PROLOGUE

Perennial Dreams

Cognition is a complex, multifaceted, and multilevel phenomenon. Unraveling it, therefore, is beyond the scope of any single discipline, the capacity of any one method, or the resources of any individual philosophy. Rather, it can emerge from the exchanges and interactions among multiple ideas, methods, models, and philosophies. In our times, cognitive science has nominally taken on the challenge to bring these strands together but, as any cognitive scientist would be willing to acknowledge, the challenge cannot be effectively met without the serious and engaged contribution of neighboring disciplines. Artificial Intelligence (AI) is one such neighbor to cognitive science, but it seems that in recent years the exchange between them has not been as active and extensive as it could be. One reason for this, I believe, is the history of developments in AI and the overall image that this history has created of the field in the scientific community. The main purpose of the current study is to examine and explain this history, not to merely criticize AI, but also to highlight its contributions to science, in general, and to cognitive science, in particular. Interestingly, these contributions are both direct and indirect. That is, there are lessons to be learned from both the content of AI – its approaches, models, and techniques – as well as from its development. We can learn *from* AI by learning *about* it.

To this end, the current study aims to put AI in context and perspective or, more specifically, to explicate the views, ideas, and assumptions it has inherited from the intellectual past (most notably, the Western rationalist tradition), and to discuss the cultural and institutional milieu of AI practices that support and promote those views. Far from a shared understanding, we shall see, these multiple origins have given rise to an enterprise the goal and character of which is constantly contended, negotiated, and redefined. In the end, the picture that will emerge of AI is a colorful fabric of interwoven

cultural, institutional, and intellectual threads that can be best understood through the relationships among its components.

AI is a dream – a dream of the creation of things in the human image, a dream that reveals the human quest for immortality and the human desire for self-reflection. The origins of AI, in one form or another, go back to the perennial pursuit of human beings to understand their place in the overall scheme of things, both as creations and as creators. In primitive societies, this quest gave rise to complex rituals, intricate religious beliefs, enchanted objects of worship, and nascent forms of philosophizing. In religious thought, it manifests itself in many guises – from resurrection and reincarnation to mysticism and other belief systems about the unity of the universe. In early modern times, the quest took the form of a question about the "nature" of animals and humans as *machines* (Mazlish 1993: 14). Today's AI is the inheritor and a major perpetuator of this perennial quest. It embodies, in the most visible shape, the modernist dream of a purified world: of a mind detached from the body and the world, of a nature detached from the self and society, and of a science detached from power and politics. As such, AI brings the intrinsic tensions of modernity – for example, the tensions between mind and body, nature and culture, and science and religion (Haraway 1991; Latour 1993) – into sharp relief. It may fail or it may succeed, but the dream will most likely stay with us in the foreseeable future, motivating similar intellectual endeavors aimed at understanding our place in the grand scheme of things. I would therefore like to think of AI as the embodiment of adream – a kind of dream that stimulates inquiry, drives action, and invites commitment, not necessarily an illusion or mere fantasy.[1]

In this introductory part of the study, my goal is to provide a brief overview of the development of AI, to highlight the common patterns of thought, attitudes, and actions observed within the AI community, and to present an initial interpretation of their meaning. In order to place the present work within the broader context of previous studies, I will examine how critics, from within and without, have reacted to this course of development. Finally, I will then try to show how the approach adopted here can contribute to the improvement of the field and its image in the scientific community.

MAINSTREAM AI IN A SNAPSHOT

AI has a short but intriguing history. Since its inception half a century ago, the discipline has had many periods of unbridled optimism followed by periods permeated by the spirit of failure. As Crevier (1993) aptly put it, this is indeed a "tumultuous history." One of the general patterns that emerges

Table P.1. Salient examples of the disruptive view of change in the history of AI.

The Approach	"Is nothing but"	"Has nothing to do with"
Physical Symbol System	Token manipulation	Body and environment
Supercomputing	Speed of Computation	Human society
Cybernetic	Control and Software	Body
Knowledge-intensive	Knowledge	Control structure
Case-based	Reminiscence	Logic
Connectionist	Neuron-like computation	Symbols and rules
Dynamical	Continuous state change	Discrete computation/ representation
Neo-robotic	Physical interaction	Representation

from the current study, for instance, is that every so often a new paradigm appears on the scene with the explicit claim of overthrowing the previous ones, and of replacing them with fresh new ideas, methods, and models that will overcome or bypass the previous problems altogether. Although this kind of punctuated development is not unfamiliar in the history of science,[2] it takes an unusually "revolutionary" character in the history of AI, in the form of all-or-nothing statements about the essence of intelligence, cognition, and mind. Most commonly, such statements take one or both of the following forms:

- Intelligence is nothing but *X*
- Intelligence has nothing to do with *Y*

where *X* and *Y* vary depending, respectively, on the specific approach that is being advocated and on the approach(es) against which it has been launched. Table P.1 provides examples of the *X*s and *Y*s that we will see throughout the coming chapters, representing the disruptive view of change prevalent within the AI community.

What turns out in each approach, however, is that none of these black-and-white dichotomies are borne out in the real world. Consequently, almost all of these so-called revolutions, after being so starkly advocated, gradually dissolve into a ripple, turn into limited reforms, and are ultimately absorbed in "mainstream AI," regenerating and reinforcing some of the old assumptions, techniques, and tensions, albeit under new guises and with new labels. As a result, the most outstanding feature of mainstream AI is that it has diverged enormously from the original vision of understanding human cognition (Sloman 2005a).

Despite the initial fervor and the revolutionary claims, therefore, the overall picture that emerges of AI is much like that of the social anarchism of

late-nineteenth- and early-twentieth-century Europe. Anarchists advocated a social system without hierarchy and governance. A major tenet of their ideology was, therefore, the elimination of power, oftentimes by resorting to force and social "shocks" such as assassination. One of these shocks, the terror of the Hapsburg Crown Prince, actually triggered World War I. Although these shocks had transient political impacts, the activities of anarchists could not do much in changing the social status quo, and anarchists themselves were gradually divided, with the majority of them joining one of the more traditional poles of the political spectrum. This is what typically happens to all the "new" approaches in AI: they begin with claims that have the disruptive character of a shock, only to gradually merge into the existing poles with the passage of time. What is the explanation for this state of affairs?

SCIENCE AND ENGINEERING IN TENSION

> The moment of truth is a running program.
>
> – Herbert Simon (1995)

The intensity of the debates in AI reflects an internal tension within the field that derives from its scientific and engineering practices. Following a three-step scheme established by Turing in his design of universal machines (see Chapter 1), Artificial Intelligence seeks to do three things at the same time:

1. as an engineering practice, AI seeks to build precise working systems.
2. as a scientific practice, it seeks to explain the human mind and human behavior.
3. as a discursive practice, it seeks to use psychological terms (derived from its scientific practice) to describe what its artifacts (built through the engineering practice) do.

The first two practices highlight the fact that AI, as a way of *knowing* and a way of *doing*, straddles the boundary between science and engineering. On the one hand, its object is typically the building of artifacts that perform specific tasks. On the other hand, those same artifacts are intended as a medium to model and explain an aspect of human behavior. These dual objectives have been present in AI ever since its inception. Herbert Simon, one of the pioneers of AI argued, "far from striving to separate science and engineering, we need not distinguish them at all. . . . We can stop debating whether AI is science or engineering; it is both" (1995: 100). The engineering principle of "understanding by building" is a strong driving force in AI, and accounts for a good deal of its appeal among its advocates. In the modernist eye, the

fact that AI theories and models of cognition are physically instantiated in computer systems provides them with an aura of being "real" – a quality that seemed to be missing from previous psychological theories. In fact, the role of technology is what distinguishes AI from any previous attempt to study human behavior (Varela, Thompson, and Rosch 1991: 5). It is what, according to some accounts, generates within AI (and cognitive science) a potential for rigor, clarity, and controlled experimentation (Crevier 1993: 247), as well as a certain degree of "honesty," which is acknowledged even by the staunchest of AI's critics (Dreyfus 1992). However, as I intend to show, engineering and scientific principles often clash with each other, generating a gap that needs to be filled by the third practice.

This third practice, which acts like a bridge, is more subjective than the other two. The term "discursive practice" refers to a crucial activity of AI practitioners – namely, their pervasive use of a discourse that connects the performances of their artifacts to their scientific claims.[3] In other words, what makes AI distinct from other disciplines is that its practitioners "translate" terms and concepts from one domain into another in a systematic way.[4] Thus, on the one hand, AI practitioners talk about their artifacts using a vocabulary that is canonically used to describe human behavior – in short, they anthropomorphize machines. On the other hand, they think of human beings as (complex) machines. Therefore, it is common to hear engineering terms applied to human behavior by some AI people. Charles Babbage, widely considered as a pioneer of modern computers, was aware of the power and appeal of this cross-domain translation. In 1838, he conceded that "in substituting mechanism for the performance of operations hitherto executed by intellectual labour [. . .] the analogy between these acts and the operations of mind almost forced upon me the figurative employment of the same terms. They were found at once convenient and expressive, and I prefer to continue their use" (in Schaffer 1994: 208).

The point is not that there is something inherently wrong or exceptionally ad hoc in this practice – for that is essentially what analogies do, and analogies may well constitute the core of human cognition (Hofstadter 2001). Nor is it even that AI practitioners are totally unaware of these translations. The problem is that this use of metaphorical language is often done without due attention to the vague and imprecise nature of the cross-domain allusions (Agre 1997a: 45–46).[5] The danger of this inattention is that it propagates a noncritical attitude, which can interfere with technical (scientific and engineering) practice (Varela et al. 1991: 133). In sum, my claim is that AI seeks to accomplish three things at the same time: a way of doing, a way of knowing, and a way of talking – and I will henceforth call these the *three practices of AI*.

These practices pull the field in different directions, generating a body of models, metaphors, and techniques that, judged by AI's history, keeps moving in cycles of fads and fashions (Ekbia 2001). This happens because the discursive practice simply cannot fulfill the role expected of it: not only does it fail to bridge the gap between engineering and scientific practices, it widens the gap by introducing more and more layers of anthropomorphic language, metaphorical imagery, and technical terminology. Indeed, the history of AI includes a number of "revolutions" that were launched against a governing paradigm (logic, computation, representations, etc.), but that in the end only reinforced some of that paradigm's fundamental assumptions in a more disguised form. Various intellectual, institutional, cultural, and socio-economic threads have heavily contributed to the entrenchment of this situation, as I will show in the coming chapters.

Furthermore, the interdisciplinary character of AI complicates the situation even further.[6] People coming from different disciplines bring with them various intuitions, assumptions, and widely disparate understandings of the same concepts and practices. Differences can begin even at the most basic level – for example, the characterization of "intelligence," which is the most fundamental notion in the field – and they then extend to other, more complex theoretical notions such as "symbol," "syntax," "semantics," "representation," and so on. The result, as we shall see, is pervasive crosstalk, unjustified redundancy displayed in reinvented wheels, mutually estranged research centers, rival research projects that work on the same topics, and, most pervasively, muddled and misleading claims.

The thesis advanced here is, therefore, that the pressures of multiple practices, together with the intellectual, cultural, and institutional threads mentioned earlier, generate great tension inside the field of AI. Engineering practice, for instance, is performance-based – what matters is the working of a system based on physical specifications. Human behavior, however, follows a different "logic" – it doesn't merely function physically; it also seeks truths, follows rules, pursues goals, wills good, and seeks beauty. A theory that attempts to explain or model human behavior will therefore have far more complicated criteria for success than engineering does. The main challenge for AI is, therefore, to how to compromise among these competing logics.

A CRITICAL STANCE

> It seems to me that the historical analysis of scientific discourse should, in the last resort, be subject, not to a theory of the knowing subject, but rather to a theory of discursive practice.
>
> – Michel Foucault, *Order of Things*

This work is a critique of AI, and a critique can take different shapes and forms. In art or literature, for instance, criticism usually means "critical appreciation," and a critic is someone with the skills and sensibility to understand a piece of art or literature as well as the process of its production. This kind of critique may not lead to unconditional praise, but it is not usually taken as a debunking of either art or literature (Biagioli 1999: xiii). A different kind of critique is that conducted by the practitioners of science studies in their analyses of science. Although not typically described as "criticism" by the practitioners, these analyses are usually seen by scientists as negative attacks on their enterprise as a whole. Biagioli attributes this kind of reaction to "the structure, scale, and social robustness of the scientific community that makes some of its members read local critiques as general ones" (ibid.). Although sympathetic to this assessment, I believe that an effective way of addressing it is to adopt a more engaged and "informed" approach to the subject matter. As will be seen throughout this study, I have applied many ideas and insights from science studies here, but I sometimes find them inadequate because of their distance from formal and technical issues that are at the heart of AI. I have, therefore, conducted this work from a perspective that involves both the formal-technical and conceptual-critical aspects of AI. Following Agre (1997), I call this approach a technical-critical practice.

Technical-Critical Practice

Technical work of the kind pursued in AI often involves problems that are incrementally solved to improve performance. Having bugs, glitches, and problems, in other words, is an inherent part of technical practice, and should not be taken as an indication of failure. One cannot get into the water, as an old Persian adage says, without getting one's feet wet. What matters in assessing technical work, therefore, is the way the work confronts and solves such bugs, glitches, and problems. A project can do this by paying explicit attention to the underlying assumptions and biases that have generated a problem, or by sticking to those assumptions and trying to find quick and accessible "fixes"; it can be patient, starting from small and well-understood domains and slowly proceeding to more sophisticated ideas, as is normally the case in science, or it can impatiently leap from raw intuitions to complicated systems; it can remain committed to well-laid-out principles (of design, performance, or psychological realism) throughout its development, or it can violate those principles, if indeed it has any, in successive stages.

From this perspective, the importance of a project in AI does not so much depend on its size, scope, or sophistication as it does on the match

between what it does, what it signifies, and what it claims to do and signify. Thus, a system that seems trivial in its ambitions may be more promising, by this measure, than a grand project with hidden disparities between word and deed. Fantasy and hype are found not only in a mismatch between claims and accomplishments, or in the emanation of unfounded predictions and undeliverable promises for the future, but also in the disparity between multiple practices (building, theorizing, and talking) and in the usage of a great deal of psychological language that often tends to make the gaps seem smaller than they are (McDermott 1976). This perspective also clarifies my notion of "dream," which plays a central role in this writing: by "dream" I do not mean to suggest illusion or mere fantasy, but something that motivates inquiry, drives action, and invites commitment.

Attribution Fallacy and Generalized Eliza Effect

My purpose, therefore, is not only to criticize the actual *programs*, but the misleading effect of the communication techniques used to talk about the programs – their discourses, to use one of the key notions of this writing. As we will see throughout this survey, there is a tendency in AI to exploit the *Eliza effect* (see later in this chapter) by smoothly conflating the real-world events being modeled with the tiny stripped-down versions that are in the models. When one reads reports, articles, and books written about various AI programs, it is often next to impossible to realize how stripped-down the world of the program really is, because of this insidious phenomenon that researchers and authors rarely try to discourage. This constant sliding back and forth between the reader's image of a real event and pieces of code that use variables whose names suggest a great deal of visual richness has the inevitable effect of contaminating an uncritical (or even a critical) reader's mind with a great deal of the imagery associated with the real event. Known in AI as the *Eliza effect*, this is a very common phenomenon that is manifested in diverse shapes and forms, many examples of which we will see in following chapters – for example, a visit to the zoo, a football game, a cooking session, a learning process, an eyeball-to-eyeball encounter between individuals, seeing a computer-generated painting, or listening to a computer-composed Chopin-style mazurka. The diversity of these manifestations points to a very broad phenomenon that I call the "Generalized Eliza Effect" (GEE). In short, GEE has to do with the often-exaggerated abilities of AI artifacts to delude the casual observer.

The flip side of this phenomenon is what I call the "Attribution Fallacy": the propensity of people to uncritically accept implicit suggestions that some

AI program or other is dealing with real-world situations. When a program supposedly explains why a shuttle launch failed, people take for granted that the program has an image or model of the situation roughly comparable to their own. The truth may be extremely remote from that – but few AI authors include in their articles a disclaimer that states: "Warning: Entities in models appear to be much more complex than they really are." And unfortunately, by not doing so, they implicitly encourage their readers to let their *own* concepts slide and glide fluidly back and forth between the real world and the model, so that in the end no clear notion is built up about how microscopic the worlds being dealt with really are.

The blame for this deception falls primarily on the researchers and authors who do nothing to try to stop the Eliza effect dead in its tracks. They may not want to do so, partly because they have too much invested in their projects, partly because, being human, they themselves fall to a certain degree for the same Eliza effect, and partly because, as the previous discussion suggests, they are under opposing pressures due to the scientific and engineering aspirations of AI. These pressures sometimes even result in a conflict of objectives within a single approach, as we will see later.

In brief, this book pursues a number of different goals – namely, to criticize AI approaches and systems in terms of the gap between their scientific claims and their engineering achievements, to expose the discourses that are used in covering the above gap, and to understand and explain the related phenomena of the Generalized Eliza Effect and the Attribution Fallacy. The last point obviously has more to do with human beings than with computers and AI systems, but that does not diminish its relevance to our purposes. As a matter of fact, to highlight the significance of this point, I sometimes have chosen to examine and critique AI systems that might otherwise be old, minor, or insignificant (even in the eyes of their originators). The advantage of such systems for our purposes here, however, is that they reveal the Eliza effect in its full scope and strength. Ironically, being "inferior" examples of AI, such systems can easily illustrate our human susceptibility to the Eliza effect, in ways that "superior" examples cannot. This, I believe, is one of the major "indirect" lessons of AI for cognitive science.

At any rate, my purpose is not to advocate certain AI approaches while debunking others. It is, rather, to contribute to a better understanding of the field of AI, to contribute to its progress and improvement, and, in so doing, to draw lessons for cognitive science. This constructive approach is inspired by, and, I believe, is in line with previous critiques that pursued a similar goal – most notably, McDermott (1976, 1987), Hofstadter (1985, 1995), Winograd and Flores (1986), Smith (1991, 1996, forthcoming), and

Agre (1997a,b, 2002) from within AI, and Dreyfus (1972, 1992), Suchman (1987, 2007) Collins (1990), Edwards (1996), Forsythe (2001), and Woolgar (1985, 1995) from outside the field.

THE CRITICAL SPECTRUM

Because of the character of its subject matter, research in AI could well come to have far-reaching implications for other disciplines such as anthropology, biology, philosophy, psychology, and sociology. Small wonder, then, that AI has been the subject of scrutiny and criticism since its early days. Such criticisms have come from a variety of quarters, with multifarious agendas and criteria, but one can broadly classify them as philosophical, social, or (as in the present case) reflexive.

Philosophical critiques fall across a wide spectrum – ranging from total embrace (e.g., Boden 1990, Dennett 1991) to somewhat supportive (Churchland 1989, Clark 1997, Dretske 1994, Haugeland 1981, 1985) to sharply critical (Dreyfus 1972, 1992, Putnam 1975, 1990) to totally dismissive (Searle 1984, 1992). The most controversial views have been those of Dreyfus and Searle. Dreyfus originally criticized AI for its failure to capture the essence of human experience, which he characterized as a set of skills for getting around in the world. This view was in sharp contrast with the common view in traditional AI that intelligence consists, by and large, in the manipulation of symbols by an isolated mind, and that it has little to do with bodily skills and with coping with the environment. Over the course of time, there has been some convergence between Dreyfus's views and those of AI community. Many AI researchers have come to agree with his point of view, and Dreyfus has shown a willingness to embrace parts of AI – for instance, connectionism. Searle, by contrast, has persistently dismissed a major strand of AI on the grounds that AI systems intrinsically lack "intentionality" (i.e., they are unconscious and the signs they manipulate are devoid of all meaning; see Chapter 2).

The overall response of the AI community to these critiques has usually been one of incomprehension or outrage but also, on occasion, cautious reflection. I would contend that these critiques have contributed to the vigor and disciplinary awareness of AI, although in rather different ways. The same could not be said about sociological critiques, which have been mostly disregarded, if at all recognized, by the AI community. This disregard doesn't mean, however, that such critiques have had no valuable points to contribute.

Social critiques of AI originate in various disciplines. Some anthropologists have tried to describe AI as a culture that projects its own values onto its systems Suchman (1987, 2007) has famously shown how the classical notion