A Functional Role for Repression in an Autonomous, Resource-constrained Agent[1]

Michael Fehling
Intelligent Systems
Laboratory
321 Terman Center
Stanford University
Stanford, CA 94305-4025

Bernard Baars
The Wright Institute
2728 Durand Ave.
Berkeley, CA 94704

Charles Fisher
The San Francisco
Psychoanalytic Institute
2420 Sutter
San Francisco, CA 94115

Abstract

We discuss the capabilities required by intelligent "agents" that must carry out their activities amidst the complexities and uncertainties of the real world. We consider important challenges faced by resource-constrained agents who must optimize their goal-directed actions within environmental and internal constraints. Any real agent confronts limits on the quality and amount of input information, knowledge of the future, access to relevant material in memory, availability of alternative strategies for achieving its current goals, etc. We specify a "minimalist" architecture for a resource-constrained agent, based on Global Workspace Theory (Baars, 1988) and on the research of Fehling (Fehling & Breese, 1988). We show how problem-solving and decision-making within such as system adapts to critical resources limitations that confront an agent. These observations provide the basis for our analysis of the functional role of repression in an intelligent agent. We show that active repression of information and actions might be expected to emerge and play a constructive role in our model of an intelligent, resource-constrained agent.

1. Introduction.

Repression involves the purposeful exclusion of information from consciousness. The repressed material is typically assumed to be too painful, shameful, or anxiety-provoking to be tolerable (Freud, 1915). Although observations suggesting repression are widely reported in the psychological clinic, hard experimental evidence for repression remains problematic. Even skeptics, however, agree that some sort of avoidance of painful thoughts is extremely common and demonstrable (Holmes, 1967; Erdelyi, 1985). This paper will not focus on the issue of psychological evidence. Rather, we will ask, could repression, or its functional equivalent, emerge naturally, in any autonomous agent, operating in goal-directed but resource-constrained fashion (Fehling et al., 1989)?

This paper presents a first report of an effort to merge three, rather distinct perspectives on the nature of intelligent agency:

- a functional theoretical perspective that combines a psychological model known as the Global Workspace Hypothesis (Baars, 1988) and a computational model of intelligent problem-solving by resource-constrained agents (Fehling, Altman, & Wilber, 1989),
- a normative theoretical perspective of reasoning adapted from decision-theory and decision analysis, and
- aspects of a Freudian theory of unconscious, psychodynamic processes.

Our discussion of repression illustrates a perspective and methodology that is emerging from our efforts to fuse these distinct theoretical approaches to the nature of intelligent agency.

2. Resource-constrained Agency.

Intelligent problem-solvers and decision makers are *resource-constrained agents*. When acting autonomously in complex, dynamically changing situations that arise in the real world, these agents must cope with the scarcity of resources that may be essential for carrying their tasks. Resource-constrained agents must act to formulate and achieve their objectives without violating constraints imposed by these limited, task-critical resources.

Time is a critical resource. In real-world situations task-performance time is *always* limited. Temporal limitations significantly impact the appropriateness and effectiveness of an agent's actions. An intelligent agent's ability to react to, and interact with, its environment may be constrained by deadlines on the time available to complete those actions. Deadlines require an agent to foreshorten the time spent deliberating about, and carrying out, its actions. Thus, an agent may need to limit the specificity, precision, or quality of its actions in order to meet such real-time constraints. An agent may also need to synchronize its actions with the independent occurrence of environmental events. For example, during diagnostic reasoning an intelligent agent may need to carefully manage the time spent in gathering and interpreting data from various sources in order to observe and interpret intermittently available data (D'Ambrosio et al., 1987).

Information is an equally critical resource. In real-world situations, intelligent agents seldom, if ever, have sufficient information to fully determine the truth of beliefs or unambiguously select the best course of action. Informational incompleteness or uncertainty significantly constrains the effectiveness of an intelligent agent's actions and deliberations. So, an intelligent agent may seek to reduce its uncertainty through *belief management*, including methods to actively gather data and to infer new beliefs from existing ones with sound methods of inference, such as logical or probabilistic deduction. Unfortunately, belief management entails costs as well as benefits (Fehling & Breese, 1988). However beneficial an agent's belief management efforts may be in reducing uncertainty, undertaking these efforts will consume time and other resources. For example, a medical diagnostician may find it quite difficult to decide whether to carry out possibly dangerous exploratory surgery, even though this procedure would return useful information. In addition, by focusing its attention on data-gathering observations or on inferential deliberation, an agent may compromise its ability to achieve other important objectives. Consider a hiker planning the quickest path back to camp before dark. Excessive time in planning could obviate the very goal it is intended to achieve.

An intelligent agent's cognitive capacities represent another set of critical resources. Even the most well-endowed agent is constrained by a finite capacity for storing and retrieving data from memory and rate of inferring new conclusions from previously stored information. While an agent can only indirectly affect the availability of exogenous resources such as time and information, these endogenous, cognitive resources are under more direct control. To cope with its finite capacities, an intelligent agent must judiciously manage the distribution of its cognitive resources among the activities to which they could potentially be applied. Selective attention phenomena may reflect very directly an agent's attempts to cope with the scarcity of its cognitive resources.

Our discussion illustrates how an intelligent agent must make the best use of scarce resources to achieve its most critical objectives. *An agent's effectiveness in reasoning and*

acting depends upon how well that agent can adapt or modify what it might do under ideal circumstances to conform to constraints imposed by the limited availability of critical resources. Theories and computational models of intelligent agency as well as specific problem-solving models must inevitably account for this type of adaptivity.

In general, to achieve this adaptivity, an intelligent agent must seek to maximize the quality of its task performance within the recognized resource constraints. Thus, the agent must employ "meta-level" processes that manage its problem-solving and decision-making activities according to this general adaptivity requirement (Fehling & Breese, 1988). These meta-level processes may

- select the actions or strategies that make the best use of available resouces from previously constructed alternatives
- modify existing actions or strategies, or construct entirely new responses, to simultaneously meet resource constraints and maximize performance quality, or
- some mix of these two basic approaches (Fehling, Sagalowicz, & Joerger, 1986).

Achieving this adaptation is made even more difficult because meta-level processes are themselves problem-solving activity consuming significant resources (Barnett, 1984). The agent must, therefore, bound its efforts to select or construct a plan for accomplishing a resource-constrained task. Meta-level deliberation must leave resources for a suitable task-performance strategy to be enacted (Maes, 1986).

Theorists and experimenters in disciplines such as philosophy, cognitive science, AI, linguistics, and the decision sciences have only recently begun to focus on resource-constrained reasoning. Few investigators have focused on how, or even whether, theories of intelligent reasoning and action might require revision in order to reflect adaptation to limited time, information, cognitive capacity, or other resources. In the psychological literature, most models of intelligent problem-solving, decision-making, etc. make no provision for managing the tradeoffs and compromises required of a resource-constrained agent. There are at least two important counter-examples to this last claim. First, the research and theories on so-called selective attention address issues that are obviously quite similar to those we raise here. In addition, research on limited capacity cognitive mechanisms such as short-term memory or sensory buffer storage focus on important types of endogenous constraints and mechanisms that subjects rely upon in coping with them.

Preliminary investigations of resource-constrained agency suggest that standard conceptions of intelligent reasoning and action may require radical revision (Good, 1983; Fehling et al., 1989). Fehling and his colleagues are conducting in-depth analyses of important cases of resource-constrained agency. Employing the computational modeling perspective of AI and cognitive science, they conclude that a general "architecture" of a resource-constrained agent must include certain features not commonly found in agent architectures such as SOAR (Laird, Newell, & Rosenbloom, 1987), BB1 (Hayes-Roth, 1985), or ACT* (Anderson, 1976). For example, in spite of extensive psychological research on selective attention, none of these architectures provide "interruptability" — the a problem-solving agent's ability to detect the occurrence of critical events, suspend or cancel its on-going activities, and respond to these critical events. Nor do these architectures substantively manage tradeoffs among multiple, independent, and possibly competing tasks. The next section sketches a minimal architecture with some of these features, critically needed by a resource-constrained agent.

Fehling and other AI researchers have argued that that models of specific problem-solving processes of intelligent agents must also be radically revised. Very few problem-solving models provide for resource-constrained, adaptive management of problem-solving activity

itself. Fehling (Fehling & Breese, 1988) has begun formulating an approach to resource constrained problem-solving that incorporates concepts from formal decision theory. This work proposes ways that an agent can decide how and when expend resources to avail itself of information that might significantly change its beliefs or its commitments to future action. Our analysis of the functional role of repression in section 4 incorporates some results of this research. We rely there upon the idea we have just outlined — an intelligent agent will attempt to limit its actions, deliberations, and meta-level deliberations to those elements whose costs are likely to be outweighed by their likelihood of improving future conditions for the agent. This principle applies, in particular, to the way in which information is stored and retained by the agent. To summarize our position

> *Repression would naturally arise in barring from use any information whose costs of use outweigh their potential benefits. In our current architecture, this would specifically imply barring access to information from the GW. Information would be repressed rather than erased if it were sufficiently likely that this cost-benefit tradeoff might change in the future.*

3. A Functional Agent Architecture.

We now describe a "minimalist" architecture for an agent who must optimize its goal-directed actions within environmental and internal constraints. Our analysis of the functional role of repression will depend the assumption that a resource-constrained agent has a limited capacity component, such as a working memory, that is roughly equivalent to conscious access in human beings (Baars, 1988).

We are investigating a *global workspace* (GW) architecture. GW architectures have been explored for about twenty years, as general problem-solving frameworks for a variety of problem-solving tasks. Several psychologists suggest analogies between GW architectures and the human nervous system (e.g. Baars, 1988: Norman & Shallice, 1980; Bower & Cohen, 1982; Reason, 1983). Fig. 1 depicts the five elements of our architecture:

1. A "society" of *Specialists* — These modules each handle a particular sub-task,
2. A *Global Workspace* (GW) — The GW manages input from the agent's sensors and outputs of the agent's effectors. It also broadcasts input information to the society as a whole so that appropriate specialists can be "triggered" to react;
3. A collection of specialized *Sensors* and *Effectors* — Active specialists manage sensors and effector. Thus, a sensor (effector) modality will be operative if its managing specialists are active and non-operative otherwise;
4. A *Dominant Goal Hierarchy* (DGH) – The DGH specifies the current precedence hierarchy of goals and objectives to which the agent is currently commited. The DGH also controls access to the GW. Thus, input relevant to top level goals is given precedence over less significant input;
5. A *Background Information State* (BIS) — The BIS includes a collection of *context hierarchies* and *goal hierarchies*. Each context hierarchy encapsulates an internally consistent collection of information that models some aspects the agent's environment or itself. The goal hierarchies may, under the action of certain specialists, become the DGH.

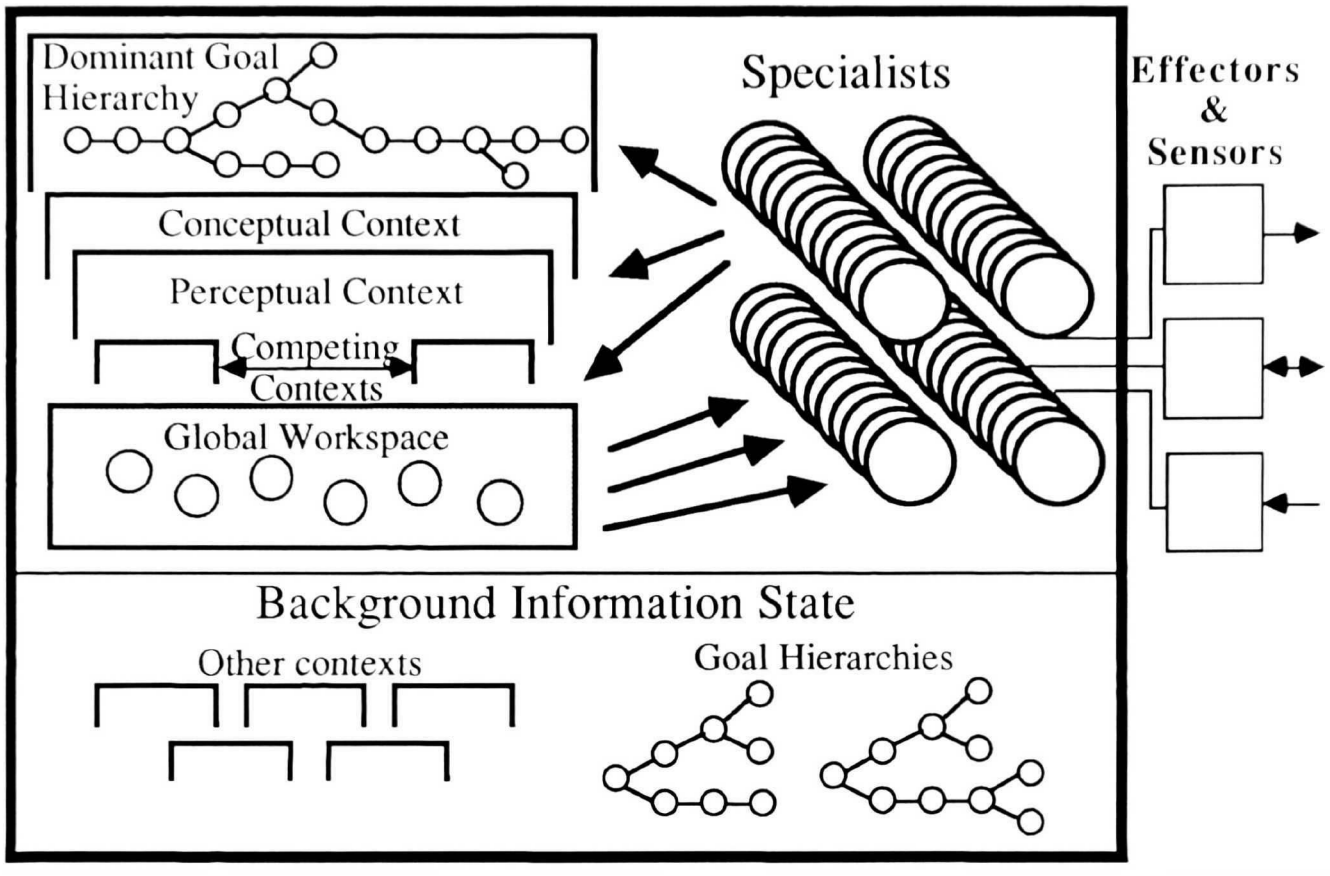

Figure 1 — An Architecture for a Resource-constrained Agent

Section two elucidated how an agent's actions must adapt to contextual constraints. In this light, an additional aspect of our agent architecture is especially important - the ability of the system to detect critical changes in the environment or in itself and react to such critical events. This "interruptability" provides our architecture with the means for coupling its actions to changes in its environment. Fehling, Altman, and Wilber (1989) discuss this aspect of the architecture in detail.

4. Repression as Information Management.

If the Global Workspace is analogous to consciousness or short-term memory (Baars, 1988), the Specialists are comparable to unconscious and isolated processors. Repression is functionally equivalent to purposeful exclusion of some potential GW input from the Global Workspace. Input that for humans provokes pain, shame, guilt, or anxiety has a functional analogue in GW messages that interrupt current GW contents by accessing very high-level goals (such as survival), and perhaps posing problems that demand a radical restructuring of the current Dominant Goal Hierarchy. The question then becomes, *from a pure systems-design perspective, is there a functional role for some mechanism like repression?* We are currently exploring two major possibilities:

1. *The need to avoid goals that have been outgrown, but which cannot be deleted.*

A robotic example. Early developmental stages of an autonomous agent may have created alarm situations (involving the goal of survival) that may no longer be realistic. Let us

suppose that we have used the cognitive architecture of Fig. 1 to implement an autonomous exploratory robot, *Alice, the autonomous agent.* Alice is exploring geological and environmental features of Mars for NASA. Upon her arrival, Alice constructed a base camp, consisting primarily of a large shelter to protect her from dust storms and high levels of solar radiation, and outfitted with solar arrays with which Alice recharge her batteries. When Alice first arrived, she was programmed with sketchy knowledge about the likelihood of danger from dust storms and radiation. In fact, she was programmed with a "cautious" goal hierarchy with higher priority elements requiring Alice to monitor carefully for conditions signalling possible dangers. Alice was also programmed with alternative goal hierarchies containing having high-priority requirements to complete experimental missions. In terms of the functional characteristics implied by Fig. 1, Alice's performance in any situation will be a function of the currently chosen DGH. To protect her from the unknown risks of her environment, Alice's designers provided her with a meta-level "rule" stipulating that the "cautious" goal hierarchy be selected as the default DGH. Operating under this DGH, much of Alice's activity is devoted to assessing possible danger, and, when any evidence of danger is noted, seeking shelter in base camp. Due to her excess caution, Alice found that she would seldom complete a task without many precautionary interruptions. After a long period of residency, Alice found that conditions did not warranted her seeking shelter. At this point Alice switched her goal hierarchy to one that specifies task completion as the highest-priority elements. Under this new DGH she would no longer regularly monitor for danger. She estimates, therefore, that she can devote more time and her own processing resources to completing her experiments. By calculating the expected value of continuing to operate cautiously, but inefficiently, and comparing that to the expected value of operating without precautionary actions, Alice determines that her expected productivity (including her chances of survival) will be higher when using the task-oriented goal hierarchy. This sort of deliberation is known in decision-theoretic terms as a "value-of-information" calculation. Alice has determined that the marginal value of precautionary data-gathering and action is negative in comparison to other modes of operation.

In terms of our architecture, Alice's switch in DGH requires barring data that could activate the cautious goal hierarchy from the GW. Otherwise, this hierarchy could supersede the new DGH. So, information and actions that could select the cautious goal hierarchy must be effectively repressed. Because the decision to repress was based on uncertain information, this goal hierarchy, and the information associated with it, should not be erased, however. Note also that Alice not only precludes the further use of the old goal hierarchy, but *avoids even testing* the relevance of the old goal hierarchy: Such tests would, after all, amount to using the same information-gathering commitments that she has determined not to be cost-effective.

A human example. For human infants, avoiding abandonment is a survival goal, and the infant is extraordinarily vigilant against any hint of rejection. If growing up involves adopting new, more mature, realistic, and self-confident DGHs, superseding the infantile DGH, then interruptions from the infantile DGH must be warded off. But, in ambiguous situations — and practically all social situations are profoundly ambiguous — the tendency to over-interpret input signals in the direction of the infantile DGH may take over again. In this case, *repression is effectively the attempt of the current DGH to guard against even momentary interruptions from the old, and currently non-Dominant Goal Hierarchy.*

But why not simply delete those parts of an outmoded goal hierarchy that have become irrelevant? The dilemma here is to determine when one can confidently delete goals. It is obviously important for an agent to attach probabilities to threats to its well-being. For example, although social rejection may be viewed as a threat to be avoided early in life, perhaps it could be deleted as a threat under evidence that it is irrelevant to well-being.

Indeed, becoming an adult involves falsification of certain childlike dependency goals. However, the threat of abandonment is, in fact, still present. Abandonment may again become a threat under circumstances which create a renewed state of helplessness. Since such circumstances are hard to predict, it would be risky to delete part of a goal hierarchy pertaining to abandonment. This inherent uncertainty may be heightened if tests of the old goal are avoided. Indeed, adult repression often exists for avoidant goals that are never tested and hence never falsified. So, to-be-avoided goals may be given a much higher probability than they actually have. Psychodynamic and cognitive-behavioral researchers have found that many people fear rejection to the point of never testing the reality of rejection, thus also losing the benefits of acceptance and social support (Erdelyi, 1985).

2. *The need to maintain a valued self-concept, i.e., one that is at or near the top of the goal hierarchy.*

A second reason to exclude material from the GW is to maintain a self-concept that ranks the self among the highest values of the hierarchy. Since GW information is widely distributed, it can also be used to revise the goal hierarchy itself. Thus, if we display GW information that shows a high-level goal to be less worthwhile than previously thought, such a goal may be pushed down in the goal hierarchy. Humans are well-known to say things like, "I'll just die from shame if I fail this exam, if I look ridiculous in front of my friends, if I act like a baby," etc. In all these cases, the value of the self is subordinated to a goal. Flexibility in the assignment of values to goals is of obvious importance, for example, in being able to make use of unanticipated opportunities to achieve goals that are not currently on top; or, if a major inconsistency is discovered between two goals, to be able to devalue one of the two goals, so as to achieve at least one. It is important to be able to value a goal nearly as much, or perhaps more than life itself. There is extensive psychological evidence from studies of depression and other disorders that people consciously under-value themselves compared to other goals quite often.

Two contrary tendencies contend. In order to perform major goals, it is useful to broadcast GW message that will place the most important goals high in the goal hierarchy. But, if we place current goals higher than self value, the system may risk self-destruction. *Repression may then protect the global workspace against messages that would serve to posit goals as more valued than the self.* From this point of view, the value of the self-concept must reside at or near the top in the goal-hierarchy, so that it can work to exclude material from the GW that will tend to devalue the self.[2]

5. Conclusions.

We have presented a model of problem-solving by intelligent, *resource constrained agents*. Our *global workspace* architecture appears to meet the minimal requirements for such an agent. We have argued that a resource-constrained agent must bar certain information from its GW, while simultaneously preserving the information in a sequestered manner for use under conditions that might arise in the future. This active exclusion of information from the global workspace corresponds to the idea of repression defined by Freud (1915) as "turning something away, and keeping it at a distance form the conscious." Repression, thus defined, is a necessary aspect of information management in *resource constrained agents* in general and in human beings in particular. While our reasoning differs from that leading to the notion of repression in psychoanalysis, our conclusion is similar. Indeed, contemporary psychoanalytic theory (e.g., Brenner 1982) regards repression as a general

[2]Recall that the DGH has both the function of controlling access to the GW and of maintaining goals for the system as a whole.

mental capacity distinct from a notion of psychopathology. We view repression as a functional concept with profound implications for the behavior of intelligent systems.

We have offered an "argument by design," that seeks to avoid controversies regarding empirical evidence for repression. This approach supports repression's plausibility by (a) proposing plausible designs, or "computational architectures," of intelligent agents, and (b) exploring a design for the role, if any, of repression. The cognitive architecture described in this paper is derived from extensive research by Baars, Fehling, and their colleagues. It is the only one of which we are aware that is explicitly formulated to address the issues of resource-constrained agency. Within this architecture, a mechanism like repression arises naturally and plays an important functional role -- that is, it appears to provide a direct, efficient way for to handle a category of highly likely information management problems.

References.

Anderson, J.R. *Language, Memory, and Thought.* Hillsdale, N.J.: Erlbaum and Associates, 1976.

Baars, B. *A Cognitive Theory of Consciousness.* New York: Cambridge University Press, 1988.

Barnett, J. "How much is control knowledge worth?" *Artificial Intelligence*, 1984.

Bower, G.H. & Cohen, P.R. "Emotional influences in memory and thinking." In M.S. Clark & S.T. Fiske (Eds.), *Attention and Cognition.* Hillsdale, N.J.: Erlbaum & Associates, 1982, pp. 291-331.

Brenner, The Mind in Conflict.

D'Ambrosio, B., Fehling, M.R., Forrest, S., Raulefs, P. & Wilber, B.M.

Erdelyi, M.H. "A new look at the New Look: Perceptual Defense and Vigilance", *Psychological Review*, 81, 1985, pp. 1-25.

Fehling, M.R., Altman, A. & Wilber, B.M. "The HCVM: An Instance of Schemer, an Architecture of Resource-constrained Problem-Solving." In

Fehling, M.R. & Russell, S. *Proceedings of the AAAI 1989 Symposium on Limited Rationality.* American Association of Artificial Intelligence, Menlo Park, CA, March, 1989.

Freud, S. Repression. In James Strachey (Ed.), Standard Edition of the Complete Psychological Works of Sigmund Freud, Vol. XIV, pp. 141-158, London, The Hogarth Press, 1957 (originally published in 1915).

Good, I.J. *Good Thinking: The Foundations of Probability and Its Applications.* Minneapolis: University of Minnesota Press, 1983.

Hayes-Roth, B. "A Blackboard Architecture for Control." *Artificial Intelligence,* 1985, 26(3), pp. 251-321.

Holmes, D. "Closure in a gapped circle figure." *American Journal of Psychology*, 80, 1967, pp. 614-618.

Laird, J.E., Newell, A., & Rosenbloom, P.S. "SOAR: An Architecture for General Intelligence." *Artificial Intelligence,* 1987, 33(1).

Maes, P. *Proceedings of the Workshop on Meta-level Architectures and Reflection,* Sardinia, 1986.

Norman, D.A. & Shallice, T. "Attention and action: Willed and automatic control of behavior." Unpublished manuscript. Center for Information Processing, UCSD, La Jolla, CA, 1980.

Reason, J. "Absent-mindedness and cognitive control." In J. Harris & P. Morris (Eds.) *Everyday Actions, Memory, and Absent-mindedness.* New York: Academic Press, 1983.

Simon, H. *The Sciences of the Artificial.* 2nd Ed. M.I.T. Press. Cambridge, Mass, 1981.