Introduction

What does it mean for a machine to be autonomous? Has any progress been made towards autonomous machines since Grey Walter's famous *M. Speculatrix*† (Walter, 1953)? In a narrow sense it is clear that there has, as evidenced by the evolution of the *M. Labyrinthea* species (of which Claude Shannon constructed an early example) into the fleet-footed trial-and-error goal seeking devices seen in successive generations of the IEEE Micromice competition. However, these devices have a predictable course and a predestined end, providing an excellent example of the old argument against artificial intelligence that "reliable computers do only what they are instructed to do". In this paper we seek autonomy in some deeper sense.

It is not surprising that dictionary definitions of autonomy concentrate on natural systems. According to the Oxford dictionary, it has two principal strands of meaning:

Autonomy 1. Of a state, institution, etc
- **a** The right of self-government, of making its own laws and administering its own affairs
- **b** Liberty to follow one's will, personal freedom
- **c** Freedom (of the will): the Kantian doctrine of the Will giving itself its own law, apart from any object willed; opposed to *heteronomy*

2. *Biol.* autonomous condition
- **a** The condition of being controlled only by its own laws, and not subject to any higher one
- **b** Organic independence

Our interest here lies in practical aspects of autonomy as opposed to philosophical ones. Consequently we will steer clear of the debate on free will and what it means for machines, simply noting in passing that some dismiss the problem out of hand. For instance, Minsky (1961) quotes with approval McCulloch (1954) that our *freedom of will* "presumably means no more than that we can distinguish between what we intend (ie our *plan*), and some intervention in our action"‡. We also refrain from the potentially theological considerations of what is meant by "higher" laws in the second part.

How can we interpret what is left of the definition? In terms of modern AI, the first meaning can best be read as self-government through goal-seeking behavior, setting one's own goals, and choosing which way to pursue them. The second meaning, organic independence, has been the subject of major debate in the biological and system-theoretic community around the concepts of "homeostasis" and, more recently, "autopoiesis".

Our search in this paper will pursue these strands separately. Goals and plans have received much attention in AI, both from the point of view of understanding (or at least explaining) stories involving human goals and how they can be achieved or frustrated, and in purely artificial systems which learn by discovery. Biologists and psychologists have studied goal-seeking behavior in people, and come to conclusions which seem to indicate remarkable similarities with the approach taken by current AI systems to setting and pursuing goals. On the other side of the coin, there are strong arguments that these similarities should be viewed with a good deal of suspicion.

The second strand of meaning, organic independence, has not been contemplated explicitly in mainstream computer science. There have been a number of well-known developments on the periphery of the subject which do involve self-replicating organisms. Examples include games such as "life" (Berlekamp *et al*, 1982) and "core wars" (Dewdney, 1984), as well as cellular (eg Codd, 1968), self-reproducing (eg von Neumann, 1966), and evolutionary (eg Fogel *eg al*, 1966) automata. However, these seem artificial and contrived examples of autonomy. In contrast, some autonomous systems have recently arisen naturally in computer software. We examine the system-theoretic idea of "autopoiesis" and then look at these software developments in this context.

† for the discerning, or "tortoise" for the profane, as its inventor took pains to point out.

‡ This seems to endow free will to a Micromouse which, having mapped the maze, is following its plan the second time round when it finds a new obstacle!

Goal-seeking — artificial and natural

In a discussion of robots and emotions, Sloman and Croucher (1981) note that many people deny that machines could ever be said to have their own goals. "Machines hitherto familiar to us either are not goal-directed at all (clocks, etc) or else, like current game-playing computer programs, have a simple hierarchical set of goals, with the highest-level goal put there by a programmer". They postulate that robots will need *motive generators* to allow them to develop a sufficiently rich structure of goals; unfortunately they do not say how such generators might work. To exemplify how goals are used in existing AI programs, we will briefly review two lines of current research.

Examples of artificial goal-seeking. Those working on conceptual dependency in natural language understanding have long recognized that stories cannot be understood without knowing about the goal-seeking nature of the actors involved. Schank & Abelson (1977) present a taxonomy of human goals, noting that different attempts at classification present a confusing array of partially overlapping constructs and suggesting that some future researcher might succeed in bringing order out of the chaos using methods such as cluster analysis. They postulate the following seven goal forms:

- Satisfaction goal — a recurring strong biological need
  Examples: *hunger, sex, sleep*
- Enjoyment goal — an activity which is optionally pursued for enjoyment or relaxation
  Examples: *travel, entertainment, exercise* (in addition, the activities implied by some satisfaction goals may alternatively be pursued primarily for enjoyment)
- Achievement goal — the realization (often over a long term) of some valued acquisition or social position
  Examples: *possessions, good job, social relationships*
- Preservation goal — preserving or improving the health, safety, or good condition of people, position, or property
  Examples: *health, good eyesight*
- Crisis goal — a special class of preservation goal set up to handle serious and imminent threats.
  Examples: *fire, storm*
- Instrumental goal — occurs in the service of any of the above goals to realize a precondition
  Examples: *get babysitter*
- Delta goal — similar to instrumental goal except that general planning operations instead of scripts are involved in its pursuit
  Examples: *know, gain-proximity, gain-control.*

The first three involve striving for desired states; the next two, avoidance of undesired states; the last two, intermediate subgoals for any of the other five forms. Programs developed within this framework "understand" (ie can answer questions about) stories involving human actors with these goals (eg Wilensky, 1983; Dyer, 1983). For example, if John goes to a restaurant it is likely that he is attempting to fulfill either a satisfaction goal or an entertainment goal (or both). Instrumental or delta goals will be interpreted in the context of the prevailing high-level goal. If John takes a cab to the restaurant it will be understood that he is achieving the delta goal *gain-proximity* in service of his satisfaction or entertainment goal.

Our second example of goal usage in contemporary AI is Lenat's "discovery" program AM, and its successor EURISKO (Davis & Lenat, 1982; Lenat *et al*, 1982). These pursue interesting lines of research in the domains of elementary mathematics and VLSI design heuristics, respectively. They do this by exploring concepts — producing examples, generalizing, specializing, noting similarities, making plausible hypotheses and definitions, etc. The programs evaluate these discoveries for utility and "interestingness," and add them to the vocabulary of concepts. They essentially perform exploration in an enormous search space, governed by heuristics which evaluate the results and suggest fruitful avenues for future work.

Each concept in these systems is represented by a frame-like data structure with dozens of different facets or slots. For example, the types of facets in AM include

- examples
- definitions
- generalizations
- domain/range
- analogies
- interestingness.

Heuristics are organized around the facets. For example, the following strategy fits into the *examples* facet of the *predicate* concept:

> If, empirically, 10 times as many elements *fail* some predicate P as *satisfy* it, then some *generalization* (weakened version) of P might be more interesting than P.

AM considers this suggestion after trying to fill in examples of each predicate. For instance, when the predicate SET-EQUALITY is investigated, so few examples are found that AM decides to generalize it. The result is the creation of a new predicate which means HAS-THE-SAME-LENGTH-AS — a rudimentary precursor to the discovery of natural numbers.

In an unusual and insightful retrospective on these programs, Lenat & Brown (1984) report that the exploration consists of (mere?) syntactic mutation of programs expressed in certain representations. The key element of the approach is to find representations with a high density of interesting concepts so that many of the random mutations will be worth exploring. If the representation is not well matched to the problem domain, most explorations will be fruitless and the method will fail.

While the conceptual dependency research reviewed above is concerned with understanding the goals of actors in stories given to a program, the approach taken seems equally suited to the construction of artificial goal-oriented systems. If a program could really understand or empathize with the motives of people, it seems a small technical step to turn it around to create an autonomous simulation with the same motivational structure. Indeed, one application of the conceptual dependency framework is in *generating* coherent stories by inventing goals for the actors, choosing appropriate plans, and simulating the frustration or achievement of the goals (Meehan, 1977). The "learning by discovery" research shows how plausible subgoals can be generated from an overall goal of maximizing the interestingness of the concepts being developed. It is worth noting that Andreae (1977) chose a similar idea, "novelty," as the driving force behind a very different learning system. Random mutation in an appropriate representation seems to be the closest we have come so far to the *motive generator* mentioned at the beginning of this section.

The mechanism and psychology of natural goal-seeking. Now turn to natural systems. The objection to the above-described use of goals in natural language understanders and discovery programs is that they are just programmed in. The computer only does what it is told. In the first case, it is told a classification of goals and given information about their interrelationships, suitable plans for achieving them, and so on. In the second case it is told to maximize interestingness by random mutation. On the surface, these seem to be a pale reflection of the autonomous self-government of natural systems. But let us now look at how goals seem to arise in natural systems.

The eminent British anatomist J.Z. Young describes the modern biologist's highly mechanistic view of the basic needs of animals. "Biologists no longer believe that living depends upon some special non-physical agency or spirit," he avers (Young, 1978, p. 13), and goes on to claim that we now understand how it comes about that organisms behave as if all their actions were directed towards an aim or goal†. The mechanism for

† Others apparently tend to be more reticent — "it has been curiously unfashionable among biologists to call attention to this charac-

this is the reward system situated in the hypothalamus. For example, the cells of the hypothalamus ensure that the right amount of food and drink are taken and the right amount is incorporated to allow the body to grow to its proper size. These hypothalamic centers stimulate the need for what is lacking, for instance of food, sex, or sleep, and they indicate satisfaction when enough has been obtained. Moreover, the mechanism has been traced to a startling level of detail. For example, Young describes how hypothalamic cells can be identified which regulate the amount of water in the body.

> The setting of the level of their sensitivity to salt provides the instruction that determines the quantity of water that is held in the body. We can say that the properties of these cells are physical symbols "representing" the required water content. They do this in fact by actually swelling or shrinking when the salt concentration of the blood changes.
>
> Young, 1978, p. 135

Food intake is regulated in the same way. The hypothalamus ensures propagation of the species by directing reproductive behavior and, along with neighboring regions of the brain, attends to the goal of self-preservation by allowing us to defend ourselves if attacked.

Needless to say, experimental evidence for this is obtained primarily from animals. Do people's goals differ? The humanistic psychologist Abraham Maslow propounded a theory of human motivation that distinguishes between different kinds of needs (Maslow, 1954). *Basic needs* include hunger, affection, security, love, and self-esteem. *Metaneeds* include justice, goodness, beauty, order, and unity. Basic needs are arranged in a hierarchical order so that some are stronger than others (eg security over love); but all are generally stronger than metaneeds. The metaneeds have equal value and no hierarchy, and one can be substituted for another. Like the basic needs, the metaneeds are inherent in man, and when they are not fulfilled, the person may become psychologically sick (suffering, for example, from alienation, anguish, apathy, or cynicism).

In his later writing, Maslow (1968) talks of a "single ultimate value for mankind, a far goal towards which all men strive". Although going under different names (Maslow favors *self-actualization*), it amounts to "realizing the potentialities of the person, that is to say, becoming fully human, everything that the person *can* become". However, the person does not know this. As far as he is concerned, the individual needs are the driving force. He does not know in advance that he will strive on after the current need has been satisfied. Maslow produced the list of personality characteristics of the psychologically healthy person shown in Table 1.

Maslow's *basic needs* seem to correspond reasonably closely with those identified by conceptual dependency theory. Moreover, there is some similarity to the goals mentioned by Young (1978), which, as we have seen, are thought to be "programmed in" to the brain in an astonishingly literal sense. Consequently it is not clear how programs in which these goals are embedded differ in principle from goal-oriented systems in nature. The *metaneeds* are more remote from current computer systems, although there have been shallow attempts to simulate paranoia in the PARRY system (Colby, 1973). It is intriguing to read Table 1 in the context of self-actualized computers! Moreover, one marvels at the similarity between the single-highest-goal model of people in terms of self-actualization, and the architecture for discovery programs sketched earlier in terms of a quest for "interestingness".

The sceptical view. The philosopher John Haugeland addressed the problem of natural language understanding and summed up his viewpoint in the memorable aphorism,"the trouble with Artificial Intelligence is that computers don't give a damn" (Haugeland, 1979). He identified four different ways in which brief segments of text cannot be understood "in isolation", which he called four *holisms*. Two of these, concerning *common-sense knowledge* and *situational knowledge*, are the subject of intensive research in natural language analysis systems. Another, the *holism of intentional interpretation*, expresses the requirement that utterances and descriptions "make sense" and seems to be at least partially addressed by the goal/plan orientation of some natural language systems. It is the fourth, called *existential holism*, that is most germane to the present topic.

---

teristic of living things" (Young, 1978, p. 16).

- They are realistically oriented.
- They accept themselves, other people, and the natural world for what they are.
- They have a great deal of spontaneity.
- They are problem-centered rather than self-centered.
- They have an air of detachment and a need for privacy.
- They are autonomous and independent.
- Their appreciation of people and things is fresh rather than stereotyped.
- Most of them have had profound mystical or spiritual experiences although not necessarily religious in character.
- They identify with mankind.
- Their intimate relationships with a few specially loved people tend to be profound and deeply emotional rather than superficial.
- Their values and attitudes are democratic.
- They do not confuse means with ends.
- Their sense of humor is philosophical rather than hostile.
- They have a great fund of creativeness.
- They resist conformity to the culture.
- They transcend the environment rather than just coping with it.

Table 1: Characteristics of self-actualized persons (Maslow, 1954)

Haugeland argues that one must have actually *experienced* emotions (like embarrassment, relief, guilt, shame) to understand "the meaning of text that (in a familiar sense) *has* any meaning". One can only experience emotions in the context of one's own self-image. Consequently, Haugeland concludes that "only a being that cares about who it is, as some sort of enduring whole, can care about guilt or folly, self-respect or achievement, life or death. And only such a being can read." Computers just don't give a damn.

As AI researchers have pointed out repeatedly, however, it is difficult to give such arguments *operational* meanings. How could one test whether a machine has *experienced* an emotion like embarrassment? If it acts embarrassed, isn't that enough? And while machines cannot yet behave convincingly as though they do experience emotions, it is not clear that fundamental obstacles stand in the way of further and continued progress. There seems to be no reason in principle why a machine cannot be given a self-image.

This controversy has raged back and forth for decades, a recent resurgence being Searle's (1980) paper on the Chinese room *gedanken* experiment, and the 28 responses which were printed with it. Searle states the thesis succinctly: "such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output". And the antithesis could be caricatured as "maybe, but does it *matter?*". Those who find the debate frustrating can always, with Sloman & Croucher (1981), finesse the issue: "Ultimately, the decision whether to say such machines have motives is a *moral* decision, concerned with how we ought to treat them".

Autopoiesis — natural and artificial

Autonomy is a striking feature of biological systems. Consequently biologists have made attempts to articulate what it means to them; to pin it down, formalize and study it in a system-theoretic context. However, this work is obscure and difficult to assess in terms of its predictive power (which must be the fundamental test of any theory). Even as a descriptive theory its use is surrounded by controversy. Consequently this section attempts to give the flavor of the endeavor, relying heavily on quotations from the major participants in the research, and goes on to describe some practical computer systems which appear to satisfy the criteria biologists have identified for autonomy.

Homeostasis. People have long expressed wonder at how a living organism maintains its identity in the face of continuous change.

> In an open system, such as our bodies represent, compounded of unstable material and subjected continuously to disturbing conditions, constancy is in itself evidence that agencies are acting or ready to act, to maintain this constancy.
>
> Cannon, 1932

Following Cannon, Ashby (1960) developed the idea of "homeostasis" to account for this remarkable ability to preserve stability under conditions of change. The word has now found its way into North American dictionaries, eg Webster's

> Homeostasis is the tendency to maintain, or the maintenance of, normal, internal stability in an organism by coordinated responses of the organ systems that automatically compensate for environmental changes.

The basis for homeostasis was adaptation by the organism. When change occurred, the organism adapted to it and thus preserved its constancy.

> A form of behavior is *adaptive* if it maintains the essential variables within physiological limits.
>
> Ashby, 1960, p. 58

The "essential variables" are closely related to survival and linked together dynamically so that marked changes in any one soon lead to changes in the others. Examples are pulse rate, blood pressure, body temperature, number of bacteria in the tissue, etc. Ashby went so far as to construct an artifact, the "Homeostat", which exhibits this kind of ultrastable equilibrium.

Homeostasis emphasizes the stability of biological systems under external change. Recently, a concept called "autopoiesis" has been identified, which captures the essence of biological autonomy in the sense of stability or preservation of identity under *internal* change (Maturana, 1975; Maturana & Varela, 1980; Varela, 1979; Zeleny, 1981). This has aroused considerable interest, and controversy, in the system theoretic research community.

Autopoiesis. The neologism "autopoiesis" means literally "self-production," and a striking example occurs in living cells. These complex systems produce and synthesize macromolecules of proteins, lipids, and enzymes, and consist of about 10^5 macromolecules. The entire population of a given cell is renewed about 10^4 times during its lifetime (Zeleny, 1981a). Despite this turnover of matter, the cell retains its distinctiveness and cohesiveness — in short, its *autonomy*. This maintenance of unity and identity of the whole, despite the fact that all the while components are being created and destroyed, is called "autopoiesis". A concise definition is

> Autopoiesis is the capability of living systems to develop and maintain their own organization. The organization that is developed and maintained is identical to that performing the development and maintenance.
>
> Andrew, 1981, p. 156

Other authors (eg Maturana & Varela, 1980; Zeleny, 1981a) add a corollary:

> a topological boundary emerges as a result of the processes [of development and maintenance].
>
> Zeleny, 1981a, p. 6

This emphasizes the train of thought "from self-production to identity" that seems to underly much of the autopoietic literature.

Operating as a system which produces or renews its own components, an autopoietic system continuously regenerates its own organization. It does this in an endless turnover of components and despite inevitable perturbations. Therefore autopoiesis is a form of homeostasis which has its own organization as the fundamental variable which remains constant. The principal fascination of the concept lies in the self-reference it implies, This has stimulated a theoretical formulation of the notion of circularity or self-reference in Varela's (1975) extension of Brown's "calculus of distinctions" (Brown, 1969). Along with other work on self-reference (eg Hofstadter, 1979), this has an esoteric and obscure, almost mystical, quality. While it may yet form the basis of a profound paradigm shift in systems science, it is currently surrounded by controversy and its potential contribution is quite unclear (Gaines, 1981). Indeed, it has been noted that an "unusual degree of parochialism, defensiveness, and quasi-theological dogmatism has arisen around autopoiesis" (Jantsch, 1981).

There has been considerable discussion of the relation between autopoiesis and concepts such as purpose and information. Varela (1979) claims that "notions [of teleology and information] are unnecessary for the *definition* of the living organization, and that they belong to a descriptive domain distinct from and independent of the domain in which the living system's *operations* are described" (p. 63/64). In other words, nature is not about goals and information; we observers invent such concepts to help classify what we see. Maturana (1975) is more outspoken: "descriptions in terms of information transfer, coding and computations of adequate states are fallacious because they only reflect the observer's domain of purposeful design and not the dynamics of the system as a state-determined system"; presumably goals are included too in the list of proscribed terms. Some have protested strongly against this hard-line view — which is particularly provocative because of its use of the word "fallacious" — and attempted to reconcile it with "the fact that the behavior of people and animals is very readily and satisfactorily described in terms of goals and attempts to achieve them" (Andrew, 1981, p. 158). In his more recent work Varela (1981) diverged further from the hard-line view, explaining that he had intended to criticize only "the *naive* use of information and purpose as notions that can enter into the definition of a system on the same basis as material interactions" [his emphasis]. He concluded that "autopoiesis, as an operational explanation, is not quite sufficient for a full understanding of the phenomenology of the living, and that it needs a carefully constructed complementary symbolic explanation". For Varela, a symbolic explanation is one that is based on the notions of information and purpose. It is clear, though, that while some allow that autopoiesis can *coexist* with purposive interpretations, it will not *contribute* to them.

Is autopoiesis restricted to *living* systems? Some authors find it attractive to extend the notion to the level of society and socio-political evolution (eg Beer, 1980; Zeleny, 1977). Others (eg Varela, 1981) stress the renewal of components through material self-production and restrict autopoiesis to chemical processes. Without self-production in a material sense, the support for the corollary above becomes unclear, and consequently the whole relevance of autopoiesis to identity and autonomy comes under question.

Artificial autopoiesis. Although one can point to computer simulations of very simple autopoietic systems (eg Varela *et al*, 1974; Zeleny, 1978; Uribe, 1981), there seems to have been little study of artificially autopoietic systems in their own right. However there are examples of computer systems which are autopoietic and which have arisen "naturally", that is to say, were developed for other purposes and not as illustrations of autopoiesis. It is probably true that in each case the developers were entirely unaware of the concept of autopoiesis and the interest surrounding it in system theory circles.

Worm programs were an experiment in distributed computation (Shoch & Hupp, 1982). The problem they addressed was to utilize idle time on a network of interconnected personal computers without any impact on normal use. It was necessary to be able to redeploy or unplug any machine at any time without warning. Moreover, in order to make the system robust to any kind of failure, power-down or "I am dying" messages were not employed in the protocol. A "worm" comprises multiple "segments", each running on a different machine. Segments of the worm have the ability to replicate themselves in idle machines. All segments remain in communication with each other, thus preserving the worm's identity and distinguishing it from a collection of independent processes; however, all segments are peers and none is in overall control. To prevent uncontrolled reproduction, a certain number of segments is pre-specified as the target size of the worm. When a segment is corrupted or killed, its peers notice the fact because it fails to make its periodical "I am alive" report. They then proceed to search for an idle machine and occupy it with another segment. Care is taken to coordinate this activity so that only one new segment is created.

There are two logical components to a worm. The first is the underlying worm maintenance mechanism, which is responsible for maintaining the worm — finding free machines when needed and replicating the program for each additional segment. The second is the application part, and several applications have been investigated (Shoch & Hupp, 1982), such as

- *existential* worm that merely announces its presence on each computer it inhabits;
- *billboard* worm that posts a graphic message on each screen;
- *alarm clock* worm that implements a highly reliable alarm clock that is not based on any particular machine;
- *animation* worm for undertaking lengthy computer graphics computations.

Can worms shed any light on the controversies outlined above which surround the concept of autopoiesis? Firstly, although they are not living and do not create their own material in any chemical sense, they are certainly autonomous, autopoietic systems. Shoch & Hupp relate how

> a small worm was left running one night, just exercising the worm control mechanism and using a small number of machines. When we returned the next morning, we found dozens of machines dead, apparently crashed. If one restarted the regular memory diagnostic, it would run very briefly, then be seized by the worm. The worm would quickly load its program into this new segment; the program would start to run and promptly crash, leaving the worm incomplete — and still hungrily looking for new segments.

John Brunner's science fiction story *The shockwave rider* presaged just such an uncontrollable worm. Of course, extermination is always possible in principle by switching off or simultaneously rebooting every machine on the network, although this may not be an option in practice. Secondly, in the light of our earlier discussion of teleology and autopoiesis, it is interesting to find the clear separation of the maintenance mechanism — the autopoietic part — from the the application code — the "purposive" part — of the worm. It can be viewed quite separately as an autopoietic or an application (teleological?) system.

Self-replicating Trojan horses. In his Turing Award lecture, Thompson (1984) raised the specter of ineradicable programs residing within a computer system — ineradicable in the sense that although they are absent from all source code, they can survive recompilation and reinstallation of the entire system! Most

people's reaction is "impossible! — it must be a simple trick", but Thompson showed a trick that is extremely subtle and sophisticated, and effectively impossible to detect or counter. The natural application of such a device is to compromise a system's security, and Thompson's conclusion was that there can be no technical substitute for natural trust. From a system-theoretic viewpoint, however, this is an interesting example of how a parasite can survive despite all attempts by its host to eliminate it.

To understand what is involved in creating such an organism, consider first self-replicating programs. When compiled and executed, these print out themselves (say in source code form); no more and no less. Although at first sight they seem to violate some fundamental intuitive principle of information — that to print oneself one needs *both* "oneself" *and, in addition*, something to print it out, this is not so. Programmers have long amused themselves with self-replicating programs, often setting the challenge of discovering the shortest such program in any given computer language. Moreover, it is easy to construct a self-replicating program that includes any given piece of text. Such a program divides naturally into the self-replicating part and the part that is to be reproduced, in much the same way that a worm program separates the worm maintenance mechanism from the application part.

View self-replication as a source program "hiding" in executable binary code. Normally when coaxed out of hiding it prints itself. But imagine one embedded in a language compiler, which when activated interpolates itself into the input stream for the compiler, causing itself to be compiled and inserted into the binary program being produced. Now it has transferred itself from the executable version of the compiler to the executable version of the program being compiled — without ever appearing in source form. Now imagine that the program being compiled is itself the compiler — a virgin version, uncorrupted in any way. Then the self-replicating code transfers itself from the old version of the compiler to the new version, without appearing in source form. It remains only for the code to detect when it is the compiler that is being recompiled, and not to interfere with other programs. This is well known as the standard Trojan Horse technique. The result is a bug that lives only in the compiled version and replicates itself whenever the compiler is recompiled.

If autopoiesis is the ability of a system to develop and maintain its own organization, the self-replicating Trojan horse seems to be a remarkable example of it. It is an organism that it extremely difficult to destroy, even when one has detected its presence. However, it cannot be autonomous, but rather survives as a parasite on a language compiler. It does not have to be a compiler: any program that handles other programs (including itself) will do†. Although presented as a pathological example of computer use, it is possible to imagine non-destructive applications — such as permanently identifying authorship or ownership of installed software even though the source code is provided. In the natural world, parasites can have symbiotic relationships with their hosts. It would be interesting to find analogous circumstances for self-replicating Trojan horses, but I do not know of any — these examples of benevolent use do not seem to benefit the host program directly, but rather its author or owner.

Viruses are perhaps less subtle but more pervasive kinds of bugs. They spread infection in a computer system by attaching themselves to files containing executable programs. The virus itself is a small piece of code which gains control whenever the host is executed, performs its viral function, and then passes control to the host. Generally the user is unaware that anything unusual is happening: as far as he is concerned, the host program executes exactly as normal‡. As part of its function, a virus spreads itself. When it has control, it may attach itself to one or several other files containing executable programs, turning them into viruses too. Under most computer protection schemes, it has the unusual advantage of running with the privileges of the person who invoked the host, not with the privileges of the host program itself. Thus it has a unique opportunity to infect other files belonging to that person. In an environment where people sometimes use each others programs, this allows it to spread rapidly throughout the system.

† As Thompson (1984) remarks, a well-installed microcode bug will be almost impossible to detect.

‡ The only difference is a small startup delay which probably goes unnoticed.

Unlike self-replicating Trojan horses, a virus can be killed by recompiling the host. (Of course, there is no reason why a virus should not be dispatched to install a self-replicating Trojan horse in the compiler.) If all programs are recompiled "simultaneously" (ie without executing any of them between compilations), the virus will be eradicated. However, in a multi-user system it is extremely hard to arrange for everyone to arrange a massive recompilation — in the same way as it is difficult to reboot every machine on a network simultaneously to stamp out a worm.

Viruses do not generally remain in touch with each other and therefore, unlike worms, are not really autopoietic. But there is no intrinsic reason why they should not be. They provide a basic and effective means of reproduction which could be utilized for higher-level communicating systems. As with the other devices reviewed above, when one hears about viruses one cannot help thinking of pathological uses. However, there are benevolent applications. They could assist in system maintenance by recording how often programs were used and arranging optimization accordingly, perhaps migrating little-used ones to slower memory devices or arranging optimization of frequently-used programs. Such reorganizations could take place without users being aware of it, quietly making the overall system more efficient.

Conclusions

We have examined two rather different directions in which autonomy can be pursued in computer systems. The first concerns representation and manipulation of goals. Examination of some current AI systems shows that they do not escape the old criticism that their goals and aspirations are merely planted there by the programmer. Indeed, it is not easy to see how it could be different, unless goals were generated randomly in some sense. Random exploration is also being investigated in current AI systems, and these show that syntactic mutation can be an extremely powerful technique when combined with semantically dense representations.

But according to modern biological thinking, the lower-level goals of people and animals are also implanted in their brains in a remarkably literal sense. Higher-level goals are not so easy to pin down. According to one school of psychological thought they stem from a single "super-goal" called self-actualization. This is remarkably in tune with the architecture of some prominent discovery programs in AI which strive to maximize the "interestingness" of the concepts being developed. While one certainly cannot equate self-actualization with interestingness, the resemblance is nevertheless striking.

The second direction concerns organizational independence in a sense of wholeness which is distinct from goal-seeking. The concept of autopoiesis formalizes this notion. Organizational independence can be identified in certain computer systems like worm programs, self-replicating Trojan horses, and viruses. It is remarkable that such applications have been constructed because they offer practical advantages and not in pursuit of any theoretical investigation of autonomy; in this way they are quite different from contrived games. In some sense self-replicating programs do have a goal, namely *survival*. A damaged worm exhibits this by repairing itself. But this is a weak form of goal-seeking compared with living organisms, which actively sense danger and take measures to prevent their own demise.

The architecture of these systems is striking in that the mechanism which maintains the artificial organism (be it the worm maintenance code, the self-replicating part of a Trojan horse, or the viral infection-spreader) is quite separate from the application part of the organism. Most people think of such programs as somehow pathological, and the application as a harmful or subversive one, but this need not be so: there are benign examples of each. In any case, separation of the organism's maintenance from its purpose is interesting because the concept of autopoiesis has sparked a debate in system-theoretic circles as to whether teleological descriptions are even legitimate, let alone necessary. In both domains a clear separation seems to arise naturally between the autopoietic and teleological view of organisms.

There have been no attempts to build computer programs which combine these two directions. The AI community which developed techniques of goal-seeking has historically been somewhat separate from the system software community which has created robust self-replicating programs like worms and viruses. What will spring from the inevitable combination and synthesis of the two technologies of autonomy?

Acknowledgements

First and foremost I would like to thank the Conference Intelligent Computer for suggesting the topic of this paper and for pressing me into writing it; it is worth pointing out that the opinions expressed in the paper are not necessarily those of the author. I am grateful to Saul Greenberg and Roy Masrani for many insights into topics discussed in this paper, and to Bruce MacDonald for making some valuable suggestions. This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Andreae, J.H. (1977) *Thinking with the teachable machine.* Academic Press, London.

Andrew, A.M. (1981) "Autopoiesis — allopoiesis interplay" in *Autopoiesis: a theory of living organization,* edited by M.Zeleny, pp 157-166. North Holland, New York.

Ashby, W.R. (1960) *Design for a brain: the origin of adaptive behavior.* Wiley, New York, (second edition).

Beer, S. (1980) Preface to *Autopoiesis and cognition* (Maturana & Varela, 1980).

Berlekamp, E.R., Conway, J.H., and Guy, R.K. (1982) *Winning ways for your mathematical plays.* Academic Press, London.

Brown, J.S. (1969) *Laws of form.* Allen and Unwin, London.

Cannon, W.B. (1932) *The wisdom of the body.* London.

Codd, E.F. (1968) *Cellular automata.* Academic Press, London.

Colby, K.M. (1973) "Simulations of belief systems" in *Computer models of thought and language,* edited by R.C.Schank and K.M.Colby, pp 251-286. Freeman, San Francisco.

Davis, R. and Lenat, D.B. (1982) *Knowledge-based systems in artificial intelligence.* McGraw Hill, New York.

Dewdney, A.K. (1984) "Computer recreations" *Scientific American, 250* (5) 14-22, May.

Dyer, M.G. (1983) *In-depth understanding.* MIT Press, Cambridge, Massachusetts.

Fogel, L.J., Owens, A.J., and Walsh, M.J. (1966) *Artificial intelligence through simulated evolution.* Wiley.

Gaines, B.R. (1981) "Autopoiesis: some questions" in *Autopoiesis: a theory of living organization,* edited by M.Zeleny, pp 145-154. North Holland, New York.

Haugeland, J. (1979) "Understanding natural language" *Journal of Philosophy, LXXVI* (11) 619-632, November.

Hofstadter, D.R. (1979) *Godel, Escher, Bach: an eternal golden braid.* Basic Books, New York.

Jantsch, E. (1981) "Autopoiesis: a central aspect of dissipative self-organization" in *Autopoiesis: a theory of living organization*, edited by M.Zeleny, pp 65-88. North Holland, New York.

Lenat, D.B., Sutherland, W.R., and Gibbons, J. (1982) "Heuristic search for new microcircuit structures: an application of artificial intelligence" *AI Magazine*, 17-33, Summer.

Lenat, D.B. and Brown, J.S. (1984) "Why AM and EURISKO appear to work" *Artificial Intelligence, 23*, 269-294.

Maslow, A.H. (1954) *Motivation and personality.* Harper and Row, New York.

Maturana, H.R. (1975) "The organization of the living: a theory of the living organization" *Int J Man-Machine Studies, 7*, 313-332.

Maturana, H.R. and Varela, F.J. (1980) *Autopoiesis and cognition.* D. Reidel, Dordrecht, Holland.

McCulloch, W.S. (1954) "Through the den of the metaphysician" *British J of the Philosophy of Science, 5*, 18-31.

Meehan, J.R. (1977) "TALESPIN, an interactive program that writes stories" *Proc Fifth International Joint Conference of Artificial Intelligence*, 91-98.

Minsky, M.L. (1961) "Steps toward artificial intelligence" *Proc IRE, 49* (1) 8-30, January.

von Neumann, J. (1966) *Theory of self-reproducing automata.* University of Illinois Press, Urbana, Illinois.

Schank, R.C. and Abelson, R. (1977) *Scripts, plans, goals and understanding.* Lawrence Erlbaum Associates.

Searle, J.R. (1980) "Minds, brains, and programs" *Behavioral and Brain Sciences, 3.*

Shoch, J.F. and Hupp, J.A. (1982) "The worm programs -- early experience with a distributed computation" *Communications of the Association for Computing Machinery, 25* (3) 172-180, March.

Sloman, A. and Croucher, M. (1981) "Why robots will have emotions" *Proc 7th International Joint Conf on Artificial Intelligence, 1*, 197-202, Vancouver, BC.

Thompson, K. (1984) "Reflections on trusting trust" *Communications of the Association for Computing Machinery, 27* (8) 761-763, August.

Varela, F.J., Maturana, H.R., and Uribe, R.B. (1974) "Autopoiesis: the organization of living systems, its characterization and a model" *Biosystems, 5*, 187-196.

Varela, F.J. (1975) "A calculus for self-reference" *Int J General Systems, 2* (1) 5-24.

Varela, F.J. (1979) *Principles of biological autonomy.* North Holland, New York.

Varela, F.J. (1981) "Describing the logic of the living" in *Autopoiesis: a theory of living organization*, edited by M.Zeleny, pp 36-48. North Holland, New York.

Walter, W.Grey (1953) *The living brain.* Duckworth, Republished by Penguin Books, Middlesex, England, 1961.

Wilensky, R. (1983) *Planning and understanding: a computational approach to human reasoning.* Addison-Wesley.

Young, J.Z. (1978) *Programs of the brain.* Oxford University Press, Oxford, England.

Zeleny, M. (1977) "Self-organization of living systems: a formal model of autopoiesis" *International J General Systems, 4* (1) 13-28.

Zeleny, M. (1978) "Apl-autopoiesis: experiments in self-organization of complexity" in *Progress in cybernetics and systems research III,* edited by R.Trappl, G.J.Klir and L.Ricciardi, pp 65-84. Hemisphere, Washington DC.

Zeleny, M.(Editor) (1981) *Autopoiesis: a theory of living organization.* North Holland, New York.

Zeleny, M. (1981a) "What is autopoiesis?" in *Autopoiesis: a theory of living organization,* edited by M.Zeleny, pp 4-17. North Holland, New York.