Herbert Simon, the anti-Philosopher

Stefano Franchi
Department of Philosophy, The University of Auckland, New Zealand, s.franchi@auckland.ac.nz

Abstract. The essay discusses the relationship between classical philosophy and Artificial Intelligence through an examination of Herbert Simon's work. Most practitioners consider AI a technical disciple aimed at the production of "smarter" artifacts. Skepticism for the high-minded ambitions of classical philosophy runs high, and the prevalent interest has been focused on specific collaborations at the level of particular theories of, say, perception, or concept-formation, etc. The essay argues that research in AI, at least in its formative years (late 40s-early 60s), had a much more complex relationship to philosophical research and even to metaphysics. On the one hand, AI arose in a self-conscious opposition to the methods of classical philosophy, and turned toward the sciences in order find both a methodological canon and concrete tools for its analyses. At the same time, however, it accepted the traditional philosophical project of providing an exhaustive theory of human nature. It saw itself as a "new philosophy" and indeed as an *anti-philosophy* that aimed at recovering the goals and scope of the millennia-old attempts toward an exhaustive account of man and his place in the cosmos, while replacing arm-chair speculations with a radically new kind of empirical approach. The claim will be illustrated and defended with reference to the development of the work of Herbert Simon.

1. The canon of AI and the work of Herbert Simon

Herbert Simon's work presents a curious anomaly to the historian and philosopher trying to understand the development of classic Artificial Intelligence (AI). Simon was one of most influential figures in AI since its birth, and yet it is always with some difficulties that his work can be made to fit within the received canon of AI's development and goals. In fact, he differed from every other figure in early AI on most counts: in terms of the recognized intellectual heritage of AI, of his own background and training, of the goals he set for his own AI work, and the assessment criteria he accepted.

I will argue that these differences provide important clues toward a reevaluation of the relationship between Artificial Intelligence and Herbert Simon's work that may change our understanding of both. Classic Artificial Intelligence provided Simon with the tool he needed to pursue a much broader research agenda that strove to encompass human beings in their cognitive, emotional, social, and political dimensions. Conversely, AI's peculiar status as the discipline that straddles the boundaries between engineering, science, and

philosophy can be recast as the tool that allowed researchers to pursue philosophy's old goals with an alternative methodology. From this perspective, Simon's version of Artificial Intelligence becomes, I will argue, a full-fledged form of "anti-philosophy" as ambitious and broad-ranging as old-fashioned metaphysics and as revolutionary as the latter in the radical refashioning of its methodology. My contention is that a philosophical assessment of AI, at least in its Simonian incarnation, must be more far-reaching than it is usually practiced. At the methodological level, it must discuss whether computer simulation as practiced in AI is really adequate to overcome the *a priori* / *a posteriori* distinction by actually producing the behavior it wants to explain. At the content level, a philosophical assessment must be willing to broaden the scope of AI theories and consider them a complete account of human nature. The present paper prepares the grounds for such a twofold evaluation, by offering a historically-based discussion of the relationship between research in Artificial Intelligence and Herbert Simon's work as a whole.

Let us begin with "received view" about AI. Although historians, philosophers, and AI practitioners have differing views on AI's intellectual content, its ancestry, goals, and methods, it is possible to extract a number of shared claims about it. The effort to produce artifacts that will simulate and reproduce the cognitive skills of humans is often linked to the long-standing effort in the Western philosophical tradition to reduce thought to the manipulation of symbols through the application of formally specifiable rules (Dreyfus, 1972; Mazlish, 1993; Franchi and Güzeldere, 2005). This emphasis on symbol-processing as the core of AI research has brought many scholars to emphasize the pivotal role played by Alan Turing's work on the theory of computation. Turing's own extension of computability to cognitive procedures and his criterion for intelligence, the so–called Turing test, are considered the cornerstones of AI's intellectual foundations. Although scholars have acknowledged that Turing did not exert a direct influence in the early years of AI's development, it is often claimed that Turing's work provides the conceptual core for research in Artificial Intelligence. This interpretation allows to locate AI within the intellectual universe of computer science, thus providing the necessary start for the assessment of AI from a conceptual standpoint (Copeland, 1993). The connection between AI and computer science so established naturally extends to engineering at large, bringing to the fore AI as the producer of artifacts whose success must be evaluated on the basis of their performance, regardless of whether the underlying processes the machine use bear any similarities to the corresponding human processes. Margaret Boden's definition nicely sums up most of these claims: *"AI is the study of how to build and/or*

program computers to enable them to do the sort of things that minds can do" (Boden, 1990, p. 1, reflecting Minski, 1968).

Herbert Simon, as most historians recognize, shared very few of these claims. In spite of the central role Simon (and Newell) played in the development of AI up to the early 60s, none of the thesis about AI briefly listed above apply to his work, or to his person. He was neither a logician nor a mathematician when, as he famously reported to a befuddled classroom of Carnegie-Mellon graduate students in 1955, he "invented a thinking machine over the Christmas break" (Simon, 1991, p. 209). His background was in economics and organization theory; he was utterly dismissive of measuring AI progress by the application of "behavioral" tests, like the famed game invented by Turing; he was not even particularly close to the work of Turing (Simon, 1991, p. 193), but rather to von Neumann's, whose game theory is not usually given much credit in the development of AI. He may have been interested in the production of artifacts, as his collaboration with numerous DARPA projects attest, but he was unwilling to tie AI research to them. From the standpoint of AI's canonical history, Herbert Simon seems to be an outsider working at the margins of the discipline.

Yet, the opposite is true. It is well-known that when Simon and Newell made their appearance at the 1956 summer seminar in Dartmouth that officially christened AI as such, they stole the show from John McCarthy, Marvin Minsky, and the other organizers. As the composition history of Minsky's manifesto "Steps Toward Artificial Intelligence" attests, Simon and Newell's approach – heuristic search– was recognized as the theoretical backbone for classic AI research for the next 30 years, until connectionist paradigms first (see McClelland and Rumelhart, 1986, and Clark, 2001 for an updated synopsis), and *nouvelle* AI later (Brooks, 1999), began to erode its unquestioned supremacy. McCorduck (1979) recounts how Minsky's original draft, already being composed in 1956, gave a very limited space to heuristic search. However, Minsky's position changed and the final version of 1963 recognized Simon and Newell's approach as the conceptual foundation of AI research, while all the other subfields (learning, planning, etc.) were to play the role of ancillary sub-disciplines (Minsky, 1963).

The distance between Simon's and AI's canonical views have often been noted. Such idiosyncrasies are usually ascribed to the versatility of a multifarious genius whose ideas could hardly be restricted to the disciplinary requirements of any given field. Conversely, scholarly reconstructions of AI have reduced Simon's reflection on it to just one position among many, perhaps one of the most extreme. Yet, these solutions cannot dispel the anomaly that Simon's work represents. It seems that either the AI canon is wrong, or Simon's extra-AI

scientific work is irrelevant. I suggest that the most fruitful approach to this interpretive problem is to reconsider Simon's career as a whole, and to reposition his work on AI. The first steps in this direction will be to determine the scope and unity of Simon's work.

2. How many Simons?

Recent scholarship on Herbert Simon's work has debated whether it underwent a sea-change in the 50s, when Simon (with the substantial collaboration of Allen Newell) developed the basic ideas of Artificial Intelligence. Some scholars have claimed that Simon's interests shifted dramatically from the analysis of decision-making processes within large organizations to the study of problem-solving in general (Sent, 2000; Guice, 1998; Mirowski, 2001). Others have claimed, with a good deal of support from Simon himself, that his research focus never wavered. As Simon once declared, he was a "monomaniac" who studied just one topic his whole life: human decision-making (Simon, 2001; Augier, 2000). These different interpretations reflect a deeper dilemma. On the one hand, it cannot be denied that Simon's prodigious scientific output was spread over several different fields: from municipal government to the theory of organizations; from philosophical topics like the theory of causality, the formation and validation of scientific theories, and the conception of rationality, to economic themes like the theory of the firm and the structure of economic organizations; from computer science subjects like list processing languages and Artificial Intelligence's concepts and techniques, to classic psychological themes like perception, memory, concept-formation, and understanding; from sociological problems like group membership and interactions among social groups, to traditional philosophical problems from ethics and politics.[1] *Prima facie*, Simon looks like a polymath, a 20th century Renaissance man whose considerable genius could not be defined by any discipline. Simon's contributions should then be considered and assessed in isolation, as different facets of the so many existing "Simons." A weaker version of the same thesis maintains that Simon's work can be construed as belonging to two different phases: a first period, roughly covering the 1940s and 1950s, in which he was interested in economics and more precisely in the analysis of organizations; a second period, from the late 50s to the end of his life, would see his interests shift radically toward psychology and cognitive science (Sent, 2000; Guice, 1998).

[1] A complete bibliography of Simon's work is available at: http://www.psy.cmu.edu/psy/faculty/hsimon/ An excellent survey of Simon's research is provided by Augier (2000).

Both the stronger and weaker version, however, run into difficulties. First of all, Simon himself strongly resisted it: "what I am monomaniac about - he stated - is decision-making" (Feigenbaum, 2001) or, equivalently, problem-solving.[2] A cursory analysis of Simon's writings easily confirms the author's interpretation. Indeed, the same general theses about the capacities, strategies, and processes needed to take decisions and solve problems recur in almost all of his works. For instance, Simon's insistence on the necessarily limited computing resources available to human beings forms the backbone of his work in organization theory as well as in AI's research. He points out that decision-seeking managers never look for optimal solutions but always settle for "satisfactory ones," the same way as AI's programs rely on "heuristic" rules instead of exploiting optimal algorithms.

There is, however, a rather large explanatory gap between the alleged topic of Simon's "monomania"—be it decision-making or problem-solving—and the scope and content of his research. Although a sustained engagement with problem-solving can be found in most of his works, this fact alone does not tell us much about the content of that research. In order to stretch problem-solving behavior to cover all of it, as Simon maintained, we are forced to turn problem-solving and decision-making into an overly generic category that can be applied to almost every kind of human or nonhuman behavioral pattern, until it becomes too broad to be still meaningful. Almost any kind of purposive behavior can be considered as an instance of "problem-solving." What does it mean to argue, as Simon repeatedly did, that it provides the red thread unifying all his scientific effort? Instead of marveling at the sheer breadth of his interests, we should wonder why such a broad research theme was confined within a handful of disciplines and why he never entered biology, or neurology, or any other discipline that may be concerned with the study, from whatever angle, of purposive behavior. Conversely, as soon as we try to provide a more precise definition of decision-making and problem-solving it becomes very difficult to make it stick to fields as diverse as ethics, perception, management, etc.

The unfortunate consequence of this situation is that both lines of interpretations seem to converge on a middle position that is bound to leave everyone unhappy. Faced with the difficulties presented by both the "unitary" and the "polymath" interpretations, many scholars (and Simon himself, at times) settle for a succession of two independent phases loosely connected by a broadly defined interest in decision-making that I mentioned above.[3] But this compromissary thesis does not really tell us much about Simon's research. Why

---

[2] See Simon (1995, p. 501). It should be noted that Simon had used the same words about 40 years earlier, according to Feigenbaum (2001, p. 2107).

[3] See Simon (1991, p. 189) and Augier and March (2002, p. 15).

did he, while in the pursuit of decision-making, switch from economics to cognitive science and not, for example, to neurology, or politics, or biology *tout court*, just to name a few theoretically compatible alternatives? Is there more than a family resemblance between business managers' decision-making processes as studied by the economist and the problem-solving strategies that the psychologist examines? These questions point to one conclusion: either Simon's career is a collection of largely independent analyses, or the motif that provides the underlying unity to Simon's research is not problem-solving.

3. *Prima philosophia*

My thesis is as follows: Herbert Simon's complex and variegated undertaking is a *general science of human beings*. The unifying subject of his research is Man "with the full glands and viscera," as Simon declared in *The Sciences of the Artificial* (Simon, 1969). "Problem-solving" (or, alternatively, "decision-making") is the *angle* from which he approached it, not the subject of his research. Since Man is essentially a problem-solver, in the rather specific sense that problem-solving is his most distinctive trait (his specific difference, so to speak), it follows that the best way to study Man is to look at his problem-solving activities. I believe that the solution to the problem of problem-solving is to be found in what Simon calls "bounded rationality." According to Simon, Man is essentially a "bounded-rational animal"; the characteristics and specific processes of bounded rationality are comprehensive enough to encompass both the individual in his rational and non-rational components (i.e. the emotions), as well as the society in which he lives (i.e. the concrete institutions). In view of such breadth, I will argue that Simon's research can be considered from three increasingly broader perspectives. In the first instance, it should be construed as anthropology; second, we should realize that the scope of Simon's anthropology is so broad to qualify as a universal epistemology; finally, we should acknowledge that the content encompassed by Simon's work, but not its form, is remarkably similar to the content of good old-fashioned metaphysics.

A change of perspective is all is needed in order to appreciate how Simon's research can be construed as a general anthropology. Let me go back to his pronouncements about his alleged "monomania" about problem-solving. The issue at stake is whether Simon was looking at "problem solving as it happens to occur in humans" or, as I suggest, at "humans as problem solvers." The two formulations point to different research programs which could be extended over different academic disciplines. Problem-solving can be studied in either its concrete instantiations in different social and institutional settings, or from the

point of view of its inner mechanisms, or of its results, etc. In the first case, problem-solving would be just one among many other relevant human activities and processes deserving our interest. In the second case, problem-solving is the single process that best explains the most relevant and unique aspects of human life.

This is exactly what Simon did throughout his career. At the beginning of *Administrative Behavior*, for instance, he writes that "insight into the structure and function of an organization can best be gained by analyzing the manner in which the decisions and behavior of [...] an employee are influenced within an organization" (Simon, 1947, p. 3). Simon extends the applicability of the theory he presented in his first book when, in the preface to the third edition published 30 years later, he points out that his work is addressed not only to organization designers but also to organization "watchers." Who are these, you may ask? It turns out that we all are:

> We are organization watchers in our role as citizens. Increasing attention has been fixed in recent years upon the functioning of society's organizations: its large corporations and its governments. Hence this could also be described as a book for Everyman–for it proposes a way of thinking about organizational issues that concern us all (Simon, 1975, p. ix).

In short, "decision-making processes hold the key to the understanding of organizational phenomena" (*cit.*, p. xl) and organizational phenomena concern us all, we have to place individuals within the organizations in which they operate in order to understand them. As Simon clearly states in a passage of vaguely Hegelian tones, "it is impossible for the behavior of any single isolated individual to reach any degree of rationality" (*cit*., p. 79). Similar passages may be found in all of Simon's writings, although they tend to be concentrated in the prefaces to his collections of essays (e.g. Simon, 1957, p. 1, 1979, p. x, 1957, p. 1, etc.).[4] The clearest formulation of the general anthropological aim of Simon's undertaking can be found in *The Sciences of the Artificial,* where Simon attempts to sketch the outline of a comprehensive scientific understanding of everything "artificial." With a somewhat counterintuitive terminology, he defines as

[4] The first reference is to *Models of Man*, the second one is to *Models of Thought*, where Simon states that "the Problem Solving Man of *Human Problem Solving* has simply been generalized to the whole (or nearly whole) Thinking Man who will be found in these pages." In their often quoted article (Newell and Simon, 1963, but originally 1958), Simon and Newell describe a chess game, which their program will attempt to simulate, by saying that "without a chance device to obscure the contest, it [the chess game] pits two intellects against each other in a situation so complex that neither can hope to understand it completely, but sufficiently amenable to analysis that each can hope to outthink his opponent. [...] If one could devise a successful chess machine, one would seem to have penetrated *the core of human intellectual endeavor* (Newell and Simon, 1963, p. 39, my emphasis.)

"artificial" any entity that can be described in terms of "functions, goals, and adaptations" to an environment. It turns out that one of the most important families of "artifacts," are what Simon calls "physical symbol systems"; an important member of this family, perhaps the most important one, "is the human mind and brain. It is with this family of artifacts, and particularly the human version of it—Simon states—that we will be primarily concerned in this book." (Simon, 1969, p. 27)

I think we need to take these well-known passages at face value to understand the nature of Simon's questioning in all of his books and articles, including his autobiography. He is not asking "What is problem solving?" and then answering: "An application of bounded rationality." Rather, he is asking "What is Man?" and he answers "Man is a problem solver." It is only in a second moment, as an answer to the question "How does problem solving actually work?", that the concept of bounded rationality enters into the fray. In other words, Simon is after an extremely general form of anthropology, which is not to be confused with what goes under this label in present-day departments carrying the label. Simon's scientific quest is not confined to the analysis of more or less culturally distant social groups. Insofar as his research queries all aspects of human activity, from artistic production (or rather, artistic reflection, see (Simon, 1994)) to political activity, from ethics to social belonging, from reasoning to perception and emotion, to the relationship between man and his environment, and so forth, it strives to provide a general form of anthropology as *science of Man in general*.

It is important to stress the scope of his inquiry. Simon finds the same mechanism at work in all the different contexts he looks at, because it is the same fundamental structure of bounded rationality that lies at the core of decision-making in organizations, at the core of problem-solving in chess, and so on. It follows that the general structure he describes transcends the single, specific field in which it is applied, and constitutes, instead, a description of the most general, systematic conditions of possibility of human cognition. Once considered in its entirety, Simon's work appears as an updated version of the transcendental inquiry inaugurated by Kant. Indeed, Simon's scope is even broader than Kant's.

In one of the first attempts to draw an explicit parallel between the transcendental inquiry and classic AI, Joëlle Proust (1987) calls attention to the Kantian themes that Simon had probably absorbed through Carnap, Frege and Hilbert. Proust is mainly concerned with what she calls "classic Artificial Intelligence," a term she applies almost exclusively to Simon's and Newell's work. Proust goes as far as drawing a structural analogy between Simon's and Newell's "Physical Symbol System" and Kant's transcendental subject. She

correctly remarks that Hubert Dreyfus and Daniel Dennett, while drawing attention to the Kantian connections of classic Al research, had missed the proper analytic level. Since his first attack on AI, Dreyfus emphasized that Kant's epistemology prescribes that the transcendental subject constitutes the object of possible experience through the application of rules to the raw data provided by sensible intuition. The same procedure is posited by classic AI, where every possible behavior is ultimately reduced to a composition of basic symbol-formation rules. This comparison targets the content of Kant's and Simon's epistemologies and, according to Proust, it fails to uncover the even stronger similarities that can be found when comparing the structure of their respective theories.

Dennett (1978) avoids Dreyfus's problem by aiming for a direct comparison between Kant's epistemology *as a whole* and AI research. Nonetheless, he fails to recognize the distinction between the empirical and transcendental levels in Kant, because he reduces the transcendental analysis to a more abstract form of psychological inquiry. For Dennett, psychology and philosophy represent the ends of a continuum that spans the range of all possible investigations on the mental, from the most "top down" inquiries carried out by the philosopher, to the most bottom-up analyses that the psychologist pursues. Dennett's characterization, in other words, misses the crucial Kantian distinction between *de jure* and *de facto* epistemological questions. The reduction of Kant's analysis of the conditions of possible experience to an abstract form of psychology leaves out the very possibility of a transcendental subject which would warrant the universality and validity of the subject's cognitions.[5]

Proust is right when she points out that a proper appreciation of the philosophical aspects of classic AI must consider the overall structure of the theory and must give pride of place to AI's claims about the universal validity of the structures it uncovers. My interpretation of Simon's work presents some similarities with Proust's, but it does differ from hers in at least two significant aspects. I agree that the analysis should target the structural level of the respective theories and be concerned with issues of universality and validity. However, while Proust compares classic AI with Kant's epistemology, I believe a more fruitful comparison may be drawn between Simon's and Kant's projects in their totality. As I argued above, Simon's work outside of AI (in economics, organization theory, etc.) is a substantial part of his overall "philosophical

[5] See for example, Kant's remark in the "Architectonic of Pure Reason": "How are we to regard empirical psychology, which has always claimed its place in metaphysics? [...] I answer that it belongs where the proper (empirical) doctrine of nature belongs, namely, by the side of *applied* philosophy, the *a priori* principles of which are contained in pure philosophy; it is therefore so far connected with applied philosophy, though not to be confounded with it. Empirical psychology is thus completely banished from the domain of metaphysics; it is indeed already excluded by the very idea of the latter science" (Kant, 1965, A848/B876, p. 664).

system," thus turning AI modeling into a crucial but limited methodological tool for the broader project. Whereas Kant pursues, at least in the first *Critique*, a general investigation of the most general conditions of possibility of *cognition*, Simon sets his sight on a general analysis of the most general conditions of possibility of *human life.* In Kantian terms, this may seems to constitute an epistemology plus an anthropology, insofar as it aims at explaining both the *formal* as well as the *concrete* conditions of possibility of knowledge. But Simon offers us a unified and consistent reconstruction of human behavior (artistic practices, social behavior, ethical imperatives, etc.) that cannot be fit into the scope of epistemology no matter how much we try to stretch the term. Rather, we can accommodate the whole spectrum of Simon's research into Kant's framework only if we consider the full scope of critical philosophy, as outlined in the first, second and third *Critique*. I suggest we use for Simon's project the label that Kant himself used for his own: namely, *metaphysics*.[6]

What sets Simon's effort definitely apart from Kant's, however, is the method he used. For Kant, a transcendental inquiry on the general conditions of possibility of judgments can only be pursued a priori, and the universal validity of the judgments themselves must be deduced, or justified, through an analysis of their general mode of functioning. On the contrary, Simon always defined his research, in whichever fields he happened to be pursuing it at the moment, as strictly empirical and *a posteriori*. If we join Simon's methodological

[6] It might be objected that the parallel between Simon's project and Kant's transcendental inquiry I am suggesting here does not allow a transition from epistemology to metaphysics, because all Kant was after was but a grounding of the results provided by natural science. This interpretation of Kant's philosophy can be traced back to the Marburg school of neo-Kantism inaugurated by Hermann Cohen and which, under Natorp and Cassirer, came to dominate the field of Kantian studies in the first half of the 20th century. Rudolf Carnap was heavily influenced by it, and it is quite likely that Simon himself, having absorbed his Kant mostly through Carnap at the University of Chicago, would have subscribed to the epistemological interpretation of Kant. However, the Kantian project as it is stated, among other places, in the "Introduction" to the first *Critique* was much broader. It aimed at a reconstruction of metaphysics from the ground up that would provide the fundamental principles of scientific, ethical, and political knowledge that would ground human behavior. The goal of Kant's analysis of science was to draw the boundaries of scientific knowledge in view of the purely philosophical foundation of ethical behavior. The neo-Kantian interpretation was a self-conscious attempt to circumscribe the scope of critical philosophy to epistemology, thus getting rid of the "thing in itself," its most embarrassing, most metaphysical concept. What is at stake here, however, is not the adequacy of Simon's interpretation of Kant, or its probable debt to the Marburg school of neo-Kantism. Rather, it is the equivalence between Simon's lifelong project as a whole- in spite of his own self-interpretation - and the similar scope and stated goals of critical philosophy. My suggestion, to put it in slightly different form, is that the scope of Simon's project is *at least* equivalent to Kant's as propounded in the introduction to the *Critique of pure Reason*. As Kant did not hesitate to use the word metaphysics to name what he was after, neither should we. For a thorough interpretation of Kant's along the non neo-Kantian line we suggest see, among many examples (Lebrun, 1970).

declarations to our thesis about the scope of his research, we come to the conclusion that he was trying to develop a universally valid metaphysics *a posteriori.* In other words, Simon's work, considered in its entirety would qualify as a form of true "anti-philosophy," because it would have the same "content" of that old-fashioned, traditional discipline (same scope and goals), but a radically opposed method (empirical vs. *a priori*).

But is it really possible to construct a universally valid metaphysics *a posteriori*? My suggestion is that Simon saw in AI modeling, the well-to-do discipline he invented in the early 1950s, precisely the tool that would solve this paradox. This is the reason why the vast majority of Simon's scientific contributions, from the mid-1950s on, took the form of reflections, analysis, and elaborations about computer programs while, at the same time, the horizon of his research kept expanding beyond the horizon of cognitive psychology. This is why most of Simon's scholars see a decisive shift happening in the 1950s and mistakenly interpret it as a turn from economics to psychology, whereas it should be read, I believe, as the transformation of a specific, limited, and narrowly circumscribed research program in economics and organization theory into a universal inquiry into the general conditions of possibility of human life which AI modeling made possible.

4. Simon's program and Artificial Intelligence's challenge

The reassessment of Simon's research program I have proposed has several important consequences for our understanding of Artificial Intelligence as a whole, which I will here simply advance as questions opening toward further research.

Methodology first. Artificial Intelligence, at least in its Simonian incarnation, represents an alternative to the binary opposition between empirical research and *a priori* analysis. This does not mean that AI modeling is a compromise between the two or some sort of Hegelian synthesis. On the contrary, AI research is not empirical, because it addresses the most general issues about human existence and is not concerned with testing hypotheses (see Marshall, 1978, and the remarks in Dennett, 1978, and Dennett, 1988). Nor is AI research *a priori* in all the various meanings that this term has taken during the course of Western philosophy: it is neither a pure analysis of concepts nor the phenomenological analysis of the first-person experiences of a subject understood as a "meaning giver." The novelty of Artificial Intelligence, and, by implication, the true meaning of the "computational turn in philosophy" that has often been associated with it, must be found in its synthetic methodology. In other words, the profound

methodological innovation of AI derives from its association of engineering techniques with age-old philosophical questions. AI researchers do not deduce concepts, nor do they test or disprove theories; rather, they *build* systems that reproduce behavior which instantiates a theory.

The "computational turn" is the aggressive transformation into a research program of the approach I attributed to Simon: computation, a "synthetic" method, becomes the new tool to do philosophy, because it preserves philosophy's goals while rescuing it from its infamous metaphysical impasses.

Let me move to the institutional and disciplinary level. This interpretation of the computational turn allows us to frame differently the endemic conflict existing between philosophy and AI. Philip Agre provides a vivid illustration of how deep the anti-philosophical feelings run (and especially ran) in the AI community. He remembers how, in his AI graduate student's days, his fellow students "would convene impromptu two minutes hate sessions to compare notes on the futility and arrogance of philosophy. 'They have had two thousand years and look what they've accomplished. Now it's our turn'." [7]

The unbridgeable distance was remarked over and over again, and was often expressed, as Agre remarks, in a difference between "doing" and "just talking" (Agre 2005). This protracted hostility is a direct consequence of AI's "anti-philosophical" position— and it could not have been otherwise, at least for those willing to subscribe to the premises of the Simonian approach. To the scholar studying the development of 20th Century thought, however, the conflictual aspect of Simon's AI program is important, because it brings to the forefront unsuspected affinities with other contemporary trends in Western culture which are often considered thoroughly alien to it. For instance, the French Structuralist movement inaugurated in anthropology by Claude Lévi-Strauss in the 1950s, and later exported to most other social sciences, shared a similar "anti-philosophical" bent. I believe that the structural similarity results from a similar theoretical move. The interesting question, however, although I must leave it unaddressed here, is whether such similarity in method translates into a similarity of doctrine or, possibly, whether the former is a consequence of the latter.

To conclude: the realignment I am advocating here allows for the reinsertion of Simon's research program into a broader cultural and philosophical context. In more specific terms, my interpretation affects the evaluation of Simon's program. The mechanical implementation of theories of human natures into computer programs becomes, in a certain sense, of secondary importance, because their "mechanizability" has a purely methodological character: it allows the theorist to do metaphysics in a novel manner. In my opinion, the almost

[7] The long-standing hostility of AI researchers against philosophy is a well-documented phenomenon, see (Agre, 1997) and (Franchi, 2005). (Simon, 1995) is perhaps the best defense of AI as empirical science.

exclusive attention upon the "methodological" aspect of AI has excessively narrowed the scope of most discussions of the discipline. Moreover, the unification under the computational banner of research programs as different as, say, Simon's (and Newell's), McCarthy's, Minsky's, etc., has obscured a number of important issues. On the contrary, I believe the gaze of the analyst should be directed upon the concrete theoretical structures produced by the individual theorists and "verified" via their software implementations. In other words, the questions I would like to ask Herbert Simon, the anti-philosophic metaphysician, and his heirs are: what is the content of his metaphysics? What does it mean to see humans as problem-solving beings, and to practice problem-solving in terms of heuristic search? Which kind of being is the being whose interaction with the world is mediated by game-like structures like those deployed by Simon?

References

Agre, P., 1997, *Computation and human experience*, Cambridge University Press, Cambridge.

Agre, P., 2005, "The soul gained and lost," in: S. Franchi and G. Güzeldere,, eds., *Mechanical Bodies, Computational Minds*, MIT Press, Cambridge, MA, 2005, , pp. 153-173.

Augier, M. and March, J., 2002, "A model scholar," *Journal of Economic Behavior and Organization* 49**:**1–17.

Augier, M., 2000, "Models of Herbert Simon," *Perspectives on Science* 8(4):407–443.

Boden, M., ed., 1990, *The Philosophy of Artificial Intelligence*, Oxford University Press, Oxford.

Brooks, R., 1999, *Cambrian Intelligence: the Early History of the New AI*, MIT Press, Cambridge, MA.

Clark, A., 2001, *Mindware: an Introduction to the Philosophy of Cognitive Science*, Blackwell, Oxford.

Copeland, J., 1993, *Artificial Intelligence: a Philosophical Introduction*, Blackwell, Oxford

Dennett, D., 1978, "Artificial Intelligence as Philosophy and as Psychology," in Daniel Dennett, *Brainstorms*, MIT Press, Cambridge, MA, pp. 109-126.

Dennett, D., 1988, *When philosophers encounter artificial intelligence*, *Daedalus* 117(1):283–295.

Dreyfus, H., 1972, *What Computers Can't Do*, Harper and Row, New York.

Feigenbaum, E., 2001, "Herbert A. Simon," *Science* 291:2107.

Franchi, S. and Güzeldere, G., 2005, *Mechanical Bodies, Computational Minds*, MIT Press, Cambridge, MA.

Franchi, S., 2005, "Hunters, cooks, and nooks," *Diacritics,* 33 (2): 98-109.

Guice, J., 1998, "Controversy and the state: Lord Arpa and intelligent computing," *Social Studies of Science* 28(1):103–138.

Kant, I., 1965, *Critique of Pure Reason*, St. Martin's Press, New York.

Lebrun, G., 1970, *Kant et la fin de la métaphysique*, Colin, Paris.

Marshall, J., 1988, "Close enough for AI?", *Journal of Semantics* 5:169-173.

Mazlish, B., 1993, *The Fourth Discontinuity*, Yale University Press, New Haven, CO.

McCorduck, P., 1979, *Machines Who Think: a Personal Inquiry into the History and Prospects of Artificial Intelligence*, W.H. Freeman, San Francisco.

Minsky, M., 1963, *Steps towards artificial intelligence*, in: E. Feigenbaum and J. Feldman, eds., *Computers and Thought*, McGraw-Hill, New York, pp. 406–450.

Minsky, M., ed., 1968, *Semantic Information Processing*, MIT Press, Cambridge, MA.

Mirowsky, P., 1986, *Machine Dreams*, Cambridge University Press, Cambridge.

McClelland J. and Rumelhart, D., eds., 1986, *Parallel Distributed Processing: Exploring the Microstructure of Cognition*, MIT Press, Cambridge, MA.

Newell, A. and Simon, H., 1963, "Chess playing programs and the problem of complexity," in: E. Feigenbaum and J. Feldman, eds., *Computers and Thought*, McGraw-Hill, New York, 39-70.

Proust, J., 1987, "L'intelligence artificielle comme philosophie," *Le Débat* 47:88–102.

Sent, E.-M., 2000, "Herbert A. Simon as a cyborg scientist," *Perspectives on Science* 8(4):380–406.

Simon, H., 1947, *Administrative Behavior*, MacMillan, New York.

Simon, H., 1957, *Models of Man*, Wiley and Sons, New York.

Simon, H., 1969, *The Sciences of the Artificial*, MIT Press, Cambridge, MA.

Simon, H., 1975, *Administrative Behavior, 3rd edition*, Free Press, New York.

Simon, H., 1979, *Models of Thought*, Yale University Press, New Haven, CO.

Simon, H., 1994, "Literary criticism: a cognitive approach," in: S. Franchi and G. Güzeldere, eds., *Bridging the Gap*, vol. 4, *Stanford Humanities Review*, Special Supplement, pp. 1–26.

Simon, H., 1991, *Models of My Life*, Basic Books, New York.

Simon, H., 1995, "Artificial intelligence: an empirical science," *Artificial Intelligence* 77:95–127.

Simon, H., 2001, "On simulating Simon, his monomania, and its sources in bounded rationality," *Studies in the History and Philosophy of Science* 32**:**501–505.