Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI

Philip E. Agre
Department of Information Studies
University of California, Los Angeles
Los Angeles, California 90095-1520

pagre@ucla.edu
http://polaris.gseis.ucla.edu/pagre/

This is a chapter in Geof Bowker, Les Gasser, Leigh Star, and Bill Turner, eds, *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, Erlbaum, 1997.

Please do not quote from this version, which differs slightly from the version that appears in print.

11600 words.

1 Introduction

Every technology fits, in its own unique way, into a far-flung network of different sites of social practice. Some technologies are employed in a specific site, and in those cases we often feel that we can warrant clear cause-and-effect stories about the transformations that have accompanied them, either in that site or others. Other technologies are so ubiquitous -- found contributing to the evolution of the activities and relationships of so many distinct sites of practice -- that we have no idea how to begin reckoning their effects upon society, assuming that such a global notion of "effects" even makes sense.

Computers fall in this latter category of ubiquitous technologies. In fact, from an analytical standpoint, computers are worse than that. Computers are representational artifacts, and the people who design them often start by constructing representations of the activities that are found in the sites where they will be used. This is the purpose of systems analysis, for example, and of the systematic mapping of conceptual entities and relationships in the early stages of database design. A computer, then, does not simply have an instrumental use in a given site of practice; the computer is frequently *about* that site in its very design. In this sense computing has been constituted as a kind of imperialism; it aims to reinvent virtually every other site of practice in its own image.

As a result, the institutional relationships between the computer world and the rest of the world can be tremendously complicated -- much more complicated than the relationships between the telephone world and telephone subscribers, or between the electric lighting world and the people who use electric lights in their workplaces and homes. The residents of these borderlands are many and varied, and increasingly so. They include the people who work on the border between the computer world and the medical world, whether because they conduct research in medical informatics or because they must encode their patient interactions for entry into an hospital's automated record keeping system. They likewise include the photographers whose livelihood is rapidly moving into digital media, the engineers who must employ computer-based tools for design rationale capture, and the social scientists who study the place of computers in society. Each of the borderlands is a complicated place; everyone who resides in them is, at different times, both an object and an agent of technical representation, both a novice and an expert. Practitioners of participatory design (Greenbaum and Kyng 1990) and requirements engineering (Jirotka and Goguen 1994), among other disciplines, have done a great deal to explore and transform them. Above all, every resident of them is a translator between languages and worldviews: the formalisms of computing and the craft culture of the "application domain".

Every resident of the borderlands has a story, and in this chapter I would like to draw some lessons from my own. In 1988 I received a PhD in computer science at MIT, having conducted my dissertation research at the Artificial Intelligence Laboratory. I started out in school studying mathematics; I moved into computing because it helped me pay my school bills and because AI appealed to my adolescent sensibilities; I moved out of computing because I felt I had said everything I had to say through the medium of computer programs; and now I am a social scientist concerned with the social and political aspects of networking and computing. This path has its geographical aspects, of course, and its institutional aspects; at each transition I was able to construct myself as a certain sort of person, and I was usually able to stay employed. Here, though, I wish to focus primarily on the cognitive aspects of my path. My ability to move intellectually from AI to the social sciences -- that is, to stop thinking the way that AI people think, and to start thinking the way that social scientists think -- had a remarkably large and diverse set of historical conditions. AI has never had much of a reflexive critical practice, any more than any other technical field. Criticisms of the field, no matter how sophisticated and scholarly they might be, are certain to be met with the assertion that the author simply fails to understand a basic point. And so, even though I was convinced that the field was misguided and stuck, it took tremendous effort and good fortune to understand how and why. Along the way I spent several years attempting to reform the field by providing it with the critical methods it needed -- a critical technical practice.

In writing a personal narrative, I am assuming some risks. Few narratives of emergence from a technical worldview have been written; perhaps the best is Mike Hales' (1980) remarkable book *Living Thinkwork* about his time as a manufacturing engineer using operations research to design work processes for chemical production workers. A sociological inquiry is normally expected to have an explicit methodology. The very notion of methodology, however, supposes that the investigator started out with a clear critical consciousness and purpose, and the whole point of this chapter is that my own consciousness and purpose took form through a slow, painful, institutionally located, and historically specific process.

A personal narrative is also open to misinterpretation. I do not wish to engage in public psychotherapy; my emotional investments in AI and its research community are illuminating in their own way, but here I simply wish to recount an intellectual passage. I am not interested in portraying myself as an victim of circumstance or an innocent party in a conflict. I am not going to confess my sins, numerous though they have been, or seek absolution for them. Nor, as Patrick Sobalvarro usefully suggested in response to an early draft of this chapter, would I wish to portray myself as Jesus among the Pharisees -- the virtuous hero who uncovers the corruption of traditional learning and yet fails to persuade the learned of their errors. Mine is not a tale of virtuous heroism, heaven knows, simply of the historical conditions of a path. Perhaps my tale will contribute to the emergence of a critical technical practice, but only if it is taken as a counsel of humility.

A final risk is that I may seem to condemn AI people as conspirators or fools. AI has a long history of conflict with critics, to whom it has often responded harshly. Although these responses may reflect the aggressive styles of particular personalities, they may also result from a lack of access to forms of historical explanation that interpretive social scientists and philosophers take for granted. Without the idea that ideologies and social structures can be reproduced through a myriad of unconscious mechanisms such as linguistic forms and bodily habits, all critical analysis may seem like accusations of conscious malfeasance. Even sociological descriptions that seem perfectly neutral to their authors can seem like personal insults to their subjects if they presuppose forms of social order that exist below the level of conscious strategy and choice.

The first few sections of the chapter will be concerned with AI as a field. Section 2 will recount some salient aspects of the field's institutional and intellectual history. Section 3 will sketch how the field understands itself; it is crucial to comprehend these self-understandings in order to work critically because they will shape the mainstream practitioners' perceptions of proposed alternatives. Section 4 will describe some of the discursive practices that have made AI so successful up to a point, while also making it difficult even to conceptualize alternatives.

The last few sections will describe my own experience and the lessons I have drawn from it. Section 5 will recount how I emerged from AI's unfortunately confining worldview and began to incorporate influences from

philosophy, literary theory, and anthropology into my technical work. Section 6 will discuss what it means in practice to develop "alternatives" to an existing technical practice; for the most part, this notion is misleading. Section 7 will conclude with my own theory of critical engagement with a technical field.

2 Historical constitution

The field of artificial intelligence arose in the years after World War II, when psychologists and others who had been pressed into wartime research returned to their labs. During the war, they had been greatly inspired by wartime technologies such as signal detection methods and tracking devices for guns. The attraction of these technologies was that they lent themselves to intentional description: a tracking device, for example, could be said to pursue goals and anticipate the future. As such, they provided the paradigm for a counterrevolution against behaviorism -- a way to make talk about mental processes scientific and precise.

Wartime research also created an important set of social networks among the military and civilian research communities. MIT in particular came to prominence as a technical university that had made significant innovations in the war effort, and after the war it found itself closely connected to the suddenly much larger government research establishment. With the transition to the Cold War, this epistemic community united around a scientific and technical vision that Edwards (1995) has usefully identified as the "closed world" -- the whole world as one large technical system. Human beings, on this view, are themselves technical entities who serve as components of organizational systems: their bodies are machines and their minds are nodes in a hierarchical command-and-control network based on rational analysis and optimization. Edwards refers to the system of practices around this idea as "cyborg discourse".

As an institutional matter, then, AI was one part of an overall movement with a coherent worldview. It would be unfair to say that AI's founders were conforming themselves to military funding imperatives, just as it would be false to say that the emerging field of AI was intellectually autonomous. Numerous factors converged to reward the search for technologies and mathematical formalisms whose workings could be described using intentional vocabulary. Several technologies and formalisms received serious attention in this regard, including information theory, simulated neural networks, signal detection methods, cybernetic feedback systems, and computational complexity theory. As the field coalesced, though, only a handful of technologies and formalisms emerged to define the first decades of AI research. Most of these technologies and formalisms -- formal language theory, problem solving in search spaces, mathematical logic, and symbolic programming -- were qualitative in nature, not quantitative, and most of them were founded on the power and flexibility of the stored-program digital computer.

As the field of AI took form, it mixed elements of science and technology in complicated ways. Many of the founders of AI were psychologists, and they explained the field in terms of computer modeling of human thought processes. Others had a more abstract, less empirical interest in intelligence, and they explained the field with formulas such as, "building systems whose behavior would be considered intelligent if exhibited by a human being". Few of them regarded themselves as engineers seeking purely instrumental solutions to technical problems. Despite this, and notwithstanding the differences among individual investigators' approaches, the research programs at the new AI labs were deeply congruent with the broader military-scientific "closed world" consensus. As a result, the research in these labs continued for over twenty years with remarkably little detailed oversight or direction from ARPA and the other basic research funding agencies in the military. Broad research problematics such as problem-solving, learning, vision, and planning formed a kind of boundary between the individual researchers (and especially their students), who experienced themselves as having considerable autonomy, and their funding agencies, who had a ready vocabulary for explaining the relevance of this research to the agencies' overall technology strategy.

Critics of AI have often treated these well-funded early AI labs as servants of the military and its goals. But it should be recognized that these labs' relative prosperity and autonomy within a deeply shared worldview also provided the conditions for dissent. In my own case, the first five years of my graduate education were paid by a fellowship from the Fannie and John Hertz Foundation. John Hertz, best-known as the founder of the Hertz car rental company, was a conservative patriot who left his fortune to a foundation whose purpose was

to support military research. The Hertz Foundation fellowship program was, and is, administered largely by scientists at Lawrence Livermore National Laboratory who are associated with the military's nuclear weapons and ballistic missile defense programs. During the late 1970's, when I received my own Hertz Fellowship, the Hertz Foundation was aligned with the military's support for AI research. Numerous other graduate students in my laboratory were also Hertz Fellows, and the Foundation officers would speak explicitly about their hopes that the nation's research base -- whose health they understood in a broad sense, not simply in terms of immediate contributions to military research programs -- would benefit from the large investment they were making in graduate students who were entering AI. They did not favor students whose ideologies were compatible with their own, although they did require us to listen to some luncheon speeches by Edward Teller, and they knowingly gave fellowships to several students who opposed militarism. The Hertz Foundation, and later ARPA, paid me a decent graduate-student salary during many months when I supplemented my technical work by reading a great deal of phenomenology and literary theory. The culture of MIT itself included a cult of "smartness", such that students deemed "smart" (including most everyone accepted to graduate school) were given wide latitude to pursue their own research directions, no matter how odd they might seem to their worried dissertation committees. If the field of AI during those decades was a servant of the military then it enjoyed a wildly indulgent master.

Within academia, the early AI pioneers were largely engaged in a revolt against behaviorism. Behaviorism, in its day, had organized itself largely in reaction against the vague and unreproducible nature of introspectionist psychology. The metaphors provided by new technologies provided a means of placing mentalist psychology on a scientific basis, and a functionalist epistemology emerged to explain what it meant to offer mental mechanisms as explanations for experimental data. Although the participants in this intellectual movement experienced themselves as revolutionaries, mechanistic explanations of mind already had a long history, going back at least as far as Hobbes. Yet, in a curious twist, these mentalists acknowledged little inspiration from this tradition. Instead they reached back three hundred years to identify themselves with the philosophy of Descartes. Although Descartes' defense of a specifically mental realm in a mechanistic universe was a useful symbol for mentalists engaged in polemics against behaviorism, the principal appeal of Descartes' theory was not its ontological dualism, but rather the explanatory freedom that Descartes' dualism afforded. The theorizing of later mechanists, from Hobbes to Locke to the associationists and reflex-arc theorists, was severely constrained by the limitations of their mechanical models. Descartes, on the other hand, could prescribe elaborate systems for the conduct of the mind without worrying how these systems might be realized as physical mechanisms. He intended his rules as normative prescriptions for rational thought; they were explanatory theories in the sense that they would accurately describe the thinking of anybody who was reasoning rationally. Although nobody has mechanized Descartes' specific theory, the stored-program digital computer, along with the theoretical basis of formal language theory and problem-solving search and the philosophical basis of functionalism (Fodor 1968), provided the pioneers of AI with a vocabulary through which rule-based accounts of cognitive rationality could be rendered mechanical while also being meaningfully treated as mental phenomena, as opposed to physical ones.

3 Self-conception

These aspects of AI's institutional and intellectual origins help to explain its distinctive conception of itself as a field. The central practice of the field, and its central value, was technical formalization. Inasmuch as they regarded technical formalization as the most scientific and the most productive of all known intellectual methods, the field's most prominent members tended to treat their research as the heir of virtually the whole of intellectual history. I have often heard AI people portray philosophy, for example, as a failed project, and describe the social sciences as intellectually sterile. In each case their diagnosis is the same: lacking the precise and expressive methods of AI, these fields are inherently imprecise, woolly, and vague.

Any attempt at a critical engagement with AI should begin with an appreciation of the experiences that have made these extreme views seem so compelling. In its first fifteen years, AI developed a series of technical methods that provide interesting, technically precise accounts of a wide range of human phenomena. It is often objected that these machines are not "really" reasoning or planning or learning, but these objections can miss the point. The early demonstrations of AI were incomparably more sophisticated than the

mechanistic philosophies of even a short time before. As a result, the people who had stayed up many late nights getting them to work felt justified in extrapolating this extraordinary rate of progress for one or two or three more decades at least. Critics of their research have often focused on particular substantive positions that have seemed unreasonable, for example the frequent use of computer symbols such as REASON and DECIDE and GOAL whose relationship to the actual human phenomena that those words ordinarily name is suggestive at best. But AI's fundamental commitment (in practice, if not always in avowed theory) is not to a substantive position but to a method. Any particular set of mechanisms will surely prove inadequate in the long run, but it serves its purpose if it forces issues to the surface and sharpens the intuitions that will guide the development of the next mechanism along. AI people generally consider that their goals of mechanized intelligence are achievable for the simple reason that human beings are physically realized entities, no matter how complex or variable or sociable they might be, and AI's fundamental commitment (again, in practice, if not always in avowed theory) is simply to the study of physically realized entities, not to production systems or symbolic programming or stored-program computers.

In relation to the rest of the intellectual and technical world, then, AI long regarded itself as simultaneously central and marginal. It understood itself as central to human intellectual endeavor, and its integral connection to the closed-world agenda ensured that its main research centers (MIT, CMU, and Stanford) would number among the most prominent scientific laboratories in the world. But it was marginal in other, sometimes peculiar senses. Not only was it intellectually autonomous to a significant degree, but it was also a small world. Research results were communicated through internal report series, IJCAI (the biannual International Joint Conference on Artificial Intelligence), and the ARPANET. Its principal archival journal, Artificial Intelligence, was an important publication venue, but the central laboratories did not emphasize journal publication, and graduate students were often not taught how to write papers for journals.

The sense of marginality extended to the culture of the field. Much has already been written about the pecularities of the "hacker culture" (Hapnes and Sorensen 1995, Turkle 1984, Weizenbaum 1976), with its extreme informality, emotional simplicity, resentment of externally imposed structures and constraints, and the leeway that the hackers afforded to one another's eccentricities. It was, paradoxically, an intensely social culture of seemingly quite asocial people. Whether explicitly or tacitly, they opposed the falseness of bureaucratic life to the principled meritocracy of their craft. Building things was truly the end purpose of the hacker's work, and everything about the methods and language and value system of the AI world was organized around the design and implementation of working systems. This is sometimes called the "work ethic": it has to work. The "result" of an AI research project is a working system whose methods seem original and broadly applicable; an "idea" is a method of building technical systems or a way of analyzing problems that motivates a promising system design; and a research "approach" is a conceptual and technical framework by which problems can be analyzed and transformed into a particular type of technical system (Chapman 1991: 213-218). The field, accordingly, reckons its history primarily as a sequence of computer systems and secondarily as a history of debates among different approaches to the construction of systems.

It is commonly supposed that work in technical fields proceeds through sharply defined rational, logical reasoning. Many technical people actually believe this to be the case, but in AI at least, it is not true. The next section will describe some consequential fallacies in the field's ideas about precision and rigor, but it is equally important to understand the role of intuition in the AI's own explicit understandings of itself as a technical practice. Whereas industrial computer programming is organized primarily around specifications that govern the input-output behavior of the various modules of a system, research programming in AI is self-consciously virtuosic and experimental. Much of the field's internal discourse has been concerned with the intuitions that guide the design of its complex, ambitious systems. The principle of modularity, for example, might be treated as an axiom or an instrumental expedient in industrial programming. But AI people understand modularity as a powerful but somewhat elusive principle of the universe, akin to a law of nature but much harder to define (Abelson and Sussman 1984, Simon 1970). The point is certainly not that AI people are mystics, or that they consciously wish to make anything obscure, but rather that they take seriously the craft nature of their work. AI people, likewise, are constantly discovering that different considerations trade off against one another. Modularity trades off against efficiency, for example, in the sense that systems can usually be made more efficient by breaking down the modularity boundaries that limit the amount of

information that two components of a system can share. The expressive power of representation schemes trades off against efficiency as well, inasmuch as symbolic matching and inference tasks become rapidly less tractable as the representation languages provide a greater variety of ways of expressing equivalent concepts. Each of these broad generalizations can be made perfectly formal in the context of particular, concrete design decisions, yet the generalizations themselves seem worthy of articulation and reification as lessons learned from research despite their informality. The enormous obstinacy of technical work -- if a method cannot be made to work in a given case then no amount of sloppiness or vagueness will make it work -- seems to back these potentially nebulous intuitions with a "hardness" and irrefutability that philosophical or literary research never seems (to AI people anyway) capable of achieving.

4 Discursive practices

The premise of AI, in rough terms, is the construction of computer systems that exhibit intelligence. One encounters different formulations of this premise at different labs, and from different individuals in the field. In philosophical and popular forums, the field is often discussed in terms of a seemingly fundamental question: can computers think? But little of the field's day-to-day work really depends on the answer to such questions. As a practical matter, the purpose of AI is to build computer systems whose operation can be narrated using intentional vocabulary. Innovations frequently involve techniques that bring new vocabulary into the field: reasoning, planning, learning, choosing, strategizing, and so on. Whether the resulting systems are really exhibiting these qualities is hard to say, and AI people generally treat the question as an annoying irrelevance. What matters practically is not the vague issue of what the words "really mean" but the seemingly precise issue of how they can be defined in formal terms that permit suitably narratable systems to be designed. If you disapprove of the way that we formalize the concept of reasoning or planning or learning, they are likely to say, then you are welcome to invent another way to formalize it, and once you have gotten your own system working we will listen to you with rapt attention. If you disapprove of the very project of formalization, or if you insist on sensitivity to the ordinary vernacular uses of the words (e.g., Button et al 1995), then, they would argue, you are simply an obscurantist who prefers things to remain vague.

In an important sense, then, AI is a discursive practice. A word such as planning, having been made into a technical term of art, has two very different faces. When a running computer program is described as planning to go shopping, for example, the practitioner's sense of technical accomplishment depends in part upon the vernacular meaning of the word -- wholly arbitrary neologisms would not suffice. On the other hand, it is only possible to describe a program as "planning" when "planning" is given a formal definition in terms of mathematical entities or computational structures and processes. The subfield of "planning research" consists of an open-ended set of technical proposals, joined by a densely organized family relationship but not by any a priori technical definition, about the implementable senses in which words in the semantic field around "planning" ("plan", "goal", "execution", "actions", "policies", and so forth) might be used. Different schools certainly differ in their standards of formalization, from "neat" (that is, explicitly and systematically mathematical) to "scruffy" (demonstrated simply through a compelling program). But they emphatically agree that the proof is in the programming, and that a proper research result consists in a method for casting planning-like tasks as technical problems that running computer systems can solve.

This dual character of AI terminology -- the vernacular and formal faces that each technical term presents -- has enormous consequences for the borderlands between AI and its application domains. The discourse of "domains" in AI is extraordinarily rich and complicated, and the field's practitioners take for granted a remarkable intellectual generativity. Once a term such as "planning" or "constraints" or "least commitment" has been introduced into the field through a first implemented demonstration in a particular domain, AI people will quite naturally shift that term into other domains, drawing deep analogies between otherwise disparate activities. Once an automated design problem, for example, has been analyzed into a large, discrete set of design choices, it immediately becomes possible to ask whether these choices can be made without backtracking -- that is, whether the choices can be made in some sequence in which earlier decisions never have unhappy implications for choices that must be made later on. Techniques that arose to support the patterns of backtracking that were discovered during research on story-telling may then find application in the automated design domain, or in a medical diagnosis domain, or in the domain of planning shopping trips.

Having proven themselves broadly useful, these techniques might be abstracted into general-purpose algorithms whose computational properties can be studied mathematically, or they might be built into a programming language. Each technique is both a method for designing artifacts and a thematics for narrating its operation.

AI researchers can build computer models of reasoning in particular domains because their discourse is, in one sense, precise. But they can only make such a wide range of domains commensurable with one another because their discourse is, in another sense, vague. At any given time, AI's discursive repertoire consists of a set of technical schemata, each consisting of a restricted semantic field and a specific family of technical methods. Among the most prominent technical schemata are "planning" and "knowledge". Each of these words might be given a wide range of meanings in different cultural or disciplinary contexts. In AI, though, their meanings are closely tied to their associated technical methods, and they are not otherwise constrained. Absolutely any structure or purposivity in anybody's behavior, for example, can be interpreted as the result of planning. This is not a hypothesis -- it is simply how the word is used. Miller, Galanter, and Pribram's *Plans and the Structure of Behavior* (1960), despite its lack of technical demonstrations, is nonetheless the field's original textbook in the rhetoric of planning. Absolutely any enduring competence, likewise, can be interpreted as a manifestation of knowledge; John McCarthy's early papers (e.g., 1968 [1958]) provided one influential AI rhetoric of knowledge in terms of the predicate calculus.

The construction of an AI model begins with these most basic interpretations, and it proceeds systematically outward from them. Having detected an element of behavioral regularity in the life of some organism, for example, one can immediately begin enumerating the unitary elements of behavior and identifying those as the "primitive actions" that the putative planner has assembled to produce its plan. Miller, Galanter, and Pribram, motivated by Chomsky's linguistic formalisms and Newell and Simon's early problem-solving programs, helpfully suggested that all plans are hierarchical: a morning's activity might comprise several distinct activities (dressing, eating breakfast, answering correspondence), and each of those activities can be understood as themselves comprising distinct subactivities, which are themselves composite activities in turn, until finally one reaches a suitably elementary repertoire of actions from which all others are assembled. Miller, Galanter, and Pribram never offered a definitive set of these primitive actions, and the field has never felt it necessary to do so. (Schank's (1975) theory of the mental representation of action for purposes of story understanding, though, includes a fixed repertoire of primitive action types.) The purpose of the theory of planning has not been to provide a single technical specification for all domains, but rather to provide a set of technical schemata that can be expanded into a narrative thematics for any particular domain. Much of the practical work of AI, in other words, consists precisely in the deployment of these technical schemata to translate, or gloss, selected features of a given domain in terms that can also be interpreted as the operation of a computer program. The vagueness of AI vocabulary is instrumental in achieving this effect.

The strategic vagueness of AI vocabulary, and the use of technical schemata to narrate the operation of technical artifacts in intentional terms, is not a matter of conscious deception. It does permit AI's methods to seem broadly applicable, even when particular applications require a designer to make, often without knowing it, some wildly unreasonable assumptions. At the same time, it is also self-defeating. It has the consequence, at least in my own experience, that AI people find it remarkably difficult to conceptualize alternatives to their existing repertoire of technical schemata. The idea that human beings do not conduct their lives by means of planning, for example, is just short of unintelligible. At best it sounds like behaviorism, inasmuch as it seems to reject all possible theories of the mental processing through which anyone might decide what to do. The term "planning", in other words, exhibits an elastic quality: as a technical proposition it refers to a quite specific and circumscribed set of functionalities and algorithms, but as an empirical proposition it refers to anything at all that can plausibly be glossed with the term. This elasticity of meaning is already found in Miller, Galanter, and Pribram. Their formal definition of a "Plan" (they capitalize the term) is "any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed" (1960: 16). A Plan is defined as a "process", and yet "process" is given no technical definition, either in their book or in subsequent planning research. Despite the broad and inclusive connotations of "any hierarchical process", in practice they use the word Plan much more specifically. In some places it refers to a "TOTE unit", which is a simple kind of feedback loop, and in other places it refers

to a parse tree of the type described by formal language theory. This latter version has been the more influential, and virtually all AI planning theories interpret a "plan" as a symbolic datastructure that functions essentially as a computer program (another connotation that Miller, Galanter, and Pribram gesture at without formally embracing). As a result of this equivocation, attempts to deny the narrow technical theory sound to the ears of AI researchers like denials that the sequential ordering of human behavior is determined by any coherent process at all.

Miller, Galanter, and Pribram's concept of a Plan also exemplifies another prominent feature of AI discourse: the tendency to conflate representations with the things that they represent. Their substantive theory is that behavior derives its structure from the structure of a Plan, and so they taught a generation of AI practitioners how to shift rapidly back and forth between talk about the structure of outward behavior and the structure of internal mental processes, and between the structure of these time-extended phenomena and the structure of static symbolic structures in the mind. This conflation of representations and worldly things is particularly encouraged by the domains that early AI research chose to illustrate its techniques. Newell and Simon's (1963, 1972) problem-solving research, for example, employed logical theorem-proving and puzzle-solving domains for which the distinction between mental representation and corporeal reality were shady. Proving logical theorems in one's head, after all, is a different activity from proving them with pencil and paper, but the essentially mathematical nature of the domain permits the distinction between logical propositions in working memory and logical propositions written on paper to be blurred. In particular, the mental representations readily capture everything about the real-world entities that can ever have any consequences for the outcome of the task. These domains appealed to early AI researchers in part because computer vision and robotics were very poorly developed, and they permitted research on cognition to begin without waiting on those other, possibly much more difficult research problems to be solved. But the privileged status of mathematical entities in the study of cognition was already central to Descartes' theory, and for much the same reason: a theory of cognition based on formal reason works best with objects of cognition whose attributes and relationships can be completely characterized in formal terms. Just as Descartes felt that he possessed clear and distinct knowledge of geometric shapes, Newell and Simon's programs suffered no epistemological gaps or crises in reasoning about the mathematical entities in their domains.

The conflation between representations and things can be found in numerous other aspects of AI research. It is found, for example, in the notion that knowledge consists in a model of the world, so that the world is effectively mirrored or copied inside each individual's mind. This concept of a "model", like that of a "plan", has no single technical specification. It is, rather, the signifier that indexes a technical schema: it provides a way of talking about a very wide range of phenomena in the world, and it is also associated with a family of technical proposals, each of which realizes the general theme of "modeling the world" through somewhat different formal means. Just as disagreements with the planning theory are unintelligible within AI discourse, it makes virtually no sense to deny or dispute the idea that knowledge consists in a world model. The word "model", like the word "plan", is so broad and vague that it can readily be stretched to fit whatever alternative proposal one might offer. AI people do not understand these words as vague when they are applied to empirical phenomena, though, since each of them does have several perfectly precise mathematical specifications when applied to the specification of computer programs.

5 Waking up

My portrait of the AI community in the previous three sections is, of course, a retrospective understanding. Although they seem commonsensical to me now, and may seem commonsensical to others who have never been practitioners in the field, as an autobiographical matter I only came to these ideas through a long struggle. I had gone to college at an early age, having been constructed as a math prodigy by a psychologist in the region of the country where I grew up. (The arrival of court-ordered school integration in that region coincided with an emphasis on identifying talented students and grouping students into classrooms based on their test scores.) I began my college work as a math major before drifting over to the computer science department. My college did not require me to take many humanities courses, or learn to write in a professional register, and so I arrived in graduate school at MIT with little genuine knowledge beyond math and computers. This realization hit me with great force halfway through my first year of graduate school, and

I took a year off to travel and read, trying in an indiscriminate way, and on my own resources, to become an educated person.

My lack of a liberal education, it turns out, was only half of my problem. Only much later did I understand the other half, which I attribute to the historical constitution of AI as a field. A graduate student is responsible for finding a thesis topic, and this means doing something new. Yet I spent much of my first year, and indeed the next couple of years after my time away, trying very hard in vain to do anything original. Every topic I investigated seemed driven by its own powerful internal logic into a small number of technical solutions, each of which had already been investigated in the literature. My attempts to investigate the area of concept learning, for example, endlessly converged back to a single idea: that all possible definitions of concepts form a mathematical lattice, and all reasonable inferences from evidence about a concept's correct scope could be analyzed in terms of lattice-theoretic operations of meeting and joining. This idea was already implicit in Winston's (1975) early research on concept induction, and had been fully worked through by others subsequently. It seemed inescapable, and overwhelmingly so.

With fifteen years' distance, I can now see that the idea of concept induction through lattice-crawling is indeed inescapable if one's ideas about concepts and evidence and learning are constrained by the ensemble of technical schemata that operated in the discourse and practice of AI at that time. But fifteen years ago, I had absolutely no critical tools with which to defamiliarize those ideas -- to see their contingency or imagine alternatives to them. Even worse, I was unable to turn to other, nontechnical fields for inspiration. As an AI practitioner already well immersed in the literature, I had incorporated the field's taste for technical formalization so thoroughly into my own cognitive style that I literally could not read the literatures of nontechnical fields at anything beyond a popular level. The problem was not exactly that I could not understand the vocabulary, but that I insisted on trying to read everything as a narration of the workings of a mechanism. By that time much philosophy and psychology had adopted intellectual styles similar to that of AI, and so it was possible to read much that was congenial -- except that it reproduced the same technical schemata as the AI literature. I believe that this problem was not simply my own -- that it is characteristic of AI in general (and, no doubt, other technical fields as well). This is not to say that AI has no intellectual resources and no capacity for originality. In recent years particularly, the field has made productive connections with a wide variety of other technical fields, establishing common cause through the sharing of technical schemata.

My own route was different. I cannot reproduce its whole tortuous detail here, and so it will inevitably sound simpler in the retelling than it was in the living. But the clarity of hindsight makes evident that I drew on the internal resources of the field, even as I struggled to find my way out of it. I began by filling my notebook with exhaustively detailed stories from my own everyday life. By this time I had grown preoccupied with planning research, d so I decided to gather some examples of real-life planning. In doing so, I was following an AI tradition of introspection that has been described aptly, if unsympathetically, by Turkle (1984). Many early AI researchers were clearly attempting, at one level or another, to reproduce their own psyches on computers, and many of them drew on introspection to motivate their programs. Introspection as a formal research method in psychology, of course, had been comprehensively discredited decades earlier. But AI people have not regarded introspection as evidence but as inspiration; since the functionality of their computer systems provides a fully adequate criterion of the success of their research, they believe, it does not matter what experiences might have motivated the systems' design. And introspection is close at hand.

But my own practice was different from introspection in one important respect: whereas introspection attempts to observe and describe mental processes under specially controlled conditions, I was trying to remember and recount episodes of concrete activity that took place in my own everyday life. Together with my fellow student David Chapman, I rapidly developed a method that I called "intermediation". Having noticed some interesting sequence of events in the course of washing the dishes or carrying out the trash, I would write it down in my notebook in as much detail as I could remember. Along the way, I would invent names for aspects of the recounted activity that seemed relevant to some technical concern. The method worked best if these names were intermediate in their degree of abstraction, thus the term "intermediation". For example, I became interested in what I called "hassles", which are small bits of trouble that recur frequently in routine

patterns of activity. Having noticed a hassle, for example an episode in which silverware tried to crawl into the garbage disposal while washing dishes, I would write out in some detail both the episode itself and the larger pattern's attributes as a hassle. Having done so, I found that I would then start spontaneously noticing hassles in other activities, particularly hassles that were analogous in some way to the hassles that I had already noticed and written out in my notebook. (I understand that linguists have this experience all the time.)

I did this regularly for a couple of years, to such an extent that I was continually noticing various aspects of the mundane mechanics of my daily life. I was also continually noticing the many small transformations that my daily life underwent as a result of noticing these things. As my intuitive understanding of the workings of everyday life evolved, I would formulate new concepts and intermediate on them, whereupon the resulting spontaneous observations would push my understanding of everyday life even further away from the concepts that I had been taught. It may be objected that a method driven by a priori concepts can only find whatever it is looking for, but that was not at all my experience. When looking for hassles, of course, I would find hassles. But then writing out the full details of an actual episode of being hassled would raise an endless series of additional questions, often unrelated to what I was looking for. It is hard to convey the powerful effect that this experience had upon me; my dissertation (Agre 1988), once I finally wrote it, was motivated largely by a passion to explain to my fellow AI people how our AI concepts had cut us off from an authentic experience of our own lives. I still believe this.

Perhaps someday I will finally write out my treatise on the true functioning of everyday routine activities, based on the great mass of anecdotes that I accumulated by this procedure. My purpose here, though, is to describe how this experience led me into full-blown dissidence within the field of AI. Given that an AI dissertation is based on a computer program, my investigations of everyday routine activities were always aimed at that goal. I wanted to find an alternative means of conceptualizing human activity -- one that did not suffer the absurdities of planning but that could be translated into a working demonstration program. To this end, I spent many months working back and forth between concepts to describe everyday activities and intuitions that seemed capable of guiding technical work. Most of these intuitions would be impossible to explain without developing an elaborate apparatus of concepts, and indeed I found that my thinking about these matters had become impossible to communicate to anybody else. A small number of my friends, most notably David Chapman, sat still for long, complex explanations of the phenomena I was observing and the intuitions they seemed to motivate. But clearly I had to bring this project back into dialogue with people who did not already share my vocabulary.

In order to find words for my newfound intuitions, I began studying several nontechnical fields. Most importantly, I sought out those people who claimed to be able to explain what is wrong with AI, including Hubert Dreyfus and Lucy Suchman. They, in turn, got me started reading Heidegger's *Being and Time* (1961 [1927]) and Garfinkel's *Studies in Ethnomethodology* (1984 [1967]). At first I found these texts impenetrable, not only because of their irreducible difficulty but also because I was still tacitly attempting to read everything as a specification for a technical mechanism. That was the only protocol of reading that I knew, and it was hard even to conceptualize the possibility of alternatives. (Many technical people have observed that phenomenological texts, when read as specifications for technical mechanisms, sound like mysticism. This is because Western mysticism, since the great spiritual forgetting of the later Renaissance, is precisely a variety of mechanism that posits impossible mechanisms.) My first intellectual breakthrough came when, for reasons I do not recall, it finally occurred to me to stop translating these strange disciplinary languages into technical schemata, and instead simply to learn them on their own terms. This was very difficult because my technical training had instilled in me two polar-opposite orientations to language -- as precisely formalized and as impossibly vague -- and a single clear mission for all discursive work -- transforming vagueness into precision through formalization (Agre 1992). The correct orientation to the language of these texts, as descriptions of the lived experience of ordinary everyday life, or in other words an account of what ordinary activity is *like*, is unfortunately alien to AI or any other technical field.

I still remember the vertigo I felt during this period; I was speaking these strange disciplinary languages, in a wobbly fashion at first, without knowing what they meant -- without knowing what *sort* of meaning they had. Formal reason has an unforgiving binary quality -- one gap in the logic and the whole thing collapses -- but

this phenomenological language was more a matter of degree; I understood intellectually that the language was "precise" in a wholly different sense from the precision of technical language, but for a long time I could not convincingly experience this precision for myself, or identify it when I saw it. Still, in retrospect this was the period during which I began to "wake up", breaking out of a technical cognitive style that I now regard as extremely constricting. I believe that a technical field such as AI can contribute a great deal to our understanding of human existence, but only once it develops a much more flexible and reflexive relationship to its own language, and to the experience of research and life that this language organizes.

My second intellectual breakthrough occurred during my initial attempt to read Foucault's *Archaeology of Knowledge* (1972). Foucault suggested that when two schools of thought are fighting, rather than try to adjudicate the dispute, one should explore whether the opposed schools are internally related components of a single intellectual formation. Having done so, it becomes possible to ask how that whole formation arose historically. I came across this idea at an opportune moment. Although the structuralism of The Archaeology of Knowledge has often been condemned by Foucault's critics, this very structuralism nonetheless ensured that I could grasp Foucault's ideas within my habitual patterns of technical thought, and that I could then employ his ideas to objectify and defamiliarize those very patterns of thought. It became possible, for example, to inquire into the nature and workings of the discursive formation that consisted of behaviorism plus cognitivism. This was an extraordinary revelation.

It may be objected that *The Archaeology of Knowledge* is only one possible theory of the history of ideas, and that dozens of preferable theories are available. My point, however, is that my technical training did not include any of those other theories. I later became a zealous consumer of those theories, but it was Foucault's theory that first pierced the darkness -- precisely because of its commensurability with the order of technical thought. Having found a means of objectifying ideas, I could then proceed systematically to extricate myself from the whole tacit system of intellectual procedures in which I had become enmeshed during my years as a student of computer science. For this reason, I have never experienced poststructuralism or literary theory as strange or threatening, nor have I ever perceived them as varieties of relativism or idealism. Quite the contrary, they were the utterly practical instruments by which I first became able to think clearly and to comprehend ideas that had not been hollowed through the false precision of formalism.

6 The fallacy of alternatives

These foreign disciplinary languages were beginning to provide an established vocabulary for expressing the intuitions that I had developed by noticing and writing out episodes of routine activity. In broad outline, my central intuition was that AI's whole mentalist foundation is mistaken, and that the organizing metaphors of the field should begin with routine interaction with a familiar world, not problem-solving inside one's mind. In taking this approach, everything starts to change, including all of the field's most basic ideas about representation, action, perception, and learning. When I tried to explain these intuitions to other AI people, though, I quickly discovered that it is useless to speak nontechnical languages to people who are trying to translate these languages into specifications for technical mechanisms. This problem puzzled me for years, and I surely caused much bad will as I tried to force Heideggerian philosophy down the throats of people who did not want to hear it. Their stance was: if your alternative is so good then you will use it to write programs that solve problems better than anybody else's, and then everybody will believe you. Even though I believe that building things is an important way of learning about the world, nonetheless I knew that this stance was wrong, even if I did not understand how.

I now believe that it is wrong for several reasons. One reason is simply that AI, like any other field, ought to have a space for critical reflection on its methods and concepts. Critical analysis of others' work, if done responsibly, provides the field with a way to deepen its means of evaluating its research. It also legitimizes moral and ethical discussion and encourages connections with methods and concepts from other fields. Even if the value of critical reflection is proven only in its contribution to improved technical systems, many valuable criticisms will go unpublished if all research papers are required to present new working systems as their final result.

Another, more subtle reason pertains to AI's ambiguous location between science and engineering. A scientific theory makes truth-claims about the preexisting universe, and so it is generally considered legitimate to criticize someone else's theory on grounds of methodological weakness, fallacious reasoning, lack of fit with the evidence, or compatibility of the evidence with other theories. Engineering design methods, on the other hand, make claims in the context of practical problems, and so the legitimate criticisms relate solely to issues of utility. AI projects are sometimes scientific in intention, sometimes engineering, and sometimes they shift subliminally from one to the other. AI people often make substantive claims about knowledge or learning or language, and yet many of them will respond with indignation to arguments that their projects fundamentally misconstrue the nature of these phenomena; in most cases (the primary exception being Newell and Simon's research group at Carnegie-Mellon University) they will argue *not* that the claims against their work are empirically false but that they are non sequiturs. Pressed to explain the seeming contradiction, they will generally state that their systems exhibit knowledge-as-such, say, as opposed to human knowledge in particular. But then, it seems, they will turn around and compare their systems' behavior to human behavior without further comment. The underlying problem is not mendacity but a conflict of languages: norms and discourses of engineering are applied to terms (knowledge, learning, language, and so on) whose meanings are inextricably rooted in the phenomena of human life. As a consequence, I have often encountered an emphatic, explicitly stated injunction against "criticizing other people's work", the idea being that the only legitimate form of critical argument is that "my system performs better than your system on problem X".

A final reason, which I have already discussed above, is that AI discourse makes it exceptionally difficult to conceptualize alternatives to the field's prevailing ideas. Indeed, AI does not have "ideas" in any sense that would be familiar from philosophy or literature or social thought; instead it has technical practices, loosely articulated intuitions about those practices, and ways of talking about the resulting artifacts that combine precision and vagueness in specific ways. If you write a program whose operation you understand in different terms then somebody will observe that your program can perfectly well be described as having knowledge, executing plans, and so on. Never mind, then, that you choose to talk about the program differently; in fact, they will say, it is nothing new. The seemingly commonsensical demand to prove alternatives in practice is thus actually a demand to express disagreements with the existing language within the existing language itself, and this is nearly impossible.

In these ways, AI's construction of itself as a self-contained technical discipline, though seemingly governed by practical-minded criteria of success and failure, is actually a powerful force for intellectual conservatism. Critics will be asked, "what's your alternative?", within a tacit system of discursive rules that virtually rules out alternatives from the start. All the same, I think that the very concept of "alternatives" is misleading, and that it is actually impossible to achieve a radical break with the existing methods of the field. This is because AI's existing language and technical practice, like any disciplinary culture, runs deeper than we are aware. Having been socialized into the field, by the time I began conceiving myself as a dissident I had acquired an extensive network of linguistic forms, habits of thought, established techniques, ritualized work practices, ways of framing questions and answers, genre conventions, and so forth. It would have been impossible to simply cast off that whole network of cultural forms, any more than I could simply decide to stop being American and start being Thai, or to become transcendentally stateless and cultureless. As a result, attempts to formulate a wholly distinct alternative worldview for AI, or even to secede from the field altogether, are bound to fail. The point is exceptionally subtle: AI's elastic use of language ensures that nothing will seem genuinely new, even if it actually is, while AI's intricate and largely unconscious cultural system ensures that all innovations, no matter how radical the intentions that motivated them, will turn out to be enmeshed with traditional assumptions and practices. When AI people look at an innovation and pronounce it nothing radically new, they will be wrong in some ways and right in others, and it will require tremendous effort to determine which is which. Critical practice is essential to make sense of such things, but its goal should be complex engagement, not a clean break.

I began to understand this once I had attained a few years' critical distance on two computer programs that illustrated the intuitions and technical ideas that arose through intermediation on the workings of ordinary activities; I have described these programs in my dissertation (Agre 1988) and more recently in my book (Agre

in press). I wrote the first of these, a program called RA that operates in a conventional AI "blocks world", as an experiment in computational improvisation; rather than constructing a plan which it then executes wholesale, it conducts a fresh reasoning-through of its situation as quickly as it can. David Chapman wrote the second, with some participation from me in the later stages of development, a program called Pengi that played a video game by a similar improvisation, but with a much more sophisticated model of visual perception.

Although intended as alternatives to the conventional theories of planning, reasoning, and vision, these programs ultimately turned out to recapitulate some subtle confusions of the conventional methods. Specifically, both of these programs relate to their "worlds" from a bird's-eye view -- or, more precisely, from an orthographic projection, as if the simulated agent were infinitely far away from the action except for the one instrument through which it moves things around: a cartoon "hand" for RA and a cartoon penguin for Pengi. Orthographic projections are ubiquitous in the diagrams of AI papers; they make it seem reasonable that the simulated agent maintains a panoptic representation of its environment. This is one particularly insidious manifestation of AI's tendency to conflate representations and things-represented. In the case of RA, the conflation was already hidden by a convention of blocks world: the blocks have their names ("A", "B", "C", etc) written on them, as if the agent's mental symbols were part of the material world, automatically connected to the things they name. In the case of Pengi, the problem was much more subtle. Pengi's mechanisms for deciding what to do were modeled on RA's. Pengi, however, employed a somewhat realistic model of vision, and so it did not automatically represent the whole of reality to itself. But the adversarial and partially random nature of the video game meant that Pengi could not rely on stable structures in the world or large-scale patterns in its interactions with the world to help it keep track of things. And since Pengi had no body, and thus no strong sense of being located anywhere in particular, every part of the visible "world" was potentially relevant all of the time. As a result, Pengi (like any player of a video game) was forever scrambling to allocate its attentional resources to the most important objects in its visual world. It could carry on reasonably well by focusing primarily on the objects closest to it. It could only keep track of individual objects, however, by physically tracking them across the screen, much as RA kept track of blocks by knowing their names. In the end, therefore, Pengi did not provide the clear-cut alternative that we had hoped.

Neither of these programs was a failure. To the contrary, each of them introduced technical methods that may have some lasting value. And each of them does point to the utility of different metaphors for the technical work of the field, even if it proves impossible to make a knock-down argument for one set of metaphors over another. At the same time, each program reflects the inherent difficulty of inventing a thoroughgoing alternative to established technical methods. It is thus not surprising in retrospect that I have found myself exchanging published arguments with another AI dissident, Terry Winograd, that one another's alternative technical ideas -- his outside of AI and mine inside -- are not actually as technically new as their associated rhetoric makes them out to be (Agre 1994, Winograd 1995). We are both right, yet neither project is discredited as a result. On a technical level they are inevitably incremental advances, just as AI people insist they are. But the conceptual analysis and philosophical critique that accompany them must be understood as intellectual contributions in their own right, grounded both in a priori analysis of the phenomena and in detailed, critically informed reflection on the difficulties encountered in getting AI models to work.

It is useful, by way of summary, to distinguish four reasons why it is difficult to create alternatives to the standard methods of AI. First, it is difficult to become aware of the full range of assumptions underneath existing practices, from technical methods to genre conventions to metaphors. Second, having formulated an alternative intuition, it is difficult to reduce that intuition to a novel technical method -- a new type of artifact, or a new way of building artifacts. Third, having invented a new technical method, it is difficult to prevent that method from being construed as "nothing new" within the elastic boundaries of existing technical schemata. Fourth, having coupled a new technical method with a new way of talking about the phenomena, it is difficult to apply the method to any real cases without inventing a lot of additional methods as well, since any worthwhile system will require the application of several interlocking methods, and use of the existing methods may distort the novel method back toward the traditional mechanisms and the traditional ways of talking about them. This litany of obstacles does not make critical technical practice impossible; it simply defines the terrain upon which such a practice must work.

7 Critical engagement

I must leave it to others to determine how effective, if at all, my attempts to reform AI have been. I know that the original Pengi paper (Agre and Chapman 1987) has been extensively cited, but one reason for this paper's popularity is that our innovations in models of situated activity happened to correlate with a shift in military strategy toward autonomous battlefield robots under the Strategic Computing Initiative (Edwards 1996: 293-301). There immediately ensued, with scant and mostly disruptive participation from us, another round of consensus-building between ARPA and the AI community about the necessity of "autonomous agents" and "reactive planning". The vocabulary of planning research soon filled with the military discourse of "uncertain, unpredictable, or changing environments" (e.g., Hendler 1990). Was our seemingly lonely work in the mid-1980's subliminally influenced by the ongoing changes in military thinking? Were we working through an immanent trend in the logic of AI work that paralleled an immanent evolution in the "closed world"? Did our laboratory's attunement to shifts in military thinking create conditions, however unconsciously, for the years of toleration of our strange investigations? I cannot know.

Whatever the case may be, we were not alone in exploring an interactionist style of AI. Authors such as Brooks (1986) and Rosenschein and Kaelbling (1986) were working on broadly similar issues, even though their technical concerns were often different and their philosophical approach was less elaborate. Additional related work has been gathered in Agre and Rosenschein (1996). Research in this style is now reasonably well established as one "approach" within the field.

Beyond these technical concerns, Chapman and I attempted in our papers and talks over several years to provoke critical reflection within the field. Since we were traveling without a map, most of our strategies are inevitably embarrassing in retrospect. We managed to make ourselves controversial in any event, and some people seem to believe that we and other dissidents and critics of the field (Randy Beer, Bill Clancey, Hubert Dreyfus, Jim Greeno, Jean Lave, Lucy Suchman, Terry Winograd, and others) constitute some kind of new establishment unto ourselves. While I cannot evaluate this belief with total precision, I can testify that this utterly disparate group has never tried to constitute itself as a school or movement. AI is still very much its own coherent center of mass, though for many reasons it is less centralized than it had been in the 1970's, and no equally coherent "critical" school has arisen to compete with it. It is a real question whether such a scenario would even make sense.

What, then, can be learned? When I first started trying to reform AI, I believed in revolutions. It seemed to me that I could clear the ground completely and start over, working out a whole alternative intellectual system that would replace everything that was there before. The concept of a generative metaphor seemed to hold out particular promise in this direction, given that so much of the underlying substantive problem with AI really can be understood as expressing a single principle, mentalism as opposed to interactionism. It seemed as though one could throw away the old foundation, and everything on top of it, and start over.

Now I do not believe that it works that way. Instead, I believe in something more like hermeneutics. The intellectual utility of technical exercises, aside from the practical utility that they might actually have in the world, lies precisely in their limitations and failures. Perhaps we can learn to approach technical work in the spirit of reductio ad absurdum: faced with a technical difficulty, perhaps we can learn to diagnose it as deeply as possible. Some difficulties, of course, will be superficial and transient. But others can serve as symptoms of deep and systematic confusions in the field. We are only aware of this possibility if we develop the critical tools to understand the depths below the ordinary practices of a technical field. Some of these critical tools will draw on the methods of institutional and intellectual analysis that have served generations of philosophers, sociologists, and literary critics (Agre 1995). Others may be responses, each sui generis, to the specific properties of technical work. Research could proceed in a cycle, with each impasse leading to critical insight, reformulation of underlying ideas and methods, fresh starts, and more instructive impasses.

But the language of hermeneutics is not adequate, either, because it suggests a solitary "reader" facing the practical reality of technical work as an individual. Technical work is performed in and by communities, and a critically engaged practitioner cannot hope to found an alternative community in which everyone shares the

same critical premises and methodologies. As I worked my way toward a critical technical practice, this was the part that I found hardest: maintaining constructive engagement with researchers whose substantive commitments I found wildly mistaken. It is tempting to start explaining the problems with these commitments in an alien disciplinary voice, invoking phenomenology or dialectics as an exogenous authority, but it is essentially destructive.

The constructive path is much harder to follow, but more rewarding. Its essence is to evaluate a research project not by its correspondence to one's own substantive beliefs but by the rigor and insight with which it struggles against the patterns of difficulty that are inherent in its design. Faced with a technical proposal whose substantive claims about human nature seem mistaken, the first step is to figure out what deleterious consequences those mistakes should have in practice. If the predicted impasses have actually detected in the technical work, then the next step is not to conclude that AI, considered as a static essence, has been debunked in a once-and-for-all fashion. Instead, research can now proceed on the basis of a radical interpretation of their significance, inevitably incremental in its practical effect but more sophisticated than it would have been otherwise, leading toward new and different problems. Or perhaps the predicted impasses have not been detected, in which case one might ask why they have been overlooked. Technical impasses can be overlooked for many reasons; they can be buried in vague or ambiguous language, in notational conventions, in experimental designs, in seemingly unproblematic assumptions, and in many other places. Critical methods might be helpful in discovering other ways in which technical troubles can be inadvertently hidden from view. But nothing can substitute for the daily work of trying to get things built and working. Technical research can only develop from within the designer's own practical work, and it will only progress when the designer's experience is neither channeled by self-reinforcing conceptual schemata from inside the field nor delegitimated by incommensurable philosophies from outside of it. Cultivating the painstaking middle way between these hazards is probably not my own path any more, but it is very much what Collins (1990) had in mind in his philosophically astute but constructively minded research on expert systems, and perhaps it will be a path for others in the future.

A critical technical practice will, at least for the foreseeable future, require a split identity -- one foot planted in the craft work of design and the other foot planted in the reflexive work of critique. Successfully spanning these borderlands, bridging the disparate sites of practice that computer work brings uncomfortably together, will require a historical understanding of the institutions and methods of the field, and it will draw on this understanding as a resource in choosing problems, evaluating solutions, diagnosing difficulties, and motivating alternative proposals. More concretely, it will require a praxis of daily work: forms of language, career strategies, and social networks that support the exploration of alternative work practices that will inevitably seem strange to insiders and outsiders alike. This strangeness will not always be comfortable, but it will be productive nonetheless, both in the esoteric terms of the technical field itself and in the exoteric terms by which we ultimately evaluate a technical field's contribution to society.

*** Acknowledgements**

This chapter has benefitted from comments by David Chapman, Harry Collins, Joseph Goguen, Warren Sack, Penni Sibun, and Patrick Sobalvarro.

*** References**

Harold Abelson and Gerald Jay Sussman, *Structure and Interpretation of Computer Programs*, Cambridge: MIT Press, 1984.

Philip E. Agre, The dynamic structure of everyday life, PhD dissertation, Department of Electrical Engineering and Computer Science, MIT, 1988.

Philip E. Agre, Formalization as a social project, *Quarterly Newsletter of the Laboratory of Comparative Human Cognition* 14(1), 1992, pages 25-27.

Philip E. Agre, Surveillance and capture: Two models of privacy, *The Information Society* 10(2), 1994, pages

101-127.

Philip E. Agre, The soul gained and lost: Artificial intelligence as a philosophical project, *Stanford Humanities Review* 4(2), 1995, pages 1-19.

Philip E. Agre, *Computation and Human Experience*, Cambridge: Cambridge University Press, in press.

Philip E. Agre and David Chapman, Pengi: An implementation of a theory of activity, *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, 1987, pages 196-201.

Philip E. Agre and Stanley J. Rosenschein, eds, *Computational Theories of Interaction and Agency*, Cambridge: MIT Press, 1996.

Rodney A. Brooks, A robust layered control system for a mobile robot, *IEEE Journal of Robotics and Automation* 2(1), 1986, pages 14-23.

Graham Button, Jeff Coulter, John R. E. Lee, and Wes Sharrock, *Computers, Minds, and Conduct*, Cambridge: Polity Press, 1995.

David Chapman, *Vision, Instruction, and Action*, Cambridge: MIT Press, 1991.

Harry M. Collins, *Artificial Experts: Social Knowledge and Intelligent Machines*, Cambridge: MIT Press, 1990.

Paul N. Edwards, *The Closed World: Computers and the Politics of Discourse in Cold War America*, Cambridge: MIT Press, 1996.

Harold Garfinkel, *Studies in Ethnomethodology*, Polity Press, 1984. Originally published in 1967.

Joan Greenbaum and Morten Kyng, eds, *Design at Work: Cooperative Design of Computer Systems*, Hillsdale, NJ: Erlbaum, 1990.

Tove Hapnes and Knut H. Sorensen, Competition and collaboration in male shaping of computing: A study of a Norwegian hacker culture, in Keith Grint and Rosalind Gill, eds, *The Gender-Technology Relation: Contemporary Theory and Research*, London: Taylor and Francis, 1995.

Mike Hales, *Living Thinkwork: Where Do Labour Processes Come From?*, London: CSE Books, 1980.

Martin Heidegger, *Being and Time*, translated by John Macquarrie and Edward Robinson, New York: Harper and Row, 1961. Originally published in German in 1927.

James Hendler, ed, Planning in Uncertain, Unpredictable, or Changing Environments, Proceedings of the AAAI Symposium at Stanford, University of Maryland Systems Research Center Report SRC TR 90-45, 1990.

Marina Jirotka and Joseph A. Goguen, eds, *Requirements Engineering: Social and Technical Issues*, San Diego: Academic Press, 1994.

John McCarthy, The advice taker, in Marvin Minsky, ed, *Semantic Information Processing*, Cambridge: MIT Press, 1968. Originally published in 1958.

George A. Miller, Eugene Galanter, and Karl H. Pribram, *Plans and the Structure of Behavior*, New York: Holt, 1960.

Allen Newell and Herbert A. Simon, GPS: A program that simulates human thought, in Edward A. Feigenbaum and Julian Feldman, eds, *Computers and Thought*, New York: McGraw-Hill, 1963.

Allen Newell and Herbert Simon, *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall, 1972.

Stanley J. Rosenschein and Leslie Pack Kaelbling, The synthesis of digital machines with provable epistemic properties, in Joseph Halpern, ed, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*, Monterey, CA, 1986.

Roger C. Schank, The structure of episodes in memory, in Daniel G. Bobrow and Allan Collins, eds, *Representation and Understanding: Studies in Cognitive Science*, New York: Academic Press, 1975, pages 237-272.

Herbert A. Simon, *The Sciences of the Artificial*, Cambridge: MIT Press, 1970.

Sherry Turkle, *The Second Self: Computers and the Human Spirit*, New York: Simon and Schuster, 1984.

Joseph Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, San Francisco: Freeman, 1976.

Terry Winograd, Heidegger and the design of computer systems, in Andrew Feenberg and Alastair Hannay, eds, *Technology and the Politics of Knowledge*, Bloomington: Indiana University Press, 1995.

Patrick H. Winston, *The Psychology of Computer Vision*, New York: McGraw-Hill, 1975.

Go back to the top of the page