ARTHUR W. BURKS

TELEOLOGY AND LOGICAL MECHANISM*

1. INTRODUCTION

There are many processes whose constituents contribute to or are likely to contribute to, or are intended to contribute to the achievement of some goal or purpose. Such processes are usually called teleological, and the goal is sometimes called a final cause. Final causality is generally contrasted with efficient causality.

Teleological explanations are very different from explanations given in terms of efficient causes, though the logic of the difference has never been made clear. It is sometimes said that efficient causes explain the future in terms of the past and the present, while final causes explain the present in terms of the future. This characterization is overly simple. In the case of a system that is deterministic both backwards and forwards, an explanation in terms of efficient causes can go in either temporal direction. In an intentional, conscious goal-directed process, present plans and hopes play the guiding role, so facts about the present explain the probable future.

Some teleological theories are equivalent to theories expressed in terms of efficient causes. Consider the motion of light from a point A to a mirror, from which it is reflected to point B. The law of least time says that the light follows that path which minimizes the transit time. Now minimizing (or maximizing) relative to a context often characterizes goal-directed activity (compare the principle of maximizing utility). The law of least time can be viewed teleologically: the light travels *as if* it were trying to reach point B as soon as possible. Yet this law is equivalent to the conjunction of two laws of efficient causality: light travels through a uniform medium in a straight line, and light is reflected from a mirror at an angle which is equal to the angle of incidence.

It has been held that all teleological explanations can be reduced to efficient causes, and I will argue for this position in the present paper. This is an in-principle claim only, like most reduction claims. Teleological explanations are very useful, and from a practical point

Synthese **76** (1988) 333–370.

of view they are indispensable. Opposed to reductionism is the thesis that teleological processes are directed by final causes which are *not* reducible to mechanisms of any kind.

I will begin with some ancient teleological doctrines: Plato's view that the Idea of the Good has a moral force, Aristotle's theory of final causes, and the Neo-Platonic Great Chain of Being. These are contrasted with the mechanisms of Pythagorean harmony and Greek atomism.

I next introduce my own form of mechanism, what I call "logical mechanism". This makes heavy use of ideas from computer science and genetics. It is important that the basic causality of computers and genetics is efficient causality.

Section 4 describes a two-part computer model of intentional goal-seeking. The static part consists of a goal representation and a plan for achieving it. The dynamic part is a repeated feedback cycle of sensed input, internal information processing, and action output. This model shows that intentionality is reducible to logical mechanisms.

In order to give a better formulation of the two basic theories of teleology, a teleological continuum of goal-seeking systems is defined in terms of robot models. This continuum runs from simple direct-response systems to intentional systems. There are two traditional explanations of this teleological continuum: the final cause theory of Plato, Aristotle, and Plotinus, and the reductionist theory of the Greek atomists and modern evolutionary biologists.

I think Charles Peirce has made the best case for the final cause theory, and his views are expounded in Section 7. In the following section I argue for the opposing theory. The reductionist theory of teleology explains low-level teleology in terms of a large, but finite, set of direct response rules which often lead to survival and genetic reproduction. Moreover, the evolution of intentional systems from direct response goal-directed systems can be explained by means of efficient causes.

Consciousness has often been held to involve final causes in an essential way, and so we need to consider whether it is reducible to logical mechanisms. Two aspects of consciousness are distinguished: immediate experience (e.g., a sensation of pain) and functional consciousness (e.g., the experienced functional connection from pain as stimulus to the immediate experience of repair action as response). This paper deals only with the nature of functional consciousness.

A survey of several types of conscious functioning shows that all are ways in which the organism controls itself and its actions. Functional pain and color experiences involve short-term control, while intentionality is a computational procedure for long-term control. But a mere collection of specific conscious functions does not have the unity of consciousness.

Consciousness is a single unified subsystem of the human person, carrying out many specific control functions intentionally and capable of being turned on (awake) and off (asleep). It is suggested that the ability to sleep evolved as an efficient general control method for protecting an animal from predators during the diurnal period (day or not) when it was not efficient for the animal to be active. This theory of functional consciousness is automaton based, involving comparisons between the human person, having mind and body, and the organization of a computer having a central control. Viewed from the perspective of computer architecture, human consciousness is a particular kind of computer control system, a relatively simple real-time control that, when the system is awake, directs short-term activities and plans longer-term activities.

To conclude, I compare ancient and modern theories of teleology. The final cause theory of teleology uses conscious intentional goal-directedness as a model of final causality. But, being limited to this model, it cannot explain how final causes operate on the unconscious and non-representational levels. The philosophy of logical mechanism shows how to reduce goal-directedness, unconscious as well as conscious, to efficient causality.

2. ANCIENT TELEOLOGY AND MECHANISM

Elsewhere I have presented a robot model of intentionality and an architectural theory of functional consciousness (Burks 1984, 1986a). In the present paper I will use these and other results to argue that teleological processes are reducible to logical-mechanical or computational processes (see also Burks 1988).

Perhaps a good way to start my presentation is autobiographical. As a graduate student and young instructor in philosophy I taught and studied several ancient doctrines which I did not understand, and which puzzled me off and on for a long time. Now I would like to deal with, and resolve, three of these puzzlements.

The first was Plato's doctrine that the Form of *The Good* directly influences and *in-forms* things and events, and that this influence is so strong that one who *knows* the good *does* the good, so that all evil is due to ignorance. The premiere example of this doctrine occurs near the end of the *Phaedo*, where Socrates says that the "true cause" of his staying in prison rather than escaping does not have to do with his bones and sinews. Rather, Socrates says, the true cause of his staying is the *reason* he does it, namely, that staying in prison to die is the right and honorable thing to do under the circumstances.

My second puzzle was over a related theory, Aristotle's doctrine that *final causes* are needed to explain purposive growth, for example, how an acorn is able to grow into an oak under favorable circumstances. I understood why Aristotle believed that efficient causality could not explain the acorn's growth – at least no one seemed to have a *good* explanation of this, as opposed to a description of the stages of growth. The problem was that I couldn't understand how Aristotle's final cause explained growth either.

Third, I puzzled long and hard over the Neo-Platonic Great Chain of Being, as used, for example, by Descartes in his *Meditations*. This doctrine makes two claims: first, that all beings are arranged in a hierarchical system, a kind of cosmic caste system; and second, that various levels of the hierarchy have different degrees of being or existence or reality. God is at the top of this hierarchy, man is in the middle, animals are below, and matter is on the ground floor. Now it wasn't the concept of hierarchy per se that was bothersome, for this concept is in Plato's *Republic*. Nor was it strange to find higher degrees of value assigned to higher levels of the hierarchy – Plato had done that in both the *Republic* and *Timaeus*. But I did find it puzzling when Descartes talked about *degrees of reality* and assumed that *degrees of existence* are logically correlated with *degrees of value* and also with *degrees of knowledge*.

There is a correlated Neo-Platonic principle of causality, that the cause must have at least as much reality as its effect. More technically, every entity must have a cause, and that cause must possess at least as much reality as the entity possesses. Opposed to this is the principle of efficient causality, which places cause and effect on the same ontological level and does not involve value considerations.

The Neo-Platonic causal principle extends derivatively to representations, with the consequence that, since *finite* man has an idea of an

infinitely perfect God, that God must exist. Indeed, most standard arguments for God's existence flow from the Neo-Platonic hierarchy and causal principle, as follows:

> The cosmological or first-cause argument: everything that exists must have a cause, and ultimately a self-caused cause, which is the one infinite God;
> The ontological argument: since perfection is correlated with existence, if God is defined to be infinitely perfect, God must exist;
> The argument from design: every entity was designed by a designer which has at least as much design-power as the entity designed, hence nature must have a designer, namely God.

Even Berkeley's argument for God depends on this causal principle: since we cannot put sense impressions in each other's minds there must be a being that does, namely, God.

What, you may be asking, does all this ancient and medieval philosophy have to do with contemporary issues of teleology, the topic of this paper? Well, the doctrines which puzzled me are all teleological in the most general sense, having to do with the role of values, goals, and purposes in creation, growth, development, causation, and action.

There was, of course, an opposite approach to teleology in ancient philosophy. In the same dialogue (the *Phaedo*) where Socrates gives a teleological explanation of his behavior, Simmias gives a non-substantive theory of mind. This account is based on Pythagoras' discovery that the harmony of the lyre depends on simple integer ratios 2 : 1 (octave), 3 : 2 (fifth), 4 : 3 (fourth). (For the Greeks these were ratios of the length of strings, while for us these are the fundamental frequencies with which the strings vibrate.) Using this knowledge, Simmias draws the analogy: the soul is to the body as the harmony of a lyre is to the lyre.

The problem with this theory of mind is that harmony characterizes some minds or persons and not others, so how can the concept of harmony be used to explain mind? As Plato said in *The Republic*, a harmonious person is one in which mind (or soul), will, and appetites all cooperate, with mind playing the controlling role. As you will see later, my architectural theory of consciousness stresses the control function of mind (Section 9).

The Greek atomists also gave a materialistic and reductionist theory of the mind, contradicting the final cause theory. Consider Lucretius' formulation in *De Rerum Natura*. Lucretius says very clearly that the mind is a special compound substance, composed of small, round, smooth atoms that dart rapidly around the body, carrying information from the senses into the body and carrying instructions for action to the limbs.

Thus Lucretius had a very different theory of mind from Plato. For Plato the mind was an indestructible mental atom influenced by the Realm of Forms above, while for Lucretius mind was a compound of material atoms influenced only by other material atoms. But it is very noteworthy that both philosophers held that the mind should control the rest of the body and that the good life results from such control.

Lucretius' view of mind seemed plausible to me. Mental activity is the functioning of a material substance which is a different kind of material substance from the body, and hence functions differently. Nerves are to bones and ordinary flesh as Lucretius' smooth atoms are to his hooked and rough atoms. But Lucretius had only analogical arguments for his view, and while these had matured into strong inductive arguments by the time I studied philosophy, no one had a very good account of how the mind works, that is, how the central nervous system thinks, reasons, solves problems, controls actions, directs the body in the pursuit of goals, and performs other control functions.

3. THE PHILOSOPHY OF LOGICAL MECHANISM

Thus in ancient philosophy there are two theories of teleology, which are forms of what are appropriately called "the final-cause theory" and the "reductionist theory". In the end I will come out on the side of the reductionist theory, but, I hope, in a way that explains teleology better than traditional reductionism, and in a way that adds to our understanding of why the ancients believed in the final-cause theory.

The argument will rely heavily on two broad areas of knowledge which have developed in the last forty years. The first area is that of computers and robots, how to program them, how to design them, and also the theory of automata, including self-reproducing automata. This subject gives a much more effective idea of the logic of mental processes, learning, and growth than existed before. Two concepts

from this subject are especially relevant, the *complexity of a computing system* and *hierarchy*. The complexity of a computer is measured by the length of its shortest description in some standard design language, or by the length of an equivalent program, assuming some standard program language. The concept of hierarchy is an important concept in understanding computer architectures. A computer organized into several levels with control flowing ever downward from the top is a little analogous to the Neoplatonic view of reality.

The second area of knowledge consists of the developments in modern genetics in the last forty years, both at the level of molecular theory and at the level of genetic theory, including the application of game theory, and the conception of the genome as a genetic program that constructs an organism which is tested in an environment. These achievements have produced a much better understanding of the logic of evolution than existed forty years ago.

It is important in a discussion of the foundations of teleology to keep in mind the bearing of these two developments on the issue of final versus efficient causality. The basic causality of an operating computer is efficient causality, and the causality of genetics is efficient causality.

Using computer ideas, one can convert Lucretius' idea of smooth atoms rushing around the body into the computer model of sensory stimuli becoming digital messages, these messages being transformed by a computer program, and the output messages of that program causing action responses. With these modern developments in the biological and computer sciences Lucretius' mechanistic account of mind is now firmly grounded.

I have expressed this basic theory about mind in a thesis which I call the man-machine thesis, or the man-robot thesis, or the man-automaton thesis: *A finite automaton can perform all natural human functions*. To compare a computer with a human, one needs to make its input-output equipment more like that of a human. Imagine television cameras for eyes, microphones for ears, and sensory devices for odors, tastes, temperatures, etc. As motor outputs, the machine has mechanical arms whose hands and fingers can manipulate objects, and it has wheels and motors for locomotion. My thesis can also be stated as: for each person, there is (in principle, at least) a machine, robot, or automaton which will perform the same natural functions.

The man-automaton thesis expresses functionalism at the behavioral

level. It says nothing about how the computer works inside. We will get to that later, in two stages: first, an account of intentionality or goal-directedness, and second, a theory of consciousness.

I have developed my argument for the man-robot thesis elsewhere (Burks 1972; 1977, Section 10.4.1; 1979). It has three parts: (1) progress in biology, (2) progress in the design of computers and computer languages, and (3) an argument based on the psychological notions of threshold of sensation and minimal accuracy of action.

The man-robot thesis can be extended to cover computer simulations and models of biological evolution, natural processes generally, intentionality, and consciousness. These views are part of a general metaphysics and epistemology which I call "the philosophy of logical mechanism" (see Salmon 1989). This philosophy is a generalization of traditional materialistic and mechanistic doctrines which relies heavily on stored-program computers and robotic extensions of them as models. In my sense of "logical mechanism", automatic devices, computers, robots, natural organisms, genetic programs, learning mechanisms, and evolutionary processes are all logical mechanisms.

4. A COMPUTER MODEL OF INTENTIONALITY

One of man's highest activities is that of formulating alternative goals, choosing one goal from among them, and then pursuing it systematically and intelligently. This activity is appropriately called *intentional goal-seeking*. I will outline here how to construct a robot with this capacity (Burks 1984, Section 2). Intentional goal-seeking employs several underlying computer capacities: sensing, reasoning (inductive as well as deductive), use of a knowledge base, and action.

The basic structure of intentional goal-seeking has (1) a relatively *static part*, consisting of a goal representation and a plan for attaining the goal, and (2) a *dynamic part*, a repeated feedback cycle of sensed input, internal information processing, and action output.

(1) A goal is some possible future state of the environment, the goal-seeking system, or a relation between the two. Often a future goal-state is represented in relation to the present state of the system and its environment, perhaps as a sequence of intermediate steps or means to the end sought. This representation merges with the sequential plan or strategy for achieving the desired end. There are alternative routes for reaching a goal, each with sub-goals. Which alter-

native is best depends on the circumstances at each step, circumstances which in turn may depend on the actions taken at earlier steps. The plan may include a procedure for modifying the goal or terminating the intention under certain conditions. The cost of the effort to attain a goal can be compared with the probable reward, and the goal modified or replaced if the price of continued efforts becomes excessive.

(2) The dynamic part of intentional goal-seeking is an iterated cycle of data collection, a process of calculation and decision, and action. The system receives information about its environment, and possibly about itself. It updates its representation of itself in relation to its goal, and evaluates that relation, makes predictions, consults the strategy (and perhaps modifies it), decides what to do, and does it. This cycle repeats until the goal is reached, modified, replaced, or withdrawn.

The preceding formulation of intentional goal-directedness is in terms of a single, fixed goal. This fits most of the automatic systems man has constructed so far, such as guided missiles. Operating systems and security systems for computers are designed to reconcile the goal of many users and hence are perhaps exceptions. But in any case natural systems stand out as typically having a complex of goals, some conflicting with each other.

A natural intentional system, such as a human, has a dynamic hierarchy of goals. Basic inborn drives occur at the lowest level. Acquired habits dominate intermediate levels. Explicit goals, and possibly a life-plan, occupy the highest levels. Moreover, the goals of this structure are only partly unified, being partly conflictive, and change over time. As John Dewey emphasized, people change not only their means but their ends as they learn from experience what they want and how to get it. This is especially the case with creative work. Also, there may be a higher-level goal of modifying and harmonizing the goals on the lower levels of the system.

The psychologist Franz Brentano held that the mental is not reducible to the physical, and thought he had found an ability unique to the mental. He said that the essence of mental activity is to be directed toward objects that may not exist, and that a material system cannot have this property. But it is easy to imagine a robot that thinks of a house that doesn't exist, and pursues the goal of making that house come into existence. A robot is a complex physical system, so a physical system can be intentional. It should be emphasized, however,

that the modern stored-program computer is a very special physical system, which had not been conceived in Brentano's time, so that his claim was plausible when he made it.[1]

This ends my analysis of intentional goal-seeking, i.e., how it operates in humans and how it can be built or programmed into robots. The analysis is only partial, and somewhat paradigmatic, with actual cases typically making incomplete use of it for various practical reasons.

The next section will contrast intentional goal-directedness with a much more elementary kind of goal-seeking mechanism, direct-response. But before this, I will make some relevant points about intentionality.

First, it is clear from our computer analysis of intentionality that, in this case at least, ordinary or efficient causes explain goal-directedness. The representation of the goal (a desired future state) plays a clear causal role in the teleological process.

Second, consciousness was not involved in our robot description of intentionality. This shows that intentionality is a high-level control procedure that can operate without being conscious. An adequate theory of consciousness must be consistent with this fact.

Third, consider the three examples of teleology I learned as a young philosopher and which puzzled me to no end: the transcendent forces of Plato's Idea of the Good, the teleological action of Aristotle's final causes, and Descartes' principle that the cause must have as much reality and excellence as its effect. All three of these ideas seem to me to be derived from an intentional model of goal-directedness. I will elaborate on this point with respect to Aristotle's final causes in the last section of this paper.

Fourth, many tasks can be carried out by either intentional systems or by non-intentional systems. Men and beavers topple trees, and so do rivers and storms. As Darwin was the first to fully appreciate, the artificial selection practiced by plant and animal breeders was paralleled by natural selection.

Fifth, humans have the best developed and most creative intentional abilities of any systems in their part of the universe, although it is likely that there are more powerful intentional systems elsewhere. Humans have evolved over millions of years from physical and chemical materials that are non-intentional. Let us now look at this process.

5. THE TELEOLOGICAL CONTINUUM

A system pursuing a goal responds differently to different circumstances, and insofar as the system is operating successfully the responses chosen tend to contribute to the achievement of the goal. An intentional system can do this by explicit representation and calculation in the elaborate manner described in the last section. But goals can also be pursued in a much simpler manner. A furnace thermostat is an example. It turns the furnace on (or off) if the temperature is below (or above) a certain level. Thus it operates by means of a fairly direct connection between stimulus and response. I call a system that functions in this manner a *direct-response goal-seeking system*.

Simple organisms are direct-response goal-seeking systems. The course of biological evolution from cells to homo sapiens has been a gradual development of intentional systems from direct-response systems. This process yields a natural dimension for classifying goal-directed systems: one can ask of each such system where it fits on this continuum. I call this *the teleological continuum* (see Burks 1984, Section 3).

"Teleology" is used here in a broad sense, to refer to the goal-seeking nature of the systems involved. It is neutral between the two opposing views: (1) teleology is reducible to efficient, material or mechanical causes, and (2) teleology is not so reducible. These views will be discussed in the next three sections of this paper.

In the term "teleological continuum" the word "continuous" is intended in a loose sense, allowing that evolution may advance by catastrophes, large or small. Furthermore, it allows that evolution uses quasi-stable or slowly changing natural building blocks, such as the four-chambered heart and the gene. Thus "continuous" means that evolution proceeds in relatively small steps, that is, that there are not large gaps in the evolutionary chain.

Evolution and growth proceed continuously, beginning with simple, isolated elements and proceeding to complex, highly integrated systems operating over a hierarchy of levels. This is an inclusive process: the mechanisms of earlier stages are preserved in later stages. For example, the human knee-jerk response is direct, and the human employs various mechanisms developed along the teleological con-

tinuum. Thus direct-response goal-seeking is not replaced by intentionality, but is incorporated in it. With this inclusion in mind, let us compare the way a direct-response system processes information and controls itself with the way an intentional system performs these functions.

There is nothing in a direct-response system comparable to the static part of intentionality, a symbolic representation of the goal and a more or less explicit plan or strategy for achieving it. The function of the dynamic portion of intentionality (the iterated cycle of input stimulus, internal calculation, and output action) is performed in the direct-response system by its set of rules about how to respond to specific stimuli. Thus the mode of computation in a direct-response system is essentially table lookup, the table entries being the simple direct-response hypotheticals: If the input stimulus is I and the internal state of the system is S, then act in manner A. As a result of executing such a rule the system moves to a new state S', and so a full statement of a direct-response rule should be: If input I and internal state S, then action A and next internal state S'.

In contrast, an intentional system employs the more complicated (but more effective in suitably complicated contexts) computational procedure described in the last section. The system contains a model of its present status in relation to its goal and regularly updates that model on the basis of the information it receives. Moreover, the data structure used by the intentional system to store its possible action responses to various environmental situations is better organized and richer than that of the direct-response system. The former uses a strategy, that is, a well-organized structure (perhaps a tree) with weights representing values and probabilities attached to its options.

These computational differences between direct-response and intentional goal-directed systems imply a significant difference in their relative abilities to adapt or learn. Each dynamic cycle of an intentional system incorporates additional information into the system, and in the case of a succcessful system that information contributes to achieving the goal. As described, a direct-response system has no learning ability. A natural extension of it is obtained by replacing each individual rule with a set of alternate rules and by providing the system with a method for evaluating the relative success of the different rules of each set.[2] Such an extended direct-response system can learn to adjust to its environment, but not as rapidly as an

intentional goal-directed system, since the process of changing rules is slower than the computational process of intentionality.

Let us return now to the concept of a teleological continuum of goal-directed systems, a continuum running from simple direct-response systems to sophisticated intentional systems. I introduced the concept by reference to biological evolution, but I want to generalize the concept of a teleological continuum and make it more abstract and logical. Then it can be connected to automata theory, on the one side, and to my three puzzling teleological examples, on the other.

This generalization is made in two steps. First, remove time and the idea of change, and think of the teleological continuum as a linear ordering of systems, arranged on a dimension of varying mixtures of direct-response operation and intentional operation.

The second step is rooted in the generalization of my man-machine or man-robot or man-automaton thesis to cover all organisms: that is, for any organism there is a finite automaton equivalent to it. This is a generalization that moves from the materials of a system to its logical structure – look at a system in terms of its logical switches and memory cells (or their analog equivalents) and how these are organized at all levels. Thus in the second step one looks at a direct-response system, or an intentional system, or any system in between, as an automaton, whether it is made of hardware, soft flesh, or even of software running in some minimal universal computer.

6. TWO EXPLANATORY THEORIES OF THE TELEOLOGICAL CONTINUUM

Next I apply the concept of a teleological continuum to the three examples of teleology that puzzled me so long: Plato's view of how the Idea of the Good produces good conduct, Aristotle's doctrine of final causes, and the Great Chain of Being. These three examples are closely related. Aristotle's final causes are Plato's transcendent Forms or Ideas made immanent. Neo-Platonism is a consolidation and blending of Platonism and Christianity.

To obtain a basis of comparison, remove the transcendent part of the Great Chain of Being – God, the angels, etc., and consider only Nature, the natural part. This part is clearly a teleological continuum, running downward from men through animals to acorns and the like. Man is at the highest level (in Nature, that is), and in rational man

teleology or final causation operates consciously. At a low level, as in the acorn, teleology or final causation operates unconsciously and non-intensionally. In comparison to the teleological continuum described by modern evolutionary biology, this ancient concept is, of course, very fragmentary. Nevertheless, it is an antecedent of the modern conception, both structurally and historically. Moreover, all three philosophies had theories to explain the teleological continuum, and in a broad sense these theories all employed the notion of a final cause.

In contrast, when Lucretius explained evolutionary facts and goal-directedness he always did so in terms of efficient causes, not final causes. Moreover, when he explained the evolutionary origin of things (compounds of atoms) he didn't appeal to final causes, but to the *absence* of efficient causes (his theory of the initial indeterministic swerve).

Thus ancient philosophy had two theories of teleology, one saying that final causes are needed to explain goal-directed action, the other saying that efficient causes are sufficient. For the most part these theories only attempted to explain the goal-directedness of individual systems in the teleological continuum. But the Great Chain of Being is a theory of the logical succession of the links of the Chain, and Lucretius did attempt to explain a few evolutionary events.

In the spirit of these two traditions I will formulate two explanatory theories of the teleological continuum, as it was defined in the preceding section.

(I) *The final-cause theory of the teleological continuum*: goal-directed action is always the result of final causes. At the human level final causes operate representationally and consciously. Final causes also explain the goal-directed actions of plants and animals at lower levels, and on those levels they do not operate consciously or representationally. Moreover, final causes explain the successive steps of biological evolution.

Teleological processes cannot be fully explained in terms of efficient causality, either deterministic or probabilistic. Final causes are not reducible to mechanisms of any kind.

(II) *The reductionist theory of the teleological continuum*: all goal-directed action is explainable by means of laws governing matter. This explanation involves two factors or types of laws: (A) efficient causality, and (B) forces or mechanisms that produce compounds that are

stable, at least statistically, so that they can evolve. The teleological continuum has been produced in nature over millions of years by laws and mechanisms of evolution which are, in principle, reducible to the laws governing inanimate matter.

All of this is familiar naturalistic doctrine, except for the requirement that the atoms be such that stable compounds of them can evolve, including logical switches, memories, and complex organisms.

Note that Lucretius postulated an atomic mechanism to explain stability: some atoms had hooks and other atoms had holes. Mental atoms, he said, were smooth and round. Smooth and round atoms cannot form stable structures, so in Lucretius' philosophy the stability of mental functioning must be explained by the bodily framework of mind. The body channels the flow of mental atoms. (Compare a fluid computer in which water moves through pipes and switching takes place at valves.)

It should be emphasized that the only important difference between these two theories concerns the foundation of causality. The reductionist agrees that in practice it is necessary to talk in terms of teleology and goal-directedness, even for lower biological forms. For most teleological phenomena, the underlying explanation in terms of efficient causes is much too complex to serve as a substitute for the teleological description. This is so even though reductions by means of theoretical analyses and computer simulations may greatly increase the scientist's understanding of the phenomena and improve his or her way of interpreting them. (We will return to this issue in Section 8.)

7. PEIRCE'S ARGUMENT FOR THE FINAL CAUSE THEORY

Charles Peirce had a good understanding of the teleological continuum established by the work of Charles Darwin and others on biological evolution. Both Peirce's pragmatism and his inductive logic were strongly influenced by the theory of evolution. Indeed, he suggested that the operation of evolution is governed by a complicated theorem of statistics and inductive logic, a generalization of the "law of gamblers' ruin". For the time, Peirce had a deep understanding of how evolution worked, and he understood the relation of Darwin's theory to geology (Lyell), economics (Smith and Ricardo), and ecology (Malthus).

In his later period (1890s on) Peirce attempted a grand metaphysical and epistemological synthesis. This was to incorporate and generalize his earlier epistemology, pragmatism, semiotic, logic (abduction, induction, deduction), and also ideas from Aristotle and the medieval realists. As part of this synthesis Peirce developed a theory of cosmic, biological, and intellectual evolution. This was his triple doctrine of tychism (chance), synechism (continuity), and agapism (evolutionary love). He depicted an evolutionary process infinite in both directions, running from an iniitial chaos through a gradual evolutionary process toward an ultimate limit of "concrete reasonableness".

Peirce's proposed synthesis is a final cause theory of the teleological continuum, as the quotations which follow will show. His *Thirds* are the final causes directing the evolutionary advance towards limits. These Thirds are akin to Aristotle's final causes and to an immanent form of Plato's Idea of the Good. For a long time I had difficulty understanding many enigmatic statements of Peirce's later period, and I think the following final cause interpretation makes sense of them.

Peirce was the first post-Darwinian thinker to give a final cause theory of the teleological continuum. Henri Bergson's theory of the *elan vital* is a competing theory, but Bergson's *Creative Evolution* was only published in 1907. (It was also very different, for the *elan vital* is an intuitive force, more like a Peircean First than a Peircean Third.) C. Lloyd Morgan held that new kinds of systems and principles *emerge* in biological evolution. He thought there was no reductive explanation of an emergent, but he didn't use final causes to explain them.

We present Peirce's explanation of how final causes operate in two parts. First, we list six points he makes about final causes, illustrating each with quotations. Then we show how his doctrines of tychism (there is objective chance), synechism (the evolutionary process is continuous), and agapism (the evolutionary process is guided by objective values) are related to his final cause theory of the teleological continuum.

First, final causes operate in concert with efficient causes, final causes providing the general goals or ends while efficient causes are the means of achieving these ends.

The mere carrying out of predetermined purposes is mechanical. (6.157)

Final causality cannot be imagined without efficient causality; but no less on that account are their modes of action polar contraries. (1.213)

There is efficient causation and there is final, or ideal, causation. If either of them is to be set down as a metaphor, it is rather the former. Pragmatism is the correct doctrine only in so far as it is recognized that material action is the mere husk of ideas. (8.272)

Second, final causes are general, leaving room for the employment of means that depend on the circumstances.

... we must understand by final causation that mode of bringing facts about according to which a general description of result is made to come about, quite irrespective of any compulsion for it to come about in this or that particular way; although the means may be adapted to the end. The general result may be brought about at one time in one way, and at another time in another way. Final causation does not determine in what particular way it is to be brought about, but only that the result should have a certain general character. (1.211)

The evolutionary process is, therefore, not a mere evolution of the *existing universe*, but rather a process by which the very Platonic forms themselves have become or are becoming developed. (6.194)

Thus a final cause only specifies and controls the general character of the object it produces, while efficient causality determines the details. The final cause in an acorn works towards the acorn becoming an oak, but the specific environment will determine how it becomes an oak and what kind of oak it becomes, if it becomes an oak.

Third, efficient and final causes operate in opposite temporal directions, efficient causes from past to present, final causes from future to present.

In the flow of time in the mind, the past appears to act directly upon the future, its effect being called memory, while the future only acts upon the past through the medium of Thirds. (1.325)

To say that the future does not influence the present is untenable doctrine. It is as much as to say that there are no final causes, or ends. The organic world is full of refutations of that position. Such action constitutes evolution .... (2.86)

Fourth, final causality is mental in nature. Human intentional goal-directedness is clearly mental. Final causes not operating through conscious beings are mental in a broader sense.

But the being governed by a purpose or other final cause is the very essence of the psychical phenomenon, in general. (1.269)

The mind works by final causation, and final causation is logical causation. (1.250)

The one intelligible theory of the universe is that of objective idealism, that matter is effete mind, inveterate habits becoming physical laws. (6.25)

... what we call matter is not completely dead, but is merely mind deadened by the development of habit. (6.158)

... *tychism* must give birth to an evolutionary cosmology, in which all the regularities of nature and of mind are regarded as products of growth, and to a Schelling-fashioned idealism which holds matter to be mere specialized and partially deadened mind. (6.102)

Fifth, final causes are needed to explain the operation of holistic, coherent systems. These are systems in which the parts are highly interdependent, each part depending on all the others.

Efficient causation is that kind of causation whereby the parts compose the whole; final causation is that kind of causation whereby the whole calls out its parts. Final causation without efficient causation is helpless; ... efficient causation without final causation, however, is worse than helpless, by far; it is mere chaos; and chaos is not even so much as chaos, without final causation; it is blank nothing. (1.220)

Sixth, final causes are objective values or ideals, guiding the evolutionary process so that nature becomes ever better and knowledge evolves toward perfection, the whole moving toward "concrete reasonableness". Peirce's expressions of this point are unashamedly romantic and sentimental.[3]

Truth, crushed to earth, shall rise again.... ideas are not all mere creations of this or that mind, but on the contrary have a power of finding or creating their vehicles, and having found them, of conferring upon them the ability to transform the face of the earth. (1.217)

No doubt Truth has to have defenders to uphold it. But truth creates its defenders and gives them strength. The mode in which the idea of truth influences the world is essentially the same as that in which my desire to have the fire poked causes me to get up and poke it. (8.272)

... as for the cosmos, only so far as it yet is mind, and so has life, is it capable of further evolution. Love, recognizing germs of loveliness in the hateful, gradually warms it to life, and makes it lovely. That is the sort of evolution which every careful student of my essay 'The Law of Mind' [6.102–163] must see that synechism calls for. (6.289)

... evolution[ary] ... development go[es] through certain phases, having its inevitable ebbs and flows, yet tending on the whole to a foreordained perfection. Bare existence by this its destiny betrays an intrinsic affinity for the good. (6.305)

Peirce's final cause theory of the teleological continuum is most systematically expounded in his 1891–93 series on metaphysics:[4] 'The Architecture of Theories' (6.7–34), 'The Doctrine of Necessity Examined' (6.35–65), 'The Law of Mind' (6.102–163), 'Man's Glassy Essence' (6.238–271), and 'Evolutionary Love' (6.287–317). There

are three aspects of this theory: tychism, synechism, and agapism.

Tychism is the doctrine that the basic laws of nature are probabilistic. In "the beginning" there were no connections at all, then some weak probabilistic connections occurred by chance, and over time these have strengthened. The properties connected by laws are themselves probabilistic groupings of simpler properties, and they evolve along with the laws. Laws that seem deterministic are actually limiting cases of probabilistic laws. Thus probabilistic laws (or "habits") are the entities which evolve.[5]

Synechism is the doctrine that this evolution is a continuous process. Peirce insisted that it was continuous in the strict mathematical sense.[6] This was so that the probabilistic connections in laws could evolve gradually. Peirce saw that to change from one deterministic law to another would involve a big jump, and evolution is more gradual than that.

Agapism is the doctrine that the evolutionary process is guided by final causes, or Thirds, and that these tend to make developments move in the direction of perfection. This is a generalization of Peirce's definition of truth as "the opinion which is fated to be ultimately agreed to by all who investigate" (5.407). Both Peirce's earlier epistemological optimism and his later cosmic optimism are forms of nineteenth century evolutionary optimism.

This completes our exposition of Peirce's final cause theory of the teleological continuum. We will conclude this section with some evaluative remarks about it.

Each of Peirce's six points has an analogue in our account of intentionality (Section 4). First, goals and means operate on the same level. Second, a goal is general, and different sequences of means can be employed to achieve it. Third, a goal is a desired future state, while a means is a present step toward it. Fourth, human goal-directedness is mental. Moreover, since it occupies a position in the teleological continuum, mentality has developed gradually over the course of evolution from lower forms of life, which developed gradually from non-living things. Fifth, a conscious intentional human being is a holistic-coherent system. Sixth, achievement of a goal is generally an achievement of some value.

Thus Peirce's six points about final causes constitute an insightful phenomenology of intentional goal-directedness. His final cause theory of the teleological continuum can be interpreted as an ap-

plication of this phenomenology to all of nature, by means of three metaphysical doctrines: the objective probabilism of tychism, the mathematical continuity of synechism, and the progress and limit claims of agapism. The result is a very interesting theory of the teleological continuum, and the first final cause theory to grapple with the problem of Darwinian evolution.

But we have argued in Section 4 that intentionality is reducible to logical mechanisms, and the reductionist theory holds that the teleological continuum is also reducible. Thus there is still the basic question: Is the final cause theory a better explanation of the teleological continuum than the reductionist theory? Do Peirce's six points and his doctrines of tychism, synechism, and agapism explain the pre-intentional forms of life and the historical development of the teleological continuum? Does Peirce show that the teleological continuum cannot be reduced? We will make some general comments before moving on to the reductionist theory of the teleological continuum.

Peirce's first, second, third, and sixth points and his agapism constitute a dualistic account of the driving force of goal-directedness. What happens is the result of two kinds of forces and controls: efficient causes, which operate forwards, and final causes, which operate backwards. Final causes are general and provide probabilistic tendencies, guiding efficient causes so they move events in the direction of an objectively good state of affairs. Contrast this dualistic analysis with that of our computer model of intentionality. In this model the concept of the goal is not per se a driving force. Rather, the motivation comes from the desire for the goal and the will to achieve it. Might not the driving forces for pre-intentional goal-directedness and the evolution of the teleological continuum also reside in efficient causes?

Consider next Peirce's fifth point, that final causes are needed to explain the emergence in evolution of holistic-coherent systems. Such systems are very complex and non-linear, with interacting dynamic feedback paths involving information flow as well as materials and energy flows. Organisms, ecologies, gestalt phenomena, coherent and well-organized thought systems, biological evolution, and cultural evolution are all holistic-coherent systems. The doctrines of final causes (Aristotle), entelechy (Hans Driesch), and elan vital (Bergson) were all created to explain such systems.

But many computer-based systems are holistic-coherent, and the

fact that human teams can design and operate them shows the limits of the argument that final causes are needed to explain the operation of holistic-coherent systems. These computer systems are very complex, involving millions of instructions and related hardware. The human design teams are necessarily large. Because of the complicated interactive character of these systems, the designers cannot proceed "in one fell swoop", either "bottom up" or "top down". Rather, they must use a cyclic feedback and successive approximation process, sometimes designing downward from a general plan of the whole, sometimes constructing from the parts up. Such a design process involves something like the mutual support relation described by the coherence theory of truth.[7]

Peirce is correct in treating evolution as a gradual process. Our concept of the teleological continuum (Section 5) was inspired by Peirce's synechism. He is right in holding that the existence of the teleological continuum refutes a dualism of mind and body. But that does not establish the reduction in either direction. If one believes with Peirce that matter is a special case of mind ("matter is mind hidebound with habit"), then it is reasonable to say that matter is mental in a broad, extended sense. On the other hand, if one believes that mind is reducible to matter, it is reasonable to say that mind is a form of matter.

Peirce recognized that ordinary mental categories are not suitable for dealing with the lower end of the teleological continuum and giving an account of the gradual evolution of conscious, intentional mental activity from lower forms of life. He developed his broad version of semiotics for this purpose. Thinking and reasoning take place in signs, and transform the information expressed by these signs. Peirce was among the first to recognize that many lower processes are semiotic: the transfer of pollen from a flower stamen to an ovule in a stigma, the neutral reaction of a frog's leg to an electrical stimulus, punched cards controlling the patterns woven by a Jacquard loom, the lower-level reasoning of a logic machine. He would have been pleased by Karl von Frisch's "language of the bees" and the use of languages in modern computers. He would have appreciated the fact that half the time span of biological evolution was needed to develop the living cell, the next quarter to devlop colonies of cells, and only the last quarter for the human mind.

I think goal-directedness, both conscious and unconscious, and the evolution of the teleological continuum, can be explained in terms of

efficient causality by using the results of two fields of study which have developed since Peirce's time. The first is modern evolutionary genetics, including the notion of a diploid genetic program directing the construction of an organism. Biological evolution is gradual (though of varying rates), but it is based on discrete genetic entities. In contrast, Peirce's strict continuity is based on the blending theory of inheritance, which R. A. Fisher proved to be inadequate (Burks 1984, p. 42). Chance plays an essential role in genetic evolution, but there are rigid connections as well.

The second subject is the interdisciplinary theory of computers, automata, robots, intelligent systems (natural as well as artificial), control systems, and complex non-linear systems. This includes information and communication theory, switching theory, the theory of finite automata, and the study of computer architectures. It also covers the theory of self-reproducing and self-repairing automata, as well as computing systems that reason, learn, and discover.

This approach to teleology is in the spirit of the philosophy of logical mechanism, and we will pursue it in the next two sections. But first we want to emphasize that at the time he gave it, Peirce's argument for final causes was plausible. He believed that evolution was progressive, and wanted an explanation of this feature of it. Mendelian genetics was not known when Peirce developed his final cause theory, and the blending theory was a natural hypothesis about the nature of inheritance. Assuming this theory to be true, Peirce saw correctly that probability (tychism) and continuity (synechism) are not enough to account for evolutionary progress, and he introduced agapism (final causality) to account for it.

Nevertheless, Peirce's doctrine of agapism does not really solve the problem, for this doctrine gives no explanation of how final causality works. "Agapism" is only Peirce's name for final causality in the context of tychism and synechism. The intentional case of final causality can be explained in terms of logical mechanisms (Section 4). But this explanation does not cover the non-intentional case of final causality.

8. ARGUMENT FOR THE REDUCTIONIST THEORY

My main argument for the reductionist theory of the teleological continuum is really an extension and adaptation of an earlier argument

for the man-automaton thesis (Burks 1972, 1986b). This argument had to do with the finiteness of the input stimuli and output responses of man, and the finiteness of the computing mechanisms inside. By the principle of the threshold of sensation I argued that a human is capable of responding only to a finite number of stimuli in any finite time span. A corresponding argument applies to a person's output actions. Finally, there is a minimum to the size and response time of human computing elements. Hence each person is, in a suitable sense, equivalent to a finite automaton.

Since the teleological continuum runs from direct-response systems at one end to intentional systems at the other end, and we have claimed that each system of the continuum is equivalent to a finite automaton – we pause to show the relation of the concept of a finite automaton to the concept of a direct-response goal-seeking system. Recall that a direct-response rule connects a given input stimulus I and a given internal state S to an action response A and a next internal state S'. Thus it is of the form

If I and S then A and S'.

But a finite automaton is equivalent to a finite set of such rules. Hence to say that an acorn is equivalent to a finite automaton is to say that it is equivalent to a finite set of direct-response rules.

A direct-response goal-seeking system, such as an acorn, is a system executing direct-response rules. It is important that each such system in nature, including an acorn, is capable of executing only a *finite number* of direct-response rules. In other words, the teleology of an acorn consists of this: an acorn is capable of a wide, but finite, variety of direct responses. In a suitable environment it will make a temporal sequence of responses that will cause it, by efficient causality, to become an oak.

In Aristotle's time there was no explanation of the goal-directedness of lower organisms in terms of efficient causality. He postulated that this goal-directedness was due to a special kind of causality, final causality, irreducible to efficient causality. Although this was a reasonable position at the time, we can now see that an organism with an appropriate set of fairly simple rules can, in a sufficient variety of circumstances, make responses which tend to contribute to short-term goals such as food, space, mating, and reproduction, and thereby perform so as to work towards long-term ends such as survival and

genetic fitness. Hence the goal-directedness of these systems – and a fortiori, of all the systems of the teleological continuum – can be accounted for in terms of efficient causality.

There were further biological phenomena that Aristotle and others after him thought required teleological causality: reproduction of organisms, and self-repair. But there are no *logical* difficulties in robots' accomplishing these functions by means of efficient causality, as von Neumann's theory of automata shows (von Neumann 1966; Burks 1970; von Neumann 1986), and biologists are well-advanced in giving detailed explanations of how organisms perform them.

Well, you may say, I now understand how a sufficiently complex direct-response system can maintain itself in an environment, grow, and reproduce. But I don't yet see how such direct-response systems could come into existence and evolve up to man as the result of efficient causality.

This is the higher-order, evolutionary question: Is efficient causality sufficient to explain how direct-response organic systems can arise out of a physical-chemical matrix and evolve into intentional organisms? Here again, I think the answer is "yes". For each organism of the teleological continuum, from micro-organisms to man, there is a finite automaton which can perform all the natural functions of that organism. Moreover, there is an automaton-like account of both pre-biological evolution and biological evolution: the evolution of purely physical entities into direct-response goal-directed organisms, and the evolution of the latter through the teleological continuum to intentional goal-seeking organisms (Burks 1984, 1986b).

It follows that, in principle, scientists should be able to simulate all these processes on computers. But these processes are very complex. The complete functioning of a single organism is computationally complex, the functioning of a group of organisms is more complex, that of competing groups even more complex, and that of evolution tremendously more complex. Hence it is not within our actual powers to simulate evolution in complete detail. My claim for the reduction of evolution to computer processes, and related claims such as the man-machine thesis, are theoretical rather than practical claims. At the practical level one must work with the unreduced theory and rough approximations to the underlying theory.

The idealized gas law illustrates this point. The law is: Pressure $\times$ volume = constant $\times$ temperature. This is a simple and useful law. The

underlying detailed theory combines the detailed state of the gas, particle by particle, with the laws of mechanics. By a statistical argument one can show that the gas law holds for any idealized gas, so that in this case the reduction is provable. But this makes no practical difference, because the underlying detailed theory is too complex for any human to grasp or use.[8]

Similarly, a teleological statement is simple enough to understand and evaluate, while the underlying detailed theory is too complicated to formulate. Thus a teleological description may be used to explain why a particular characteristic appeared in a species, and may provide a factor for evaluating the survival possibilities of that species, whereas a rigid proof of these matters is typically beyond the simulation powers of existing computers.

These points may be illustrated with our computer model of intentionality (Section 4). Suppose one develops an algorithm that will make a robot function intentionally. This intentionality algorithm may be expressed as a program and put in a general-purpose robot, or one can design a special-purpose robot to execute the algorithm. One can place the robot in an environment, study how well it works, and modify it in various ways to improve it. Each robot is a machine operating (by efficient causality) as a system of components (logical switches, memory elements, communication wires, etc.), but it would not be useful in this context to view it as a detailed system of components. The programmer works at some software level, usually with a well-developed system which facilitates programming, and leaves it to the hardware designer and maintenance personnel to deal with the basic hardware components.

Let us now trace some implications of our broad claim about the automaton character of evolution. Consider intentionality. Man, with his brain, hands, power of speech, and social organizations, is the most complicated and advanced organism of evolution we know of. He employs intentional control. Since evolution is selective, the presumption is that this method of control arose because it is an efficient way for an organism to adapt to its environment.

Our earlier computer analyses of direct-response and intentional systems makes this result plausible. Any finite automaton is equivalent to a finite number of direct-response rules, connecting input state and internal state to output state and next internal state. These rules constitute the state table which defines the step-by-step operation of

the machine. Moreover, there is a finite algorithm for converting any state table into a computer. That is, given a state table and a sufficient number of connectable computer components, an engineer can in principle construct a computer that functions according to that state table. In this manner, any automaton, and hence any intentional automaton, is in principle reducible to a direct-response automaton. An alternative but equivalent method would be to enter the state table into a general-purpose computer which could keep track of the internal states of the direct-response computer, and for each input state look in the state table for the next internal state and the output state.

But as any computer designer knows, this is not a practical way to design (or simulate) a computer, and the resulting computer would be terribly inefficient. Except for very small finite automata one cannot work with state tables and table-lookup because the combinatorial explosion from switches and memory cells to possible states is just too great. Rather, one must work with the logical structure of the components and the uniformities of blocks of these that enable one to move up and down the architectual levels of a design.

For an example, consider a finite automaton equivalent to Socrates, a highly intentional human. It follows from the claims made earlier that there is a direct-response automaton equivalent to him, but it would be terribly large and inefficient. That is why man evolved as an intentional, conscious, free being.

Similar efficiency considerations explain why teleological explanations are needed in practice even for reducible phenomena. The intentional account of Section 4 describes how a robot could operate in a goal-directed manner. A reductionist account would show how the robot goes through successive states related by cause and effect. But the simplicity, the generality, and the focus on essentials of the intentional account are not preserved by the reduction.

This completes the main part of our case for the reductionist theory of the teleological continuum, general considerations about automata theory and computer simulation, resting on the impressive results of biological science and modern genetics. But there are further issues that need to be addressed. These involve the human capabilities of intentionality, free choice, and consciousness. These have often been held to involve final causes in an essential way, and if they in fact do, the reductionist theory of the teleological continuum is wrong. I have

already dealt with the intentionality issue (Section 4). I will make a few comments about the free choice problem here, and then argue in the next section for the reduction of an important aspect of consciousness.

There are two traditional views about the free will issue, and they are correlated closely to the two theories of the teleological continuum. The *free will thesis* is that a free choice is partly uncaused, and that the absence of complete causality is essential to freedom and responsibility. This thesis is closely connected to the final cause theory of the teleological continuum. *Compatibilism* is the view that free choices only require inner conscious control, and that such control is compatible with determinism. This doctrine is closely associated with the reductionist theory of the teleological continuum.

Both theories are compatible with an indeterministic phenomenology of free choice. The act of choosing among alternatives is a spontaneous event. It is influenced by reasons and other conscious factors, but it is not determined by the totality of the free person's conscious contents. Relative to this totality the subject could have chosen otherwise.

The free person picks his or her goals and his or her means. Thus at the phenomenological and conscious level, "final causes" and "efficient causes" are on a par, as Peirce held. But the conscious free choice process is part of a larger context, a system consisting of the whole person, the relevant environment, and society. The issue between the believer in free will and the compatibilist concerns this whole system.[9] The free willer says that it cannot be deterministic, the compatibilist says that it can. Similarly, some believers in free will think that the human capacity for free choice depends on final causality, while compatibilists assume that efficient causes are sufficient to account for free choice.

I think the compatibilist's position is correct, so that logical mechanisms are capable of free choice, but there is not space to present my arguments here.

9. AN ARCHITECTURAL THEORY OF FUNCTIONAL CONSCIOUSNESS

In Section 4 I showed that consciousness goes beyond intentionality. Next we need to distinguish two aspects of consciousness: *functional consciousness* and *immediate experience*.

Consider an instance of pain. Suppose one's toe is injured. This is on the physiological side. On the experiential side, one feels a sharp pain in the toe, sees that the toe is bleeding, and puts a bandage on it. It seems to me that this experience of pain has two aspects: the felt pain as such, an immediate feeling of pain; and the experienced functional connection from pain as stimulus to the immediate experience of repair action as response.

Another example is taken from Karel Ĉapek's 1921 play R.U.R. ('Rossum's Universal Robots'). The heroine Helena Glory is upset by the fact that the robots have no self-interest. A robot does only what it is told. It has no desire to accomplish anything, not even to continue operating. Thus a robot doesn't care when it is told that it is to be dissected and then put in the stamping mill so its materials can be reused. Horrified by this, Helena persuades the chief psychologist Dr. Gall to change a "physiological correlate" so the robots will have goals and can look after their own welfare. The ultimate consequence is that the robots revolt. Goal-seeking and the will to live are functional. Whether the modified robots have the immediate feelings which accompany the exercise of these functions in humans is a question that Ĉapek does not address.

The revolt of the robots illustrates the fact that there are at least two fundamental aspects to a living organism, reason and will. These two aspects or factors are reflected in the title of Arthur Schopenhauer's book *The World on Will and Idea*. I think that the will is ultimately rooted in the replicating or copying feature of evolution.

I will now sketch a theory of functional consciousness in two stages. The nature of immediate experience and its relation to functional consciousness are equally important problems, but there is not time to address them here.

The functional aspect of conscious pain may be illustrated by the experience of lepers. Leprosy damages the nerves which carry signals from the periphery to the central nervous system. A leper may injure his toe, and because he feels no pain is not aware of it. Consequently, he does nothing to repair and protect it, and it ultimately deteriorates and falls off.

It has been known for a long time, both practically and theoretically, how to make computers that detect and correct their own errors and malfunctions. A robot could have circuits which detect the state of its appendages and send reports to the central processor, some

central control unit responsible for reliability, or to various regional units with this function. The responsible unit would decide on a method of correction and supervise its implementation: switching in an alternative circuit, transferring the job to another appendage, or even manually replacing the damaged part.

The functional aspect of sense experience lies in the organism's ability to use the information thus gathered from the environment and to respond appropriately to it. To be successful this ability must relate sense reports to one another and to thoughts and possible actions in various ways, and the sense reports must be of the proper generality to make these interrelations useful. This is the problem of pattern recognition – which I think will be solved – someday.

The will-to-live was lacking in Rossum's universal robots as they were originally manufactured, but was present after Dr. Gall changed a "physiological correlate". What did he do? Presumably he gave the robots desires by designing them so they would pursue various goals, including self-preservation. To deal with the problem of conflicting goals, Dr. Gall might have assigned weights or relative priorities to the different goals and organized the robot so that the amount of effort it devoted to a goal was influenced by the weight assigned to that goal.

The foregoing shows in a general way how to construct a robot that would perform the internal functions associated with pain, color experiences, and desires. This general design procedure could be applied to other types of conscious experience as well. On that basis I advance the following general claim: There could be a robot that performs the functional aspect of every specific type of conscious human experience.

It does not follow, of course, that this universal robot would be conscious. Our robotic depictions of pain, color experiences, and desires, as well as our earlier account of robot intentionality, contained nothing of consciousness in them. The universal robot would only have more of the same, and hence would be no nearer to being conscious. Moreover, a collection of specific functions would not have the unity of consciousness. Clearly, something essential has been left out. There must be some feature of consciousness or some particular way in which a human carries out its conscious functions that has so far been omitted from our robot account. We need to investigate what it is.

To fully comprehend a concept one might understand its relations to associated concepts and to incompatible concepts as well. The words "conscious", "aware", "awake", and "consciousness' are closely associated, while all are to be contrasted with "asleep", "unconscious", "comatose", and "the unconscious". A person may be awake (conscious), asleep, unconscious, or comatose. Consciousness is functioning in the first case but not in the case of sleep, or being unconscious, which differ in the difficulty with which consciousness may be restored.

We are now closer to understanding why a robot which can perform all the specific functions of human conscious experience and is, for good measure, intentional as well, might not be conscious. It might not have a single conscious subsystem carrying out all of these functions, a system which can be turned on (awake) and off (asleep). Our descriptions of the robot did not contain a description of such a subsystem, and did not even employ the distinction between waking and sleeping.

We need therefore to analyze the higher order functional role played by consciousness. Why do humans perform certain functions consciously, and hence only when they are awake? Is this the best way for robots to perform these functions? More generally, why are animals conscious – why are they sometimes awake and sometimes asleep, and should robots be designed similarly?

Before answering these questions we need to establish a basic fact about the "size" or complexity of consciousness. Consciousness is only a small "part" of the person, in the following sense. Consider the whole person, body and mind, as a system, and compare its complexity with that of the subsystem of consciousness. Consciousness is essentially vague, so that one cannot give a precise description of a present state of consciousness or the conscious rules governing its transitions. But it seems intuitively clear that an approximate description of these would be very much shorter than a correspondingly approximate description of the whole person. This intuition is confirmed by psychological measurements of the capacity of short-term memory, that is, the amount of information that is in consciousness at one time. Short-term memory can hold about ten items, while long-term memory can store many thousands. Assuming a parallelism between mental and bodily activity, the neural subsystem involved in consciousness is small compared to the whole body.

I am now ready to state my architectural theory of functional consciousness. A survey of the examples of functional consciousness analyzed earlier shows that they have a common property. All are ways in which the organism controls itself and what it does. Functional pain and color experiences involve short-term control, while intentionality is a computational procedure for long-term control. This observation, combined with earlier comments, suggests our theory of human consciousness: functional consciousness is a real-time system of relatively small capacity that exerts short-term control of the person and is capable of long-term, intentional control. In goal-directed behavior the details of a long-term plan are stored in main memory and accessed, used, and revised in short-term conscious memory.

Consciousness developed in earlier animals. We saw when analyzing the teleological continuum that evolution is gradual. Hence there is no sharp line dividing the preconscious from the conscious. Let us focus on the early mammals. Our theory explains the value of consciousness to them. Consciousness is a simple system for real-time control of their immediate interactions with the environment. But why should it be turned on for only part of the day-night cycle and turned off for the rest? Why does the state of being awake alternate with sleep?

A common answer to this question is that animals sleep so they can rest and repair themselves. But one can design robots that need no rest and can repair themselves without having a periodic sleep mode, and can even repair themselves while they continue to work. Hence we must look elsewhere for the answer to the question of why animals sleep.

Given the large differences between day and night, organs adapted to one would generally be different from organs adapted to the other. For example, an eye good mainly in daylight would be different from an eye good mainly at night, and one good in both environments would be more complex than either. Hence at some stage evolution would produce organisms that performed differently during the day and the night, in particular, that were active during one of these periods and inactive during the other.

The explanation of sleep I like best is that it reduces a mammal's vulnerability to predators during the inactive period. During this period the mammal would naturally hide to reduce its danger, perhaps in a hole, and predators would be searching for it. When a predator came near, the mammal's best strategy might be to remain immobile

and not respond at all. Sleep accomplishes this. As Carl Sagan expresses it: "Animals who are too stupid to be quiet on their own initiative are, during periods of high risk, immobilized by the implacable arm of sleep" (Sagan 1977, p. 131; Webb 1975; Meddis 1975). The annual hibernation of a bear or woodchuck serves the same function.

This explanation may not apply to all cases of sleep, and even as it stands it is incomplete. On it, the behavior mode to be achieved during the inactive period is that of being immobile, not responding externally, or "playing dead". The method of turning consciousness off during the inactive period and on again for the active period is one way of accomplishing this. Note that on this method the on-off switching between the active and the inactive mode of performance is shared by consciousness and the rest of the system. Consciousness can make decisions that lead to sleep, but it is not able to switch the organism directly into the sleep mode, and when the organism is very tired consciousness cannot prevent sleep.

An alternative method of achieving the inactive behavior mode is to turn the output off while leaving the input on. Consciousness would then have two modes of operation: the ability to make quick external reactions during the active period together with the ability to abstain from external reactions during the inactive period. Presumably evolution chose the former method because it was easier to develop. During the evolution from animals to primates and thence to man, consciousness added the function of intentionality, or rational control.

We have said that an intelligent robot could be conscious. But would that be an efficient design, and if so, should the robot's consciousness be made very much like human consciousness or quite different? These are organizational design questions, the answers to which depend very much on the nature of the technology used and design process. And these are very different in the cases of the robot and the human.

Human consciousness was constructed by a gradual evolutionary design process, beginning with primitive forms of consciousness, and if it could be redesigned ab initio there might very well be a much better design. Equally important, the sizes and speeds of hardware components are very different from those of biological components. Since human consciousness is a real-time control system of small capacity, these hardware-fleshware differences are relevant to how much in-

formation a robot consciousness should process (how large its short-term or cache memory should be) and how a robot consciousness should relate to the rest of the robot. Perhaps a good robot design would make its short-term memory very large and would devote much more computation to long-range planning than a human does. Note that these issues belong to the general subject of computer control, computer organization, and system management.

I have now sketched my theory of the functional aspect of human consciousness: its nature, its origin, and its function. The theory is computer based, consciousness being approached in terms of the organization of the human mind and body. Viewed from the perspective of computer architecture, human consciousness is a particular kind of computer control system, a relatively simple real-time control which, when the system is awake, directs short-term activities and plans longer-term activities.

Let us now apply this theory of consciousness to robots. We argued earlier that a robot could be built to perform the functional aspect of every specific type of conscious human activity, including intentionality, and raised the question of whether a robot might not be constructed so as to have a single conscious system that carries out all human conscious functions. The preceding analysis makes this plausible. I believe that someday it will be practical to build a robot capable of performing all natural human functions and to organize the control system of that robot in such a manner that the robot will be conscious. Thus the answer to the question "Can a robot be functionally conscious?" is *yes*.

10. ANCIENT AND MODERN EXPLANATIONS OF TELEOLOGY

The reductionist theory of the teleological continuum is a central thesis of the philosophy of logical mechanism (cf. Section 3). Our argument for this thesis (Sections 4, 8, 9) is incomplete on two points: the compatibilist account of free choice, and the reducibility of immediate experiences. Many others have argued for these later two claims, and we have done so elsewhere.[10] For the purpose of comparative history, let us assume that these two gaps have been filled, and return to the three ancient puzzles with which this paper begins.

Expressed on questions, these were: How can Plato's idea of the

Good operate on Socrates' reason to control Socrates' behavior? How do Aristotle's final causes actually control and direct the acorn's growth? What sense is there to Plotinus' Great Chain of Being, with its hierarchy of descending levels, each level having associated degrees of reality, value, and knowledge?

As a graduate student I came to philosophy with a background of mathematics and physics and some knowledge and interest in Christian religion. When I studied and taught these doctrines of Plato, Aristotle, and Plotinus I had difficulty in understanding them. I couldn't see how to develop or explain them or grasp the models used by those who did seem to understand them.

In this paper I have developed a computer model of human intentionality and, on this basis, suggested that human intentionality was the model for the force of Plato's Idea of the Good, for the directivity of Aristotle's and Peirce's final causes, and the model for the creative emanations of the Great Chain of Being. Plato and Aristotle, the originators of the concept of final causality, derived their understanding of teleology by using the Greek craftsman as a model.

Consider Aristotle's doctrine of final causality: final causes are needed to explain all purposive or goal-directed activity of the teleological continuum, final causes are as basic as efficient causes, and hence final causes are not reducible to efficient causes.

In the computer model of intentionality (Section 4) one can discern a final cause and distinguish it from the stream of efficient causes and effects. The final cause is the regulative goal plan. The stream of efficient causality is the sequence of sensory inputs, steps of internal information processing, decisions, and actions that are taken in the process of working towards the goal. I suggest that Aristotle was aware of this high level distinction of final versus efficient causality, and that to explain the growth of non-intentional systems like acorns he moved the distinction back down the teleological continuum to its early stages. Peirce followed Aristotle in this.

There is, however, a fatal weakness in any such attempt to explain how final causality works at the pre-intentional level. As one moves back down the teleological continuum, intentionality gradually weakens and finally disappears, and so does the regulative goal plan. Hence the intentional basis for final causality gradually disappears. In a direct-response teleological system final causality consists only of a

set of direct responses which tend to fulfill the goal, and there is no supervisory or directing plan.

It is nevertheless true at the *practical* level of macroanalysis that final causes are as basic as efficient causes. Final causes are essential at this level because a reductionist description of a final cause would be much too complex to work with. Aristotle and Peirce were correct to this extent. But it is a mistake of level to think that because final causes are on a par with efficient causes at the macro-level they are also on a par at the micro-level.

Finally, consider Plotinus' Great Chain of Being. When the transcendent part of the Great Chain of Being is removed the result is a teleological continuum, running downward from men through animals on down to simple cells, that is, running opposite in direction to the evolutionary teleological continuum (Section 5). According to Plotinus' doctrine, there is a gradual variation of degrees of reality, value, and knowledge (all in close association) along the continuum. What sense can be made of this doctrine in terms of our reduction of the teleological continuum to an evolutionary continuum of automata?

Each automaton is finite, and it has a measurable complexity. This complexity can be measured in various ways, but by any intuitively plausible definition of automaton complexity the evolutionary continuum involves gradually increasing complexity. The human is the most complex automaton of the continuum. He clearly has the greatest ability to acquire knowledge. Most of us believe that man's values are more important than the values of lower organisms. Thus the teleological continuum is a continuum of gradually increasing complexity, value, and knowledge, all in close association.

Now if we replace the term "complexity" by the term "reality", we obtain this result: the teleological continuum is a continuum of gradually increasing reality, value, and knowledge, just as the Great Chain of Being says.

The account I have given you of teleology is logical, mechanical, and reductive, in the general spirit of my philosophy of logical mechanism (Section 3). This account of teleology is very different in both language and substance from the idealistic and non-reductive teleological philosophies of Plato, Aristotle, and Plotinus. Still, I feel that there is enough formal and logical similarity between my account and theirs that I finally understand what they were driving at, and the ways in which they went wrong.

Let me express this conclusion about ancient and modern explanations of teleology in a different way. Philosophy is a field characterized by radical, long-standing disagreement even on the most basic matters – perhaps we philosophers are fated to substantial eternal disagreement. For this reason the following two criteria should be applied to each philosophical theory. First, that it give a plausible account of the phenomena involved, an account which is somewhat constructive and somewhat logical in structure. Second, that the theory explain why the opposing theory seemed plausible to those who advocated it.

The philosophy of logical mechanism shows how modern computer and genetic concepts can be used to reduce the goal-directedness of each system of the teleological continuum to efficient causality and also to reduce the evolutionary development of the teleological continuum to efficient causality. Furthermore, this reductive account explains why Plato, Aristotle, Plotinus, and Peirce advocated non-reductive theories of teleology. Each was, in his own intellectual context and in his own way, extending human intentional goal-directedness to nature as a whole. This was a simple and natural approach, and in ancient, medieval, and even modern times it worked about as well as the reductionist theory, for in those times neither theory could really explain teleological processes. This situation has changed drastically in the present century with the development of evolutionary genetics and theoretical and practical knowledge of computers.

NOTES

* A draft of this paper was presented at the Conference on Teleology in Natural Science, Center for Philosophy of Science, University of Pittsburgh, December, 1984. An expansion of Section 9 has been published in the proceedings of that conference under the title 'An Architectural Theory of Functional Consciousness', in Nicholas Rescher (ed.), *Current Issues in Teleology*, University Press of America, New York, pp. 1–14.

This paper was also presented at the Institute of Philosophy, Chinese Academy of Social Sciences, Beijing, and at the Institute of Psychology, Chinese Academy of Sciences, Beijing, in May, 1985. *Acta Psychologica Sinica* 19 (1987) 158–66.

This research was supported by National Science Foundation grants SES82–18834 and DCR83–05830.

[1] Brentano also believed that intentionality involves consciousness. While my reduction of intentionality does not cover consciousness, I offer a reductive account of consciousness later (Section 9).

[2] As in John Holland's bucket-brigade algorithm. See Holland (1985), Holland et al. (1986), and Burks (1988).

[3] For an interpretation of Peirce's rhetoric, see Burks (1981).
[4] There are many other references, including: 1.22, 1.204, 1.207, 1.231, 1.615, 2.66, 6.163, 6.320, 6.477, 6.582, 6.604, 8.2, and 8.318.
[5] For further analysis of tychism, see Burks (1977, Section 9.3.4).
[6] Peirce had a rather rich and strange sense of the mathematical continuum, one that incorporated a notion of the infinitesimal. He never made clear what this concept was, but I do not think the difference between his concept of the mathematical continuum and the standard one is relevant to his final cause theory of the teleological continuum.
[7] See Burks (1988). Section 11 argues that there is a holistic kind of causality, *inverse causality*, which contrasts with direct causality, but is reducible to a system of direct causality.
[8] See the analysis of embedded subsystems in Burks (1977, Section 9.4.1).
[9] Thus it is an issue concerning the relation of an embedded subsystem to its underlying system. See Burks (1977, Section 9.4.1).
[10] Salmon (1989, Sections 3.3 and 3.4 of my replies).

REFERENCES

Burks, A. W.: 1943, 'Peirce's Conception of Logic as a Normative Science'. *The Philosophical Review* **52**, 187–93.
Burks, A. W.: 1967, 'Ontological Categories and Language', *The Visva-Bharati Journal of Philosophy* **3**, 25–46. Visva-Bharati (University), Santiniketan, West Bengal, India.
Burks, A. W. (ed.): 1970, *Essays on Cellular Automata*, University of Illinois Press, Urbana.
Burks, A. W.: 1972, 'Logic, Computers, and Men'. *Proceedings and Addresses of the American Philosophical Association* **46**, 39–57.
Burks, A. W.: 1977, *Chance, Cause, Reason – An Inquiry Into the Nature of Scientific Evidence*, University of Chicago Press, Chicago.
Burks, A. W.: 1979, 'Computer Science and Philosophy', in Peter Asquith and Henry Kyburg (eds.), *Current Research in Philosophy of Science*, Philosophy of Science Association, East Lansing, Michigan, pp. 399–420.
Burks, A. W.: 1981, 'Man: Sign or Algorithm? A Rhetorical Analysis of Peirce's Semiotics'. *Transactions of the Charles S. Peirce Society* **16**, 279–92.
Burks, A. W.: 1984, 'Computers, Control, and Intentionality', in Donald Kerr, Karl Braithwaite, N. Metropolis, David Sharp, and Gian-Carlo Rota (eds.), *Science, Computers, and the Information Onslaught*, Academic Press, New York, pp. 29–55.
Burks, A. W.: 1986a, 'An Architectural Theory of Functional Consciousness', in Nicholas Rescher (ed.), *Current Issues in Teleology*, University Press of America, New York, pp. 1–14.
Burks, A. W.: 1986b, *Robots and Free Minds*, University of Michigan, Ann Arbor, Michigan.
Burks, A. W.: 1988, 'The Logic of Evolution, and the Reduction of Coherent-Holistic Systems to Hierarchical-Feedback Systems', in William Harper and Bryan Skyrms (eds.), *Causation in Decision, Belief Change and Statistics*, Kluwer Academic Publishers, Dordrecht, Holland, 135–91.
Holland, John: 1984, 'Genetic Algorithms and Adaptation', in O. G. Selfridge, E. L.

Rissland, and M. A. Arbib (eds.), *Adaptive Control in Ill-Defined Systems*, Plenum Press, New York, pp. 319–33.

Holland, John: 1985, 'Properties of the Bucket-Brigade Algorithm', in John Grefenstette (ed.), *Proceedings of an International Conference on Genetic Algorithms and Their Applications*, Carnegie-Mellon University, Pittsburgh, Pennsylvania, July 24–26, pp. 1–7.

Holland, John, Keith Holyoak, Richard Nisbett, and Paul Thagard: 1986, *Induction: Processes of Inference, Learning, and Discovery*, MIT Press, Cambridge.

Meddis, Ray: 1975, 'On the Function of Sleep', *Animal Behavior* **23**, 676–91.

Peirce, Charles: 1931, *Collected Papers of Charles Sanders Peirce*, Charles Hartshorne and Paul Weiss, editors of vols. 1–6, 1931–35; Arthur Burks, editor of vols. 7 and 8, 1958. Harvard University Press, Cambridge. References are given in the standard form, 8.272 referring to vol. 8, paragraph (not page) 272.

Sagan, Carl: 1977, *The Dragons of Eden*, Random House, New York.

Salmon, Merrilee (ed.): 1988, *The Philosophy of Logical Mechanism*, a festschrift for Arthur Burks, Kluwer Academic Publishers, Dordrecht, Holland.

Von Neumann, John: 1966, in Arthur W. Burks (ed.), *The Theory of Self-Reproducing Automata*, University of Illinois Press, Urbana, Illinois.

Von Neumann, John: 1986, *Papers of John von Neumann on Computers and Computer Theory*, William Aspray and Arthur Burks (eds.), MIT Press, Cambridge.

Webb, W. B.: 1975, *Sleep, the Gentle Tyrant*, Prentice Hall, Englewood Cliffs, New Jersey.

Department of Philosophy
2205 Angell Hall
The University of Michigan
Ann Arbor, MI 48109
U.S.A.