundational level	Communality of problems	Basic concepts of AI	Theory revision
Methodological level	Communality of tools	Critical role of philosophy	Paradigm implementation

task of explaining the mind to the design stance toward the mind, namely the investigation of mind through its mechanisms, capabilities and evolutions (Churchland, 1990).

– **Methological Level:** individuates the philosophy of AI as a new methodology with the characterization of a *paradigm implementation*. Some programs and tools of AI are utilized as practical means for approaching philosophical problems. An interesting example is the use of neural nets to evaluate the level of coherence of some scientific explanations (Thagard, 1989).

As said, one of the main advantages of this framework is that it captures several different approaches to the problem in the same scenario and gives direction to a more complete and vast systematization of the philosophy of AI. It represents one of the first systematic step for the foundation of the philosophy of AI as an autonomous and mature discipline. In the next section the PAI framework will be evaluated at the light of a concrete example integrating the top-down approach adopted until this point with the bottom-up approach promoted by the analysis of a case-study. This further step has two purposes: to pragmatically test the PAI framework in the classification of an example and, from the considerations derived, to expand and revise it.

4. The Agency Paradigm

This section addresses the description of a paradigm derived by the adoption of a particular multiagent system, called agency, in a scientific context. I will show how this paradigm completes the PAI framework by testing and reviewing it. The result will be the elimination of the spurious separation between the influence of philosophy on AI, from the one side, and the influence of AI on philosophy, from the other one.

4.1. MULTIAGENT SYSTEMS AND AGENCIES

Multiagent systems are becoming an increasingly important paradigm for developing "intelligent distributed systems" (Ferber, 1999; Weiss, 1999; Woolridge, 2002). Originated from distributed artificial intelligence (Bond and Gasser, 1988), multiagent systems constitute now an autonomous area with a number of techniques, methods, technologies, and tools.

In this section I do not exhaustively survey all the issues of multiagent systems, but I concentrate only on those that make them an appropriate paradigm to be employed within modern scientific discovery scenarios. In particular, the attention is concentrated on a special class of multiagent systems in which the agents cooperate and are oriented toward a single global goal. Those cooperative multiagent systems can be conveniently called *agencies* to stress their unitary nature when they address a single global problem (Amigoni et al., 1999b).

An agency can be considered as a unique complex machine devoted to a single task. It is complex because its components are agents, namely processing machines, such as computers and robots, able to perform inferential abilities (i.e., to automatically infer conclusions by means of axioms and inferential rules). It is unique because the agents are not acting independently but are cooperating in a coordinated way to achieve a global goal.

The nature and the role of an agency can be better understood when the origin of this concept is considered. It was firstly introduced by Marvin Minsky (1985) under the metaphor of "the society of minds". Minsky's goal was to overcome the difficulties posed by the complex nature of the phenomena of human intelligence in order to reach their deep understanding and their satisfactory representation within given models. Minsky considered an agent as an individual entity, where a particular and specific way (paradigm) of modelling a given phenomenon of intelligence is embedded into the functional architecture of the agent itself. Both the plurality of the phenomena to deal with and the variety of reasonable paradigms that can be adopted for modelling a given phenomenon suggest a scenario in which a high number of agents coexist and collectively contribute to set up a rich, comprehensive, and precise description of human intelligence. Minsky adopted the term 'agency' to denote such system of agents, each one representing a descriptive paradigm of a given phenomenon.

Starting from this initial abstract characterization, the concept of agency has been concretely employed in distributed artificial intelligence and robotics. In this perspective an agency is a unitary machine whose agents, although having complex natures to perform high-level functions, cooperate in order to achieve a global goal.

An agency is characterized by two central properties that allow its use appropriated in a different number of applications: multiparadigmatic nature and flexibility. These two properties derive from the particular architecture that characterizes each agent of an agency. Each agent is structured as a couple of semiagents: the *op semiagent* and the *co semiagent* (where op is for "operative" and co for "cooperative"). The first one is devoted to perform specific tasks and may be different for each agent composing the agency, whereas the second one is devoted to cooperate and must be the same for each agent. The interplay between these two components allows the simultaneous presence of different paradigms and the easy insertion and elimination of other ones.

Let us consider now the multiparadigmatic nature of agency. As in the case of Minsky's initial approach, it is natural to embed in the agents composing an agency different paradigms for solving a problem or for achieving a goal. Cooperation among the agents harmonizes these different paradigms in a coordinated effort to solve a global problem or to achieve a global goal, thus enlarging the range of problems that an agency can tackle.

Let us consider then the flexibility of an agency. Since an agency is composed of complex and relatively independent components like the agents, it is usually easy to modify its modular composition in order to exploit the best combination of agents for tackling a given problem. This argumentation is supported by the observation that many of the cooperation mechanisms presented in the literature usually scale well to large numbers of agents.

4.2. SCIENTIFIC AGENCY

4.2.1. *Agency and Scientific Discovery*

That of multiagent systems is a research area in which AI and philosophy present strong connections. AI is involved since the presence of autonomous and intelligent agents. Philosophy is involved since the attention to the notions of intelligence and rationality, both in the classical meaning of the "intelligence" of a single agent and in the new perspective of the "intelligence" and interaction of societies of agents. In particular, the structure of an agency allows it to afford a variety of problems. The basic idea of each application is that of having a sophisticated device which, due to its flexibility, is able to emulate some of the human intellectual activities. Among the activities in which agencies can play an important role, one of the most interesting is represented by scientific discovery which is, at the same time, one of the peculiar expressions of human creativity and a challenging field of application for "intelligent" machines.

In this context AI programs and devices may cover a wide range of roles: basically for all of them the purpose is to emulate some human intellectual activities performed during the scientific discovery process, such as hypothesis construction, theory revision, law induction, and theory formation. This application has promoted what has been called *computer-supported scientific discovery* (de Jong and Rip, 1997), where the emphasis is not on the autonomy of machines (Langley et al., 1987), but rather on their role as supports for scientists in complicated scientific processes (Langley, 2002).

Moreover, the theme of scientific discovery process is one of the traditional areas of interest for the philosophy of science that has constantly tried to explain the mechanisms and processes presupposed by scientific activity in order to give account for the development of scientific knowledge (see Popper, 1959, as significative example).

Scientific discovery represents thus a field of interest for both AI and philosophy. In particular, an agency applied to scientific discovery and called *scientific agency* is able to play a central role, having interesting implications both in the case of AI and in the case of philosophy (Amigoni et al., 2002). As an AI tool it can support with high success scientists in their activities (*assistant agency*): the properties previously described give account for the possibility of having a particularly successful device in performing some of the processes involved in scientific discovery. Besides that, the particular architecture of an agency allows for the possibility of representing the obtained scientific results (*representational agency*), giving account to a more rigorous approach for the philosophy of science in the explanation of scientific discovery processes. The dialectics between assistant agency and representational agency will show the role of the agency paradigm as a case-study within the PAI framework and in the direction of a better foundation for the philosophy of AI.

4.2.2. *Assistant Agency*

Usually scientists exploit a wide number of tools in carrying on their work. Information machines (e.g., computers and robots) are in a prominent position among these, since a larger and larger number of not only practical, but also intellectual, activities can be delegated to them both for necessity (e.g., huge quantity of data) and for convenience (e.g., speed increasing). Scientific agency, according to its nature of concrete, flexible, and powerful machine, represents a particularly useful support for scientists during the process of scientific discovery. In this case it is called assistant agency.

Besides being a collection of information machines supporting scientists, an assistant agency is a cooperation machine that offers a valid support for the social nature of the contemporary scientific research. Even if there are implemented agencies to address different applications, the agency technology has not yet been fully developed in the scientific context. However, for its paradigmatic nature and its flexibility it represents a promising trend in this direction.

The role of agency as a support for scientists within scientific discovery can be interpreted by using the PAI framework. It gives an account for the contributions that philosophy offers to AI, since assistant agency allows observing the role of philosophy on AI, more precisely on the specific area of AI represented by multiagent systems. First of all, it is worth noting that sophisticated and complex tools, such as agencies and scientific agencies, are stimulated by a general philosophical contribution to AI both on a foundational and a methodological level.

The main contribution is represented by the philosophical investigation and the critical inquiry philosophy provides of concepts like cooperation, interaction and coordination. In order to develop an agency, which is a cooperation machine, a coherent framework for the concepts of interaction and cooperation is needed as the natural starting point in conceiving, designing and building agents that coexist and act in the same environment.

4.2.3. *Representational Agency*

The second role of agency, as description of scientific results, is perhaps less intuitive, but fundamental in approaching explanations of scientific discovery. This agency describes, in a concrete way, the set of models resulting from a scientific effort in accordance with van Fraassen (1980) and Giere (1988) to see scientific discovery as a creation of adequate models to describe phenomena. In the case of a representational agency these models are embedded in the agents, providing a descriptive (when the models are simply stored in the agents) or a more powerful operational (when the models result from the agents activity) representation of scientific knowledge. The adoption of a representational agency in a descriptive function offers not only a more formal description of the models resulting from a scientific effort, but also an improvement in managing the interaction among the different models produced by the scientific process.

The role of an agency as representation of scientific discovery can be interpreted within the PAI framework as well: representational agency allows to observe the role of AI on philosophy, namely the influence of agencies in promoting a new approach to philosophy of science by answering to the demand of rigor of philosophy. An agency therefore may represent the set of models resulting from the scientific effort both in a metaphorical way, as a descriptive representation, and in a concrete way, as an operational representation. If in the first case the models resulting from the scientific discovery process are just conceptually represented as agents of an agency, in the second case the models are physically inserted and implemented in the composing agents. According to the PAI framework, the metaphorical description is related to the foundational level of the contribution of AI to philosophy as theory revision: the models composing the scientific discovery process are described in a sort of computational manner. Moreover, when the description promoted by scientific agency is not only metaphorical, but also concrete and implemented in an agency, the methodological level of the framework is presented as paradigm implementation. The key point is represented by the fact that the description provided by an agency is in this case concrete, namely is embedded in a physical agency machine which in its architecture and its mechanisms displays the description itself.

4.3. SOME FURTHER CONSIDERATIONS ON THE PAI FRAMEWORK

Through the concepts of assistant and representational agency and their explanation in the context of the philosophy of AI, the PAI framework has been tested in a concrete case. The result is its validity as a founding block for the philosophy of AI. However, a further interesting point must be stressed with the help of the agency paradigm: the extension and the completion of the PAI framework in the direction of a progressively better adherence of it to the current trends of the philosophy of AI.

A first interesting point is the possibility to mutually integrate the two roles of assisting and of representing of a scientific agency: *circular evolution* is the property expressing this integration. The property is related to the possibility of implementing both the assistant agency and the representational agency in a unique physical agency, which is able to contemporaneously perform both roles. In this way, the representation of new results, provided by the representational agency, and the discovery environment, based on the assistant agency, can mutually improve each other. Some results of the scientific enterprise, expressed by agents of a representational agency, can be physically inserted in an assistant agency. Therefore, this new enhanced machine supports the production of new results that, in turn, are employed to further empower the tool in an endless evolutionary process.

The property of circular evolution puts under the light the deeper integration of philosophy and AI with respect to other topics of interest in the philosophy of AI. It represents a concrete articulation between the two disciplines since the two roles a scientific agency is able to perform, which correspond to the role of philosophy on AI and the role of AI on philosophy, are implemented in the same physical machine, namely the scientific agency. So, the analysis of the innovative agency paradigm does not represent an alternative framework, but just a particularly interesting case-study in which observing, at the same time, the mutual influence of philosophy and AI. As a consequence it offers a contribution for the expansion of the framework adopting the property of circular evolution as a starting point. The expansion can be both along the foundational level and the methodological one.

- **Foundational Level:** individuates the philosophy of AI as a new field and can be labelled as *agency topics*. In the same research area, the agency paradigm deals with the traditional founding themes of AI (such as the concepts of intelligence, rationality, autonomy) in the light of one of the last frontier of AI represented by multiagent systems. Moreover, it presents promising approaches for the philosophy (such as the metaphorical representation of scientific models by means of scientific agency).
- **Methodological Level:** individuates the philosophy of AI as a new methodology and can be labelled as *agency methods*. In the same research area the agency paradigm offers philosophical tools to critically evaluate and improve the AI practice, some of which have been specifically developed to deal with the problems deriving from the adoption of multiagent systems (such as the

Table 2. The PAI framework revised

	Philosophy→AI		*AI→Philosophy*	*Philosophy and AI*
	Historical level	Conceptual level		
Foundational level	Communality of problems	Basic concepts of AI	Theory revision	Agency topics
Methodological level	Communality of tools	Critical role of philosophy	Paradigm implementation	Agency methods

analysis of the various forms of interaction like cooperation and competition). Moreover, it offers concrete tools to revolutionize the analysis of some philosophical problems (such as the operational representation of scientific models by means of scientific agency).

In conclusion, the analysis of the agency paradigm stimulates further reflections. First of all, the necessity to constantly integrate the general and somehow abstract framework with the analysis of concrete examples. In this view, the PAI framework plays the role of a structure of reference which can be updated and improved in each specific case. Secondly, the two new parts of the framework (agency topics and agency methods) are not just the simple union of the previous cases, but, although they derive from those, express some new features which are observable within the agency paradigm. In accordance to that, the labels of the new fields of the framework contain the reference to the idea of agency. This is why at the moment the real integration between the AI component and the philosophy one is achieved within this field. That does not exclude in the future to expand it to other fields of interest and as a consequence to insert some more general labels. Finally, it is interesting to note how from this perspective the trends in the current philosophy of AI are observable: I should say a trend in the direction of areas of interests and research that exploit several levels of interconnection in the same topic or application.

5. Conclusions

In this paper I have presented a framework to give reason in a systematic way to the different forms of interaction existing between philosophy and AI. The motivation was in starting to put the basis for a well-founded philosophy of AI, in the same direction for instance of the philosophy of mathematics. The starting point was the acknowledgement of a variety of different contributions that, however, do not find a stable and coherent placing in a traditional state of the art.

From that, it has emerged the idea of proposing a framework articulated at different levels capable of giving reason both to the influence of philosophy on AI and to the influence of AI on philosophy. Moreover, I have illustrated an application to concretely integrate and complete the PAI framework in the direction of the simultaneous influence philosophy and AI can exert one on each other in specific areas of the philosophy of AI. The application of agencies in particular, and of multiagent systems in general, to scientific discovery offers an example of what the philosophy of AI is today: a complete and fruitful field of integration between philosophy and AI.

Future research work will address the refinement of the PAI framework in the direction of other specific areas of the philosophy of AI: that will be progressively tested with the application to a larger number of concrete examples, where to observe interesting forms of influence between philosophy and AI. Moreover, the scientific agency context will be an object of interest with the implementation of the first prototypes of scientific agencies to real world examples of scientific discoveries.

Acknowledgements

I wish to thank the anonymous reviewers for their useful comments.

References

Akman, V. (2000), ‘Introduction to the Special Issue on Philosophical Foundations of Artificial Intelligence’, *Journal of Experimental and Theoretical Artificial Intelligence* 12, pp. 247–250.

Amigoni, F., Schiaffonati, V. and Somalvico, M. (1999a), ‘Processing and Interaction in Robotics’, *Sensors and Actuators A: Physical* 72, pp. 16–26.

Amigoni, F., Somalvico, M. and Zanisi, A. (1999b), ‘A Theoretical Framework for the Conception of Agency’, *International Journal of Intelligent Systems* 14(5), pp. 449–474.

Amigoni, F., Schiaffonati, V. and Somalvico, M. (2002), ‘Multiagent Systems for Supporting and Representing Social Creativity in Science’, in *Proceedings of AISB’02 Symposium on Artificial Intelligence and Creativity in Arts and Science, The Society for the Study of Artificial Intelligence and Simulation of Behaviour*, London.

Boden, M. (1990), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Bond, A. and Gasser, L. (1988), *Readings in Distributed Artificial Intelligence*, San Mateo, CA: Morgan and Kaufmann.

Bratman, M. (1987), *Intention Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Bratman, M. Israel, D. and Pollack, M. (1991), ‘Plans and Resource-Bounded Practical Reasoning’, in R. Cummins and J. Pollock, eds., *Philosophy and AI*, Cambridge MA: MIT Press, pp. 7–21.

Bynum, T.W. (1985), *Computers and Ethics*, Oxford Readings in Philosophy, Oxford: Oxford University Press.

Churchland, P. (1990), ‘Some Reduction Strategies in Cognitive Psychology’, in M. Boden, ed., *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, pp. 334–367.

Cummins, R. and Pollock, J. (1991), *Philosophy and AI*, Cambridge, MA: MIT Press.

Dahl, O. (1989), ‘Contextualization and De-Contextualization’, in R. Studer, ed., *Natural Language and Logic*, Lecture Notes in Artificial Intelligence, Berlin: Springer-Verlag, pp. 62–69.

Dennett, D. (1991), *Counsciousness Explained*, Boston: Little Brown & Company.

de Jong, H. and Ripp, A. (1997), 'The Computer Revolution in Science: Steps Toward the Realization of Computer-Supported Discovery Environments', *Artificial Intelligence* 91, pp. 225–256.

Elgot-Drapkin, J., Miller, M. and Perlis, D. (1991), 'Memory, Reason, and Time: The Step-Logic Approach', in R. Cummins and J. Pollock eds., *Philosophy and AI*, Cambridge, MA: MIT Press, pp. 79–103.

Ferber, J. (1999), *An Introduction to Distributed Artificial Intelligence*, Reading, MA: Addison-Wesley.

Floridi, L. (2002), 'What is Philosophy of Information?' *Metaphilosophy* 33(1/2) pp. 133–145.

Giere, R. (1988), *Explaining Science. A Cognitive Approach*, Chicago: The University of Chicago Press.

Langley, P., Simon, H., Bradshaw, G. and Zytkow, J. (1987), *Scientific Discovery: Computational Explorations of the Creative Processes*, Cambridge, MA: The MIT Press.

Langley, P. (2002), 'Lessons for Computational Discovery of Scientific Knowledge', in *Proceedings of First International Workshop on Data Mining Lessons Learned*, Sydney, Australia.

Longino, H. (1990), *Science and Social Knowledge*, Princeton, NJ: Princeton University Press.

McCarthy, J. (1995), 'What Has AI in Common with Philosophy?' in *Proceedings 14th International Joint Conference on AI*, Montreal, Canada, August.

McCarthy, J. (1989), 'Philosophical and Scientific Presuppositions of Logical AI', in H.J. Levesque and F. Pirri, eds., *Logical Foundations of Cognitive Agents: Contributions in Honor of Ray Reiter*, Berlin: Springer.

Minsky, M. (1985), *The Society of Minds*, New York: Simon & Schuster.

Pollock, J. (1991), 'OSCAR: A General Theory of Rationality', in R. Cummins and J. Pollock, eds., *Philosophy and AI*, Cambridge, MA: MIT Press, pp. 189–211.

Popper, K. (1959), *The Logic of Scientific Discovery*, New York: Harper and Row.

Ringle, M. (1979), *Philosophy Perspectives in AI*, Atlantic Highlands, NJ: Humanities Press.

Searle, J. (1980), 'Minds, Brains, and Programs', *The Behavioral and Brain Sciences* 3, pp. 417–424.

Simon, H. (1995), 'Artificial Intelligence as an Empirical Science', *Artificial Intelligence* 77, pp. 95–127.

Sloman, A. (1995), 'A Philosophical Encounter', in *Proceedings 14th International Joint Conference on AI*, Montreal, Canada, August.

Sloman, A. (1978), *The Computer Revolution in Philosophy*, Atlantic Highlands, NJ: Humanities Press.

Thagard, P. (1988), *Computational Philosophy of Science*, Cambridge, MA: MIT Press.

Thagard, P. (1989), *Conceptual Revolutions*, Princeton, NJ: Princeton University Press.

Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind LIX* 2236, pp. 433–460.

van Fraassen, B. (1980), *The Scientific Image*, Oxford: Clarendon Press.

Weiss, G. (1999), *Multiagent Systems: An Introduction to Distributed Artificial Intelligence*, Cambridge, MA: MIT Press.

Winograd, T. and Flores, F. (1987), *Understanding Computers and Cognition. A New Foundation for Design*, Reading, MA: Addison-Wesley.

Wooldridge, M. (2002), *An Introduction to Multiagent Systems*, New York: John Wiley & Sons.