# Trends in Epistemology

Peter Wegner

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-91-31
April 1991

# Trends in Epistemology

Peter Wegner, Brown University, April 1991

**Table of Contents**

# Trends in Epistemology

Peter Wegner, Brown University, April 1991

**Abstract:**

This paper has two primary objectives:

(1) To review trends in epistemology, emphasizing twentieth-century developments but starting from Greek beginnings.

(2) To reexamine the criticisms by Searle and Penrose of Turing's hypothesis that machines can think.

Section 1 reviews Greek epistemology, section 2 considers rationalist and empiricist philosophies in the 17th-19th centuries, and section 3 examines trends in twentieth-century analytic philosophy. These sections provide a concise overview of the evolution of epistemology not available elsewhere in the literature, and a framework for the analysis of the debate concerning the Turing test in section 4.

Objectors to Turing's position that machines *can* think can be classified as *intentional* and *extensional* skeptics who respectively question the inner quality and extensional power of computation. Searle [Se] is an intentional skeptic, believing that computers cannot think because symbol processing is an inherently mindless activity. However, his view ignores the fact that neurons are just as mindless as computational agents. Intelligence resides in organizational structure, while primitive agents are almost always mindless. Soldiers in an army or genes in a chromosome perform their function without knowledge of its purpose.

Penrose is an extensional skeptic believing that mental behavior cannot be completely modeled by computers. In *The Emperor's New Mind* Penrose suggests that incomplete physical laws can be completed only by making them noncomputable. We question Penrose's hypothesis on the following grounds:

(1) Penrose's hypothesis is not the only one compatible with the evidence. We suggest two alternative hypotheses: nonlinear laws (chaos) that are computable in theory but not in practice, and infinitely scalable (Mandelbrot) universes whose particles at one level of scale are galaxies at another level of scale.

(2) The assumption that "complete" laws exist is based on a false analogy with mathematical models. For formal systems, completeness can be defined in terms of a correspondence between truth and derivability of formulae. There is no notion corresponding to truth for physical models and therefore no notion of completeness.

(3) The existence of a reality that can be "completely" modeled is questionable. Do laws of mechanics, electromagnetism, optics, and acoustics determine physical properties like texture and appearance? Is the behavior of neurons modeled by the same physical properties as the behavior of mindless particles? We can model physical properties that determine constraints on reality, but reality itself cannot be modeled or even defined.

In exploring the evolution of epistemology we consider the following kinds of questions:

- What kinds of explanations of natural phenomena are offered by philosophy?
- What are the roles of reasoning and observation in acquiring knowledge?
- What is the status of primitive assumptions (laws or axioms) of an empirical model?
- What role should deduction and induction play in model construction?
- What do we mean by causality (for example, does gravity cause motion)?
- What are the relations among physical, mathematical, mental, and computational knowledge?

## 1. Greek Epistemology

Greek thought owes its distinctiveness to a new epistemological paradigm whose fundamental assumptions about the nature of knowledge are still accepted in the late twentieth century. The questions asked by early (Pre-Socratic) philosophers on the constitution of matter, the nature of mathematics, and the nature of change have a remarkably modern flavor.

### 1.1. The First Epistemologists

Thales is known for his prediction of an eclipse of the sun in 585BC and for his belief that everything is made of water. His successor Anaximander believed that everything is made of a primordial substance, that life arose from moisture evaporated by the sun, and that man is descended from the fishes. Anaximenes believed air to be the fundamental substance, fire to be rarified air, and water, earth, and stone to be progressively more condensed versions or air.

These speculative hypotheses, motivated by the desire to account for the observed properties of matter, were not systematically verifiable. But the view that everything is made of water is not too different from the theory that everything is made of hydrogen, while the view that fire, air, earth, water, and stone are progressively more condensed versions of the same substance is a plausible, empirically motivated account of the nature of matter.

Pythagoras believed that the structure of objects is determined by numerical relations among their parts. He relied on reason as the means for acquiring knowledge about the world. The discovery by his disciples that certain numbers, such as the square root of 2, could not be expressed as the ratio of integers eventually discredited Pythagoreanism. But his idea that reasoning rather than observation is the basis for true knowledge greatly influenced future philosophy.

Heraclitus and Parmenides held opposing views on the nature of change. Heraclitus believed that fire was the primordial element that caused all things to be continually changing. Parmenides thought change was logically impossible because words and thoughts have a constant meaning that can only describe unchanging phenomena. Sense impressions were an illusion because of their transitory nature. The paradox of Achilles and the tortoise (a problem of limits of infinite series) is due to Zeno, a disciple of Parmenides. Parmenides' views were influenced by Pythagoras and in turn strongly influenced Plato.

Empedocles identified air as a separate substance by observing that water does not displace air in submerged containers. He viewed air, earth, fire, and water as separate elements that could be combined in different proportions to produce earthly substances. He believed that the moon shines by reflected light, that light travels at a finite speed, and that evolution occurs by survival of the fittest.

Leucippus and Democritus developed an atomistic model of the physical world constructed of indestructible atoms of different shapes and sizes separated by empty space. Their ideas of matter, space, and motion are not unlike Newton's but lack Newton's empirical and mathematical infrastructure. Atoms are in motion and may collide like billiard balls, generating vortices which generate bodies and eventually worlds. There are many worlds, some growing, some decaying, some with many moons, and some with no moons. Although their model has a modern flavor, it is based on lucky speculation rather than on empirical foundations.

## 1.2. Plato and Aristotle

Plato thought sense impressions were an illusion because true knowledge required knowing the unchanging abstract essence of objects. Just as appreciating beautiful things is inferior to knowing the concept of beauty, so experiencing particular objects like tables is inferior to knowing the underlying idea of tablehood. His parable of the cave suggests that those who believe in the reality of sense impressions are like prisoners in a cave who perceive reality only by shadows on its walls, confusing these shadows with reality.

Abstract concepts like beauty and tablehood that capture the unchanging essence of classes of perceived phenomena are called *universals*. If we believe with Plato that universals are more real than their instances, we downplay the role of sense impressions in acquiring knowledge about the world. If we believe that specific tables are more real than universals, we have an entirely different attitude to the construction of models. The question whether universals are real affects our attitude to a large class of related questions that together determine a paradigm for interpreting and managing knowledge.

Are universals real, or are they simply auxiliary scaffolding, like constructions in a geometrical proof, that facilitate reasoning without being existentially necessary? Are the abstractions we use to formulate a theory more real than the enntities that the theory is about, or are the entities being modeled the true reality? This is a pervasive epistemological question that is relevant today in interpreting the relation between physical theories and the ''reality'' they describe (see section 4.3.). Plato's view that universals are real while their instances are an illusion accords with the fact that the primitives of physical science like mass, energy, electrons, and genes are unobservable abstractions having a very indirect connection to everyday objects.

Aristotle ascribed greater importance than Plato to objects and less to abstractions. He viewed universals like tablehood as a convenient way of describing properties common to many objects, but did not believe that universals existed apart from their embodiment in specific objects. This elevation of the status of observed objects over abstract universals is conducive to the development of empirical science, and Aristotle in fact contributed to almost all scientific disciplines.

Aristotle made extensive empirical tabulations of properties of plants and animals and developed models for their classification. His treatises on physics and cosmology present models of natural phenomena that are firmly in the empirical tradition. His contributions to logic included the method of reasoning by syllogism:

''Socrates is a man'' and ''all men are mortal'' implies ''Socrates is mortal''

Aristotle's classification techniques span both mathematical and biological classification. The syllogism may be viewed as way of reasoning about class membership (''all men are mortal'' asserts that men are a subclass of the class of mortals). But his more important contribution to taxonomy lies in the development of systematic classification techniques for plants and animals based on the observation of structural similarities:

> it is by the resemblance of the shapes of their parts, or of the whole body, that the groups are marked off from each other.

Aristotle strongly distinguished between logical classification by dichotomies and physical classification into groups with common structure. In differentiating between logical analysis based on reason and empirical investigation based on observation, Aristotle nicely balanced mathematical and empirical techniques of investigation.

The Platonic view that knowledge is acquired by reasoning is called rationalism while the view that knowledge is acquired through empirical observation is called empiricism. Rationalism and empiricism determine different paradigms of modeling [Ku] in the sense that their criteria for model construction and for evaluating the adequacy of models are very different. The choice among these paradigms has a profound effect on the kinds of models that are considered worth while. The dominance of Platonism during much of the Middle Ages impeded the development of science because of the Platonic aversion to observation. The espousal of Aristotelianism by Thomas Aquinas in the 13th century and William of Ockham in the 14th century helped to create a climate for the emergence of science in the 16th and 17th centuries. Ockham developed ''Occam's Razor'', the principle that if two theories or models have equal explanatory power the simpler one is to be preferred.

## 2. Rationalism and Empiricism

Francis Bacon and Rene Descartes respectively initiated modern empiricism and rationalism. We trace the evolution of this tradition through the empiricism of Locke, Berkeley, and Hume, the idealism of Kant and Hegel, and 20th-century formalism and common sense.

### 2.1. 17th Century Epistemology

Bacon's identified induction as a basis for inferring that regularities of finite sets of observations are true in general. He was not satisfied with induction by simple enumeration as a method of inference (we cannot infer that all humans speak English by observing that one thousand Englishmen speak English). To discover the nature of heat, which he believed to consist of small particles, he made lists of bodies of varying degrees of heat to determine features common to hot and cold bodies, hoping by this method to arrive at general laws.

Bacon's empirical philosophy in *The Advancement of Learning* is in some respects remarkably modern. His assertion that ''we ought to be neither like spiders which spin things out of their own insides, nor like ants which merely collect, but like bees which both collect and arrange'' is incomplete, since induction requires a creative act beyond collecting and arranging. But the view that science proceeds by collecting and arranging, supplemented by incompletely understood creative acts like induction, is close to the modern view of scientific method.

Descartes' dictum ''cogito ergo sum'' states that thinking rather than observation is the key to acquiring knowledge. Descartes was a dualist, believing that mind and matter were two entirely different kinds of substances, with mind being more fundamental than matter. Thinking could continue even when a person's body ceases to exist, but existence was not possible without thinking.

Descartes' reduction of geometry to numbers (Cartesian coordinates) was both an important mathematical contribution and a triumph of the rationalist scientific method. The reduction of physical to mathematical phenomena lies at the heart of rationalist philosophy. Descartes viewed Euclidean geometry as a physical theory and Cartesian geometry as a reduction of a physical to a mathematical theory. This practical demonstration of the power of rationalism rested on a misconception that was not dispelled till the development of noneuclidean geometry in the nineteenth century.

Descartes' rationalism appears to conflict with his dualism. Dualism asserts that matter and mind are distinct substances, while rationalism asserts that the physical properties of matter can be determined by reasoning, and that physical phenomena are in this sense reducible to mental phenomena. This contrast between dualism of the substance of minds and bodies and monism at the level of reduction of physical to mental phenomena is difficult to reconcile.

The scientific ideas of Copernicus, Kepler, Galileo, and Newton provided new directions for philosophy. Copernicus, Galileo, and Kepler each dispelled deepseated beliefs that stood in the way of developing empirical theories of natural phenomena:

(1) Copernicus dispelled the notion that the earth is the center of the universe, replacing it by the idea that planets rotate around the sun. This required a profound change in the view of the role of man in the universe.

(2) Kepler dispelled the view that the natural motion of heavenly bodies is circular, replacing it by the notion that planets move in ellipses. While this idea was not as threatening as Copernicus' idea, it required dislodging deepseated Ptolemaic prejudices.

(3) Galileo dispelled the idea that the natural state of objects is rest, replacing it by the idea that the natural state is uniform motion. This completed the dethronement of man from the center of the universe by demonstrating that we cannot distinguish whether the earth is at rest or in uniform motion. Note that empirical evidence generally contradicts Galileo's hypothesis: carts quickly come to rest when no longer pulled by horses and ships come to rest when there is no wind to propel their sails. Galileo's hypothesis is about an ideal (Platonic) world rather than about the actual world of appearances.

Once these preconceptions were dislodged, Newton was able to develop a unified theory, showing that the force of gravity could account for both planetary and terrestrial motion. Newton's laws of motion have the flavor of Euclidean axioms but were viewed by Newton as justified by their explanatory power rather than by their a priori inevitability. This view of "laws of science" constitutes the *hypothetico-deductive method.* The hypothetico-deductive method makes use of Baconian empiricism to determine basic postulates and chooses these postulates to allow deduction of the complete range of natural phenomena from a small number of assumptions.

Newton's synthesis focused philosophical attention on the nature of scientific explanation. Philosophers found hypotheses like "Force = Mass * Acceleration" difficult because this relation between unobservable abstract entities does not directly correspond to sense impressions. The hypothesis that these abstract entities satisfy a relation that is always true goes beyond observing that the law is true in specific instances. Scientific explanation in terms of unobservable entities like force and mass required new forms of philosophical justification beyond Bacon's collecting and arranging or Descartes' rationalism.

## 2.2. Locke, Berkeley, and Hume

The relation between raw sense impressions and abstract concepts occurring in scientific theories was a central preoccupation of the empiricist philosophers Locke, Berkeley, and Hume. Locke's *Essay Concerning Human Understanding* (1690) admits intuition, reason, and sensation as sources of human knowledge. The goal of determining the degree of certainty of different forms of knowledge, formulated by Locke, became a focus for investigation by Berkeley and Hume. Locke asserts that the origin of ideas is sensations (sense impressions), that sensations are transformed by the mind into perceptions, and that complex concepts are the result of combining atomic sensations. The mind is viewed as a blank tablet (tabula rasa) at birth that is furnished through experience with an increasing store of knowledge, ideas, and understanding.

Berkeley believed that knowledge is based on sense impressions, but was an idealist in denying the existence of matter. In his *Dialogues of Hylas and Philonius* (1713), Hylas' realist view of educated common sense is demolished by Philonius who argues that we cannot see the causes of color, hear the causes of sound, or infer the causes or sense impressions. Philonius "proves" to Hylas that the causes of sensations of heat, taste, color and even concrete objects must be subjective.

Since the reality of sensible things consists in their being perceived, Berkeley concludes that unperceived things have no independent existence. In the absence of God's omniscience, material objects would at best have an intermittent life, coming into existence when someone perceives them but ceasing to exist when no one is looking. Fortunately, God perceives all objects even when there is no human observer, so that objects do in fact have a continuous existence.

There are two kinds of idealists: *subjective idealists* who believe that reality is a subjective illusion in the mind of each individual and *objective idealists* who believe that reality is a shared illusion (for example, because it is in the mind of God). Berkeley is an objective idealist.

Hume took the skepticism of Berkeley to its logical conclusion, denying not only the existence of an external world but also the existence of a ''self''. In his *Treatise on Human Nature* he distinguishes between raw sense impressions and ideas in the mind derived from them. Simple ideas correspond to sense impressions, while complex ideas created from simple ideas need not correspond to possible objects (for example, a winged horse). Universals like ''tablehood'' are likewise composed by generalizing from instances of tables to create abstractions that do not correspond to any real object. Hume goes beyond both Berkeley and Descartes in repudiating the self on the grounds that it has no independently observable existence apart from the bundle of perceptions it sustains:

> For my part, when I enter most intimately into what I call myself, I always stumble on some perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch myself at any time without a perception and never can observe anything but the perception.

Causality, like the self, has no independently observable impression from which the idea can be directly derived. We can perceive that A is above B or that A always precedes B but never that A causes B. Hume rejects inferences of causality based on the constant conjunction of events, and therefore rejects induction as a basis for generalizing from experience or predicting the future.

Modern philosophers of science agree with Hume that causation cannot be inferred from mere conjunction of events, but believe that statements like ''force causes acceleration'' and ''heat is caused by the motion of molecules'' reflect a legitimate notion of causality based on stronger evidence than the mere conjunction of events. Such causal statements express relations among a carefully constructed network of unobservable abstract concepts chosen for their explanatory power. Causality is justified by the wealth of empirically verifiable consequences of the hypothesis of causation. Statements like ''force causes acceleration'' are viewed by behaviorists as abbreviations for infinite collections of behavioral statements about observable phenomena about billiard balls, planets and other objects. Causality and inductive inference lead to useful scientific theories that may be pragmatically justified by principles such as Occam's razor. The step from thinking of causality as an abbreviation to thinking of causality as something real is thus methodologically questionable, but is pragmatically useful and agrees with common sense.

## 2.3. Kant and Hegel

Kant countered Hume's skepticism by arguing that some empirical knowledge can be certainly known because of the structure imposed on sense impressions by the mind. Knowledge is dependent on experience, but not entirely derived from experience. The mind is a filter that strongly determines our perceptions (if fishnets have a mesh of one centimeter then fishermen will conclude that all fish have at least this size). Kant's *Critique of Pure Reason* makes use of this principle to show that some synthetic a priori propositions like ''the earth revolves around the sun'' can be known a priori. Euclidean geometry and Newton's laws of motion are true because of the structure imposed by the mind on sense impressions. The fact that both have been shown to be wrong (only approximately true) does not invalidate Kant's position, but it merely alters the structure of a priori synthetic perceptions. Induction is likewise a valid mechanism for inferring laws of physical causality because it reflects a priori mental structure. Kant's criterion for deciding which synthetic propositions are a priori does not seem precise, but includes induction and physical laws.

Kant's argument appears to demolish Hume's view that we cannot know synthetic propositions, but at the price of changing their information content. Kant does not believe we can have certain knowledge of reality itself; his notion of "synthetic" refers to mental rather than real phenomena. His acceptance of synthetic a priori propositions counters Hume's skepticism about empirical knowledge by eliminating it altogether and replacing it by mental knowledge. Thus Kant's attempt to realize empirical ceertainty leads him to an idealist position. Kant's synthetic propositions provide knowledge about the structure of our minds rather than about the structure of reality. Kant's position is close to Berkeley's position of objective idealism, substituting Kantian categories for the mind of God as the basis for a shared (objective) view of the world. Kant did not have the tools to examine the nature of the categories imposed by the mind on reality; their nature is only now beginning to be empirically investigated by cognitive and neural scientists.

Hegel is more extreme (his supporters would say consistent) in his idealism than Kant. According to Hegel, the rational is the real, and this reality takes the form of an absolute, logically consistent, interconnected world of ideas. The "real" real world is a network of necessary interconnections among objects related to each other in a way that can be deduced by logic. Hegel's ideal absolute world can be "logically" derived from the principle that it is not contradictory. However, Hegel's view of logic is a good deal less formal than the modern idea of a formal system. Its principal form of inference, dialectical reasoning, is totally imprecise about the kinds of inferences it permits.

A dialectical reasoning step consists of a thesis and antithesis yielding a synthesis. This dialectic process was elevated into a fundamental process for acquiring knowledge of an absolute reality. Any partial thesis about this absolute reality has an antithesis that, combined with the thesis, yields a synthesis that is a more complete description of reality. For example, the thesis "the absolute is pure being" and the antithesis "the absolute is nothing" leads to the synthesis that "the absolute is becoming" (the union of being and not being). Further dialectical steps concerning the nature of what becomes gradually reveal the properties of Hegel's absolute reality by pure reasoning. The principles for choosing the antithesis and the kinds of synthesis that may legitimately be inferred are imprecise, and it is this imprecision that gives Hegelian logic its power.

The dialectic mode of reasoning appealed to materialists like Marx, who applied Hegel's "dialectical idealism" to develop a materialist social theory, arguing that violent (revolutionary) change of society by the clash (synthesis) of opposite forces was an inevitable consequence of dialectical evolution. The British reaction against Hegel was more intellectual. Bertrand Russell and G. E. Moore, after an early infatuation with Hegelian idealism, adopted the view that common sense reality is in fact the true reality. Russell describes his "dehegelization" as a form of emancipation that transformed a world that had been thin and logical into one that was rich and varied and solid.

In retrospect, the enormous intellectual and social influence of Kant and Hegel is difficult to explain. Both theories appear simplistic and unscientific from the standpoint of modern ideas. Kant's synthetic a priori knowledge seems almost a cruel trick in its realization of certainty by making the objects about which we are certain subjective. Hegel's idealism seems even less securely based, but his problematic dialectical method provided an impetus for the most significant social experiment of the 20th century. The influence of epistemological theories does not seem to depend on their intellectual soundness.

## 3. Twentieth-Century Analytic Philosophy

In the twentieth century, logical positivism redefined the role of philosophy to be the clarification of language and concepts. This was reflected in mathematics in an emphasis on formalism. Philosophies of common sense and of mental phenomena retained the more ambitious goal of making substantive pronouncements about the nature of the world, but believed analysis was the fundamental tool of philosophical investigation. We examine the following topics:

(1) philosophies of common sense

(2) philosophies of mental phenomena

(3) mathematical philosophies

(4) logical positivism (logical empiricism)

### 3.1. Philosophies of Common Sense

G. E. Moore advocated common sense realism as an antidote to the idealism of Hegel. He had in fact started out as a Hegelian, passed through a phase of Platonic realism and moved towards an increasingly strong form of material realism culminating in the philosophy of common sense. Moore thought philosophy should describe the things we know in the universe and the relations among them. The starting point for such a description should be the beliefs we take for granted called common sense. Common sense affirms the existence of material objects and the existence of consciousness and of other minds. The evidence for the common sense belief in material objects and other minds is overwhelming and confirmed every day by experience. Compared to the certainty with which we hold common-sense beliefs, the arguments of skeptics who deny the existence of material reality are at best uncertain. Common-sense knowledge that one is holding a pencil or that one had breakfast this morning is more certain than the abstract arguments used to prove such knowledge uncertain.

According to Moore, the statement "we do not know that there are material objects" is contradictory because the "we" implies that the speaker and other persons exist. He proved the existence of his hands in a lecture by holding them up and pointing with one to the other, and then proved the existence of memory by reminding his audience that he had previously held up his hands. He examined the role of philosophical analysis in determining the meaning of sentences like "this is a human hand". Moore's view of the role of philosophy as a tool for analysis and clarification is in the mainstream of analytic philosophy.

The notion of common sense has been made more precise in the last 20 years by knowledge-representation models of artificial intelligence. These models formalize the common sense model that each of us has of how people and objects behave, based on logical reasoning from the following kinds of premises:

persons feel pain when pricked by a needle
persons who feel a sharp pain tend to wince
persons in pain take actions to remove the cause of pain

Such common-sense rules allow people to predict that a person pricked by a needle may wince. We use literally thousands of such rules in predicting human behavior. These rules collectively form the basis of common sense reasoning and are sometimes referred to as "folk psychology". Common sense may be viewed as a theory that, like a physical or mathematical theory, establishes meaning by a network of relations among terms. But the number of terms of common-sense reasoning and of relations among them is much larger than that of a physical theory, and the nature of reasoning is more informal.

Basing the meaning of terms like pain entirely on their relation to other terms ignores the inner qualitative nature of the sensation of pain. Theories of common sense must be empirically based on direct qualitative sensations, referred to as qualia. Qualia like pain are more directly perceived than experimental observations of physics, but visual qualia like redness or emotional qualia like jealousy do not have the directness of pain. Thus theories of common sense, just like physical theories, are based on both direct sensations and interpreted perceptions.

Although theories of common sense are built up from shared human perceptions, it is possible that organisms with qualitatively different sensory mechanisms, like martians or intelligent computers, could have precisely the same theory of common sense. The qualia of such organism would be isomorphic to ours while being qualitatively different. In logical terms this corresponds to the idea that the theory of common sense admits many different isomorphic interpretations. Beings whose qualia determine isomorphic internal structures in dealing with common sense are much more like humans than beings or mechanisms that merely behave identically.

Common sense provides a shared model of the world that serves as a basis for human communication. But its model of the world is very different from that of scientific theories whose primitives are abstract objects like force and mass. One of the weaknesses of common sense theories is the difficulty of mapping common sense into scientific theories. Another more practical weakness is the enormous number of common-sense hypotheses needed to formalize common sense for even the simplest tasks.

## 3.2. Philosophies of Mental Phenomena

Kant's view that the mind acts as a filter on sense impressions suggests that focusing on the nature of this filter is a key to perception. This motivated Husserl. the founder of phenomenology, to develop a philosophy of mental processes that captures systematic introspection. Phenomenology systematically suppresses (disconnects) the common sense view of the external world to obtain a clearer picture of the inner mental world. Husserl calls the common sense world the *natural standpoint*. To introspect we must disconnect the natural standpoint by mentally focusing on self-awareness of our inner processes of consciousness. The object of introspection is referred to as the "cogito", suggesting an analogy of the introspective method to Descartes' method of extreme doubting. However, Descartes repudiates the natural (common sense) standpoint, while Husserl merely suspends or disconnects it.

Husserl emphasizes the importance of studying mental phenomena and develops what he believes to be a systematic technique for undertaking such a study. He views understanding of mental phenomena as a key to understanding the nature of the synthetic a priori propositions hypothesized by Kant, which in turn is the key to understanding the nature of our perceptions of the physical world. In principle, adopting multiple views of natural and mental phenomena is neutral with regard to realism and idealism. However, Husserl's later work on "pure phenomenology" tended increasingly towards idealism, adopting the view that consciousness was the primary reality.

The emergence of psychology and cognitive science permitted Husserl's introspective approach to be replaced by systematic empirical investigation. The scope of philosophy was correspondingly narrowed, but certain problems, like the mind-body problem, still remained a legitimate philosophical concern. Four appraoches to the mind-body problem are briefly described below:

(1) Dualism: Mental phenomena are nonphysical (not subject to physical laws).
The two primary forms of dualism are substance dualism, which views the mind as a nonphysical substance and property dualism, which views mental attributes as nonphysical

properties of brains.

(2) Behaviorism: Mental phenomena are simply the behavioral effects of physical processes. The term "mind" is an abbreviation for a rather complex set of potential behaviors, and minds have no existence apart from the behavior to which they give rise. Behaviorism is a form of monism that dominated psychology in the first half of the 20th century and motivated Turing's 1950 definition of what it means for machines to think.

(3) Materialism: Mental phenomena are reducible to physical states of the brain. The simplest version of materialism asserts that mental states may be directly reduced to physical states. However, empirical research suggests that the relation between mental and physical states is indirect, with a given mental state being physically distributed in the cerebral cortex. Conversely, a given physical state persisting over time could be a component of multiple mental states.

(4) Functionalism: Minds have mental (cognitive) states and are definable by their state transitions, their external behavior, and their effects on the brain and body. Functionalism does not eliminate the mind, but defines it as an unobservable entity, like electrons in physical theories, is defined in terms of its effects on itself and on other unobservable entities. It ascribes cognitive states to the mind, and defines the effect of a given stimulus in terms of its effects on the brain, body, and mental state:
(mental state; stimulus) → (effect on brain, body, and environment; new mental state)

According to functionalism, people construct models in their minds having cognitive states with the same relational structure (state transition relations) as the structure being modeled. The mind therefore functions as a computer, and the model of any particular object or situation serves as a program. Cognitive models are in a "high-level language" and must be translated into the "machine language" of the mind for execution. The term "functionalism" for this view of the function of the mind was introduced by Craik in the 1940s [Cr].

## 3.3. Philosophies of Mathematics

Mathematical philosophy in the early 20th century was influenced by the success of non-Euclidean geometry in finding interesting new interpretations of the Euclidean axioms by abandoning the idea that they were self-evident. The view of Euclidean axioms as formal hypotheses about mathematical structures as opposed to self-evident truths about the real world marked a profound paradigm shift concerning the role of formal models. The view that models are intended to capture properties of the real world was replaced by the view that they should explore the consequences of interesting hypotheses independent of their connection to reality. However, in the second half of the twentieth century, there has been a reaction against formalism, sparked in part by Godel's incompleteness theorem and Turing's demonstration that there are noncomputable functions. We consider four views of the role of formalism in mathematics:

(1) Logicism: the view that mathematics can be reduced to logic.

(2) Formalism: the view that mathematics is concerned solely with formal derivation rather than with the inherent meaning of mathematical objects that are being modeled.

(3) Intuitionism: the view that mathematical reasoning must be justified by intuition and that modes of reasoning not justified by intuition should be excluded.

(4) Platonic realism (model theory): the view that mathematical phenomena have an independent existence that can be mathematically captured by interpretations of a formal system in a semantic domain.

Russell and Whitehead's *Principia Mathematica* (1913) set the tone for twentieth-century philosophy in its attempt to reduce mathematics to logic. This task was motivated by a desire for certainty in mathematics that parallels the quest for certainty of physical theories by Hume and Kant. It fits in nicely with the goal of clarifying concepts in a discipline, in this case mathematics, by reducing its concepts to a more primitive logical framework. To accomplish this is was necessary to extend traditional logic to handle numbers so that the propositions of arithmetic could be deduced from primitive propositions about numbers. Russell borrowed from Frege the idea of representing numbers as classes, representing numbers like 13 by the class of all things having 13 members. The properties of numbers are then described by a logical set of axioms for number classes from which theorems like "2 + 2 = 4" can be derived.

Russell's logic included a theory of types for avoiding paradoxes like "Is the class of all classes that are not members of themselves a member of itself?" The key idea in resolving the paradox is that the class of all classes that are not members of themselves is not an entity of the type for which we may meaningfully ask whether it is a member of itself. Russell's theory of types, like type systems in modern programming languages, restricts the arguments of functions so that they are meaningful only if they belong to a designated type called the domain of the argument. Assertions (or programs) that specify the application of functions to inappropriate arguments of the wrong type are meaningless.

Formalism, associated with David Hilbert, views the primary task of mathematics to be that of exploring the consequences of arbitrary axiom systems. This approach had liberated Euclidean geometry from the constraints of Euclidean space, and had been successfully applied in the late 19th century to the formalization of the real numbers and the differential calculus. Hilbert viewed mathematics as a game of formal manipulation that should be played without any regard to the meaning of what was being manipulated. He applied this approach with great success to geometry and to the cleaner formal development of many branches of mathematics. But the limitations of the formalist method were demonstrated when Godel in 1930 showed that any mathematical theory formalizing natural number arithmetic is incomplete in the sense that it contains true theorems that cannot be proved within the theory itself. Thus any sufficiently powerful formalist theory requires a more powerful metatheory to talk about the truth and falsity of its assertions.

The intuitionist view of mathematics, associated with Brouwer and Heyting, asserts that traditional formal reasoning is inadmissible because it goes beyond mental constructs derived from intuition about natural numbers and time. In particular, logical laws such as the excluded middle (that every proposition is either true or false) are inadmissible. Existential arguments like "there exists an x such that ..." go beyond intuition. Mathematical reasoning based on intuitionism is weaker than traditional formal reasoning and leads to a form of constructive mathematics that permits only constructive and not existential proofs.

Platonic realism, the oldest of the philosophies of mathematics in this section, has again become popular in the second half of the twentieth century as a reaction against formalism. It accepts the existence of an independent mathematical reality whose entities are discovered rather than invented. The Platonic mathematical world is essentially richer than the formal worlds of Russell, Hilbert, or Brouwer, and provides scope for insight and creativity in the discovery of mathematical truths and mathematical phenomena.

Platonic realism can be mathematically captured by models that interpret the syntactic primitives and relations of a formal system in a semantic domain. A formal system may in general have many models, including a standard model and many nonstandard models. For example, Peano's axioms for the integers have a standard model in which the numeral "0" is interpreted as the number zero. But there are also an infinite number of nonstandard models (satisfying the axioms) in which "0" is interpreted as some other integer.

Mathematical model theory views formal systems as constraints that guarantee that phenomena being modeled have certain properties but have nothing to say about the essential nature of modeled phenomena. The view that theories are constraints on modeled phenomena can be carried over to the interpretation of physical theories. Newtonian mechanics or quantum theory determine very limited properties of physical reality (the motion of particles) but tell us little about other properties like texture and appearance and even less about reality itself (whatever that may mean). The relation of formal systems to mathematical reality and the weaker relation of physical theories to physical reality is further examinrd in section 4.3.

## 3.4. Logical Positivism (Logical Empiricism)

Logical positivism scorns traditional philosophical subdisciplines like ethics and metaphysics, while venerating logical analysis and empiricism. Logic (including mathematics) and empirical science are viewed as the only respectable cognitive activities. Wittgenstein's *Tractatus Logico-Philosophicus* (1921) powerfully presents a form of logical positivism. Its power derives from a direct but simplistic mapping between the real world, its logical representation by propositions, and thoughts representable by "pictures":

The world is all that is the case
The world is the totality of facts, not things
A fact is the existence of states of affairs
A state of affairs is a combination of objects
The elements of a picture are the representations of objects

Wittgenstein viewed the world as a set of facts composed of states of affairs (atomic facts) representable by propositions that are logically independent of each other. Meaningful sentences express propositions about facts by propositions whose truth or falsity is determined by the truth or falsity of atomic propositions. Propositions are the expression of thoughts and thoughts are logical pictures of facts. Wittgenstein did not specify precisely what he meant by pictures, but he clearly had in mind that mental representations are in some way picture-like.

The *Tractatus* presents a narrow but elegant model of the world in terms of a well-known logical system of limited power that is essentially the propositional calculus. Although Wittgenstein admits facts about the world as meaningful propositions, his model is entirely deductive. He shares with Humian skeptics the desire to exclude induction and other empirically derived mechanisms from the permitted ways of acquiring knowledge about the world. Wittgenstein's view that metaphysical propositions are meaningless, which makes the modeled world tidy by narrowing the kinds of phenomena being modeled, became a tenet of logical positivism. His direct and simple model exerted great influence on philosophers during the 1920s and 1930s, although the model is imprecise and unduly restricted from the perspective of the 1990s.

The Vienna Circle (1923-1938) is a group of philosophers at the University of Vienna who developed and popularized the philosophy of logical positivism. They had a common view of philosophy as being concerned with clarification of statements rather than the propounding of new ideas. That is, philosophy should not make pronouncements about the nature of the world, but should rather contribute to clear thinking by the analysis and elucidation of the propositions of others. They believed philosophy to be analytic rather than synthetic, emphasizing logical analysis but not ignoring the prevailing empirical paradigms of science. For every discipline X, the philosophy of X should complement the work of practitioners of X by clarifying and exploring the nature of the X discipline. The twin goals of logical analysis and empiricism led to the philosophy of logical empiricism (logical positivism) based on the view that meaningful statements of discourse were either formal propositions of logic and mathematics or factual propositions of the natural sciences. All other statements, including those of ethics and aesthetics, are considered metaphysical and therefore meaningless.

Karl Popper's *Logic of Scientific Discovery* (1934) rejected verifiability as a criterion of meaning for empirical propositions on the grounds that no finite number of positive instances could ever verify it. On the other hand, just a single false instance can falsify a proposition, and Popper therefore proposed falsifiability in place of verifiability as a criterion of meaning. According to Popper, the concept of verifiability fails the test of being empirically verifiable and is therefore itself a metaphysical notion. Rejection of verifiability is closely related to rejection of induction as a tool for verification and is particularly logical for those who reject induction as a mechanism of reasoning.

Kuhn's *The Structure of Scientific Revolutions* (1961) is concerned with the processes by which scientific theories evolve. He introduces the notion of a paradigm as an outstanding scientific achievement that serves as a model for research by a community of researchers, characterizes normal research in a mature scientific discipline in terms of stability of its paradigm, and examines transitions from one paradigm to another, such as the transition from Ptolemaic to Copernican astronomy or from Newtonian to relativistic physics. Kuhn is concerned primarily with the sociology of behavior of scientists during the transition from one paradigm to another. He suggests that scientists are as prone as any other group to adopt self-interest as a basis for action even when it conflicts with rationality.

Kuhn's requirement that a paradigm be an outstanding achievement seems unnecessary. It is sufficient that a paradigm be a set of principles that determine a pattern of research or beliefs and of criteria for the evaluation of research and scholarship, even when there is no outstanding achievement to serve as an example. The notion of paradigms of philosophical and scientific investigation is useful in understanding the relation among ideas and theories. The philosophical positions of realism, idealism, rationalism, and empiricism constitute patterns of thought and belief that qualify as paradigms.

## 4. Computational Knowledge

A computer is a cognitive mechanism, information is a form of knowledge, and computation is a form of thinking:

computer = brain
information (programs + data) = knowledge
computation (information processing) = thinking

These analogies have been criticized as too anthropomorphic, but are justified because computability was proved by Church and Turing to be a remarkably robust concept. Models that realize computation in very different ways, as well as all commercial computers, turn out to have precisely the same computational power. The power of effective procedures computable by Turing machines was found to be the same as that of mathematical systems like the recursive functions or the lambda calculus. We can talk about the computational power of abstract (Platonic) computers without worrying about differences among specific computers. The concept of computability has a Platonic reality comparable in its robustness to mathematical or physical reality. "Computation" becomes a very precise concept whose expressive power and limitations can be analyzed.

Once we accept the idea that information is a form of knowledge, it is natural to ask how it is related to physical, mathematical, and mental knowledge. We ask questions like "Can computers model the real world?" and "Can the real world model computers?" The answer to the second question is obviously yes, but the related question "Do computers occur in nature?" cannot be answered definitively until we understand the extraordinary reproductive ability of DNA. The question "Can mental phenomena be reduced to computational phenomena?" rephrased as "Can machines think?" has been a topic of intense philosophical debate.

The question "Can machines think?" has little direct relevance to the construction of intelligent programs for specific tasks. It does, however, raise important issues concerning the nature and limitations of mechanical processes and of thinking. The debate about whether intelligence can be characterized by behavior and whether the inner quality of thinking and consciousness can be captured by mechanical processes has contributed to our understanding of the nature of knowledge.

### 4.1. The Turing Test

Alan Turing made two major contributions to our understanding of the nature of computation. He defined a precise notion of computability in terms of a simple computational device now known as a Turing machine, and he developed a precise notion of what it might mean for computers to think in terms of a thought experiment now called the Turing test.

In his 1937 paper [T1], Turing introduced Turing machines as a prototypical computing device, demonstrated that their computational power was robust, and proved that the halting problem for Turing machines was undecidable. Turing was uniquely qualified to examine whether computability was sufficiently powerful to capture thinking [T2], but the notion of thinking was not as precisely defined as computability. To avoid metaphysical entanglements, Turing defined thinking behaviorally in terms of a thought experiment (the Turing test) for testing whether the responses of a computer to a wide range of questions were indistinguishable from those of humans.

The Turing test views the mind as a black box or data abstraction whose interface is specified by its stimulus-response behavior and whose internal structure is irrelevant. The problem of modeling the mind becomes that of designing an abstract data type sufficiently powerful to realize the stimulus-response behavior of thinking. Turing gives the following sample dialog between a human questioner and a computer:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one, I never could write poetry.
Q: Add 34957 and 70764.
A: (Pauses about 30 seconds and then gives the wrong answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have a K at my K1 and no other pieces. You have only a K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds): R-R8 mate.

The Turing test avoids questions of tone of voice, facial expression, etc. by restricting communication to an impersonal interface. The computer is allowed to procrastinate, give wrong answers, and lie in its attempt to persuade the questioner that it is human, and is deemed to have passed the test if the questioner believes it is human.

Turing believed computers would in time be able to respond indistinguishably from humans to such questions and concluded that computers were in principle capable of thinking. Turing's view is regarded as simplistic by his critics, but his formulation has nevertheless defined the framework for debate during the forty years since its publication. His position has been questioned on two grounds:

(1) Computability is not sufficiently powerful to capture the behavior of thinking. Computers will never be able to completely simulate human intelligent behavior (they will not be able to pass a sufficiently demanding version of the Turing test).

(2) There is more to thinking than merely behavior. Thinking has intentional properties of self-awareness and consciousness as well as extensional properties of behavior. The Turing test is an insufficient criterion for determining whether machines can think.

In order to handle these two kinds of objections to Turing's position, we break down the question "can machines think?" into two separate questions:

1. Can machines simulate the behavior associated with thinking?
   Can machines simulate thinking extensionally?
2. Can machines that simulate thinking be said to think?
   Do machines capture the intentional properties of thinking when they compute?

The first question appears to have an empirical flavor, while the second is definitely metaphysical. But closer examination suggests that the Turing test doesn't have an empirical decision procedure, since precise criteria for passing the test are not defined. Moreover, even a Turing test with precise criteria could at best affirm the proposition that computers can think but not disprove it. The failure of a given computer to pass the Turing test can always be blamed on the inadequacy of particular machines rather than on inherent impossibility. Thus both questions are metaphysical; neither can be empirically verified. The debate concerning the question "can machines think" is therefore essentially metaphysical.

Turing answered yes to both questions. Skeptics who believe that machines cannot think can be divided into two classes according to which of these questions they deny.

*extensional skeptics* who question the ability of machines to simulate the behavior of thinking (they deny the first question).

*intentional skeptics* who question the ability of machines to realize the inner qualities of processes of thinking, understanding, and consciousness (they deny the second question).

The physicist Roger Penrose and the philosopher John Searle are respectively the spokesmen of the positions of extensional and intentional skepticism. The extensional question logically precedes the intentional one, since there is no point in discussing intentional properties of mechanical processes unless we are convinced of their extensional power. We consider intentional skepticism first because it historically preceded extensional skepticism.

## 4.2. Intentional Skepticism (Searle)

Searle [Se] believes that computers are incapable of thinking because symbol processing is an inherently mindless activity. Computers cannot realize the intentional qualities of consciousness, self-awareness, or cognition because no amount of higher-level organization can possibly convert mindless symbol processing into such higher-level functions. To dramatize the mindlessness of such symbol transformation, Searle proposed the "Chinese room" thought experiment.

Suppose a person with no knowledge of Chinese (locked in a "Chinese room") is able to answer a broad range of questions by mechanically applying rules that transform Chinese symbol strings representing questions about a Chinese story into Chinese answers. Searle argues that producing "intelligent" answers by this method does not imply understanding of the Chinese language or the Chinese story, and asserts that a Chinese speaker would produce answers through a process of understanding quite different from the occupants of Searle's Chinese room.

According to Searle, computers perform symbol transformations without understanding the symbols they transform, just like the occupants of the Chinese room. Answering questions by syntactic symbol transformation does not constitute semantic understanding, no matter how accurate (or human) the answers. Computers cannot think because mere symbol processing does not constitute understanding or cognition; world-champion chess-playing programs are mere symbol transformation devices that do not *really* understand chess.

However, proving that computers cannot have cognitive states is difficult, requiring a precise definition of "cognitive states" and a proof that no possible computation can realize cognitive understanding. Showing that a particular computation does not simulate thinking is insufficient. Thus Searle's attempt to demonstrate that *all* computations are mindless by appealing to the fact that symbol transformations are mindless may at first appear convincing, but has fundamental flaws.

### 4.2.1. The Society of Mind

Mindless low-level agents, whether silicon transistors or humans locked in a Chinese room, can be organized to realize goal-oriented, purposive behavior. Research on neural networks has the goal of realizing purposive behavior in terms of the mindless firing of neurons. The neurons that cause human intelligence are considered to be mindless agents whose power lies in the rules by which they are organized. Soldiers in an army are likewise not aware of the high-level goals that result from their low-level actions. World-class chess playing programs are nothing but mindless symbol transformation rules organized to realize high-level chess strategies.

Intelligence, whether in the human mind or in a computer, does not reside in individual agents but in the way that agents are organized. In computers intelligence resides in the program that organizes individual symbol-processing steps to perform a specific task. In an army intelligence resides in the strategic plan that causes soldiers to cooperatively win a battle. In the human mind intelligence resides in the as yet unknown mechanisms that cause the brain to behave intelligently.

Any abtract process that hides "irrelevant" internal structure while revealing "relevant" interface behavior is in some measure intentional. The Chinese room is such an abstraction, as are procedure and data abstractions in programming languages. The question whether information hiding in silicon modules differs in some essential way from information hiding in biological organisms is a legitimate one. But the argument that symbol transformations in silicon modules are mindless does not address this question because neuron firings are as mindless as symbol transformations. Searle agrees that the difference is not simply a matter of what modules are made of (he accepts in principle that silicon Martians could be intelligent). Thus the difference would have to be in qualitative properties of consciousness that can be realized by neuron structures but not by silicon structures in computers. When put in this way, Penrose's contention that the difference, if there is one, must be extensional appears reasonable.

Minsky's *Society of Mind* [Mi] uses the metaphor of a society as the mechanism for organizing mindless agents into intelligent systems. He formulates the following questions about agents of the mind:

Function: How do agents work?
Embodiment: What are they made of?
Interaction: How do they communicate?
Origins: Where do the first agents come from?
Heredity: Are we all born withthe same agents?
Learning: How do we make new agents and change old ones?
Character: What are the most important kinds of agents?
Authority: What happens when agents disagree?
Intention: How could such networks want or wish?
Competence: How can groups of agents do what individual agents cannot do?
Selfness: What gives them unity of personality?
Meaning: How could they understand anything?
Sensibility: How could they have feelings and emotions?
Awareness: How could they be conscious or self-aware?

Organisms become self-aware by having a model that allows them to examine and control their cognitive states. Such organisms are said to be reflective; they can reflect on their own structure and behavior. Reflective computational systems have a computational model of their own behavior which they can examine and manipulate to modify their behavior. Reflective computational systems have been extensively studied, providing insight into the nature of computational introspection that may be used in modeling human introspection.

Darwinian natural selection provides a mindless natural mechanism for the emergence of purposive, intentional organisms. It imposes an evolutionary direction on random (mindless) variation of organisms by selecting those best fitted for survival. The selection process has the effect of imposing a purpose and direction on evolution. Natural selection provides an explanation of the apparently intentional behavior of living organisms and a mindless mechanism for the evolution of such organisms.

Silicon modules for performing purposive tasks are created by a design process that requires repeated selection among design alternatives. Repeated selection steps during a single lifetime has a similar, though more rapid, effect to a single selection step in each generation. Non-natural selection of design alternatives during design, implementation, and adaptation (enhancement) of silicon modules fosters purposive, goal-oriented evolution similar to that of Darwinian natural selection.

Searle's conclusion that occupants of the Chinese room do not understand their task is correct. But his inference that computers cannot therefore exhibit intentional behavior is wrong, since a defining feature of intentionality is precisely that the low-level agents realizing higher level intentionality do not understand the higher-level task. Searle may possibly be right about mental processes being in some sense essentially different from computational processes, but his Chinese room argument does nothing to support his claim that computers cannot realize intentional behavior.

### 4.3. Extensional Skepticism (Penrose)

Penrose's thesis is that the laws of physics are incomplete, that complete physical laws will be noncomputable but deterministic, and that minds are subject to noncomputable physical laws [Pe]. The evidence for noncomputability includes the nonlocality (action at a distance) of gravity and quantum phenomena and the unpredictability of mental phenomena like insight and creativity.

### 4.3.1. Computational, Mathematical, Physical, and Mental Knowledge

Penrose believes that the philosophical issues surrounding the concept of "mind" cannot be understood without an appreciation of the physical and mathematical laws governing the behavior of the universe. *The Emperor's New Mind* therefore examines the nature of computational, mathematical, physical, and computational knowledge in some detail.

(1) Computational knowledge:
Because algorithmic and formal techniques have limited power, computational phenomena cannot model the complete range of mathematical and physical phenomena. Both the undecidability of the halting problem for Turing machines and Godel's incompleteness theorem are proved by similar diagonalization techniques. Formal systems of mathematics are subject to the same kinds of (diagonalization) limitations as models of computation, and neither are powerful enough to model physical or mental phenomena. Deductive systems and computational systems are qualitatively similar, representing similar kinds of knowledge. Formal and computational knowledge is qualitatively different from mathematical, physical, and mental knowledge.

(2) Mathematical knowledge:
New mathematical objects are generally discovered rather than invented; discovery generally requires nonalgorithmic insight and creativity. The Platonic world of mathematics is essentially richer than the algorithmic computational world. The Mandelbrot set, which exhibits intricate geometric patterns generated by mathematically simple rules, is an example of an independently existing mathematical object. Mathematical structures like the Mandelbrot set are part of an eternal Platonic reality that can be discovered by mathematicians in much the same way that the planet Neptune or Newton's laws are discovered.

(3) Physical knowledge:
Physical laws are phenomenally accurate and their mathematical form suggests that reality, or at least our perception of reality, is in some degree mathematical. But physical laws are accurate only over a limited scale of phenomena, gravity and quantum phenomena exhibit puzzling forms of nonlocality, and physical laws of quantum theory are incomplete. The incompleteness of current physical laws can be remedied only by introducing an element of noncomputability. The apparent nondeterminism in measuring the position and momentum of particles can be eliminated by a more complete (noncomputable) law for particle motion.

(4) Mental knowledge:
Mental phenomena are subject to the laws of physics but are so complex that physical explanations are still quite primitive. One of the goals of neural science is to model cognition in terms of the behavior of neurons. Mental models must explain the firing of neurons

and also the brain plasticity that is the basis of memory. Quantum considerations appear to affect the action of the brain (we can perceive a single photon), so that the incompleteness of physical laws of quantum theory affects our ability to model the brain completely.

Penrose's goal of developing a unified theory of physical and mental phenomena is incredibly ambitious, more ambitious than attempts to develop unified physical theories. His suggestion that physical laws are noncomputable but deterministic is provocative but not very helpful in practice because there are no practical examples of useful noncomputable theories. His critics [BBS] argue that physical laws are in fact computable and that, even if they were not, our ability to work with noncomputable laws is so limited that we are better off with approximate physical laws that are computable.

Does Godel's incompleteness result imply that mathematical insight is nonalgorithmic? Penrose carefully demonstrates that there are undecidable true propositions of any formal system that can immediately be seen to be true by insight. However, it does not necessarily follow that there is no algorithm for formalizing such insight in a more powerful reflective metasystem. Martin Davis argues in [BBS] that such insight is indeed formalizable, making use of Godel's completeness result (that the predicate calculus allows all semantically true propositions to be proved). The debate concerning the formalizability of insight for undecidable but true Godel propositions is inconclusive, depending on precise definitions of the notion of formalizability. However, there are other, more fundamental grounds on which Penrose's hypothesis of noncomputable physical laws may be questioned.

Penrose's evidence for the incompleteness of physical laws is convincing, but his faith that they can be completed and his very specific prescription for completing them are in doubt. We consider three kinds of objections to the hypothesis that a new paradigm of physical explanation based on noncomputable but nondeterministic physical laws should be the basis for "complete" physical and mental models:

(1) Even if a new paradigm of physical explanation is needed, Penrose's hypothesis of computable nondeterministic laws is not the only hypothesis compatible with the evidence.

(2) The belief that incomplete physical laws can be completed tacitly assumes that completeness is a meaningful concept for physical models and that complete physical models exist. Completeness for mathematical models rests on a correspondence between truth and derivablility for which there is no physical analogue.

(3) Noncomputable models, if they exist, must be usable as a basis for discovering, predicting, and controlling physical and mental phenomena. The existence of complete models is insufficient without the additional requirement that people and/or computers can use them effectively.

### 4.3.2. Alternatives to Noncomputability

Penrose's evidence for incompleteness does not in itself imply any particular way of completing physical laws, or even that completeness is possible. His hypothesis of noncomputability appears compatible with the evidence, but he does not claim that it is the only compatible hypothesis. We propose two other compatible hypotheses: chaos (nonlinear laws) and infinitely scalable (Mandelbrot) universes. Chaos is a weaker hypothesis, since nonlinear laws are in principle, though not in practice, computable. An infinitely scalable universe implies a stronger hypothesis: that not even noncomputable laws are globally valid at all levels of scale.

Nonlinear differential equations are extremely sensitive to changes in initial conditions, so that the future is in practice not computable from the past. Extreme sensitivity to initial conditions gives rise to *chaos*, a phenomenon that turns out to be more predictable than its name suggests. Chaos was first noticed by Lorenz in the late 1950s in weather data, but is present even in phenomena that appear very regular, like the motion of pendulums. The study of chaos has made considerable progress since the 1960s, and global stability (attractors) independent of the local unpredictability of the future on the past gives models of chaos an interesting structure.

Quantum theory has been developed largely for linear systems, the linear principle of superposition of quantum states has played a crucial role. A nonlinear quantum theory that abandons superposition could conceivably provide the missing ingredient of physical laws required to handle nonlocality. It is also plausible that creativity and insight are the result of extreme physical sensitivity to initial conditions rather than noncomputability. Compared with the apparent intractability of noncomputable models, chaos has the pragmatic merit of already being a substantive research area with well-defined experimental and mathematical techniques that has uncovered unexpected physical regularities (created order out of chaos).

Penrose's noncomputability hypothesis could be wrong because physical reality, although locally amenable to mathematical approximation, is globally too complex to be modeled even by noncomputable laws. Models in molecular biology, genetics, and neural science seem to involve a never-ending process of approximation that could turn out to be inherently approximate theoretically as well as practically. Newtonian mechanics is approximately valid for galaxies, the solar system, terrestrial bodies, and atomic phenomena, but breaks down for velocities approaching the speed of light and for very large masses. Cosmology breaks down for spatial singularities like black holes and for temporal singularities like the big bang origin of the universe. Computational models are limited not only by nonalgorithmic phenomena but also by the limited accuracy of initial conditions and the phenomenon of chaos. Any theory inevitably encounters limitations in the scale of phenomena to which it applies.

To illustrate problems of scalability, consider the hypothesis that the universe is infinitely scalable like the Mandelbrot set: that is, each atom contains a universe of atoms that in turn contains universes of atoms ad infinitum. This hypothesis appears compatible with physical laws, and would be yet another blow to our view that we occupy a special place in the universe. There is no reason to suppose that the level of scale at which humans operate is special compared with other levels of scale at the level of electrons or galaxies other than the fact that our observation processes operate at this level. The property that every fragment of the Mandelbrot set, no matter how small, contains a structural replica of the complete set is suggestive of genetic codes that replicate the complete adult organism in each cell of the embryo. Such infinite scalability, though not directly suggested by Penrose, accords with his Platonic preference for a physical reality that emulates mathematical reality.

Scalable universes (call them Mandelbrot universes) have no "fundamental particles" in any absolute sense. But limitations on observations at one level of scale about phenomena at other levels of scale might cause universes at one level to behave like fundamental particles and exhibit quantum effects when viewed from a higher level. Current physical theories are in principle compatible with this idea. The scale factor between successive universes could possibly be related to to the total number of particles in the universe ($10^{80}$), or perhaps, through quantum effects, to Planck's constant ($10^{35}$).

If the universe were in fact infinitely scalable like a Mandelbrot set, then fundamental particles would not exist and the existence of physical laws that apply uniformly at all levels of scale would be questionable. The best that we could get would be an approximation that works for a finite local region of an infinitely scalable universe. The hypothesis of infinitely scalable universes has not yet been seriously explored by scientists as a model of the universe, but could lead to the discovery of interesting scale invariants that relate phenomena at different level of scale.

### 4.3.3. The Search for Completeness

Are "complete" physical laws possible? Is the concept of completeness meaningful for physical (as opposed to mathematical) models? Laws for the motion of particles tell us little about their other physical properties like texture and appearance and even less about what they are "really" like. We can learn something about the inherent limitations of modeling physical reality by considering formal systems as models of mathematical reality. However, the gap between physical models and the reality they describe is greater than that between mathematical models and their associated mathematical reality. In particular the mathematical notion of completeness, which is based on a correspondence between truth and derivability in formal systems, has no analogue in physical theories.

Formulae of formal systems have a syntactic notion of being true in all interpretations (validity) that can be defined independently of the notion of provability. A formal system is said to be complete if all valid formulae are provable from the axioms by rules of inference (if truth implies derivability). Physical theories have no corresponding notion of truth and validity and therefore no notion that corresponds to mathematical completeness. Possibly a nonmathematical notion of completeness is definable for physical theories in terms of an independently defined notion of adequacy for a particular purpose. However, any such notion would be a relative rather than absolute notion of completeness.

Formal systems have many possible models (possible worlds) while physical theories are about a single real world. However, the assumption that there is a single reality cannot be inferred from the physical laws themselves. Physical laws are constraints on the behavior of natural phenomena that determine a set of possible worlds rather than a unique physical reality. Newton's laws describe some specific properties of the physical world rather than its complete behavior. The laws of electromagnetism, thermodynamics, optics, and accoustics describe other complementary properties that interact with Newton's laws of mechanics in a compatible way. Each branch of physics determines a qualitatively different set of constraints on behavior that provides a different perspective on the physical world.

The collection of constraints determined by subdisciplines of physics are compatible with each other and with many different realities corresponding to particles with different shape, texture, color, constitution, etc. Observation of the texture and color of atoms and molecules is probably impossible in principle for the same kinds of reasons that the position and momentum of particles cannot be precisely determined.

The unique real world whose existence we assume can never be completely described by physical theories. The best we can do is to describe constraints that determine a set of possible worlds. The view of physical theories as constraints allows the observed nondeterminacy of quantum theory to be attributed to constraints inherent in the process of observation rather than to reality. Nondeterminacy arises from our inability to determine which possible world we inhabit rather than from the inherent nondeterminacy of our given possible world. Einstein's belief that God does not play dice with the universe thus becomes compatible with quantum theory.

Sometimes a formally defined theory that is thought to be physical, like Euclidean geometry, is enriched by conceiving it as a mathematical theory for which new interpretations can be discovered by modifying its axioms. Most physical theories are not formally defined in the sense of Euclidean geometry. But our discussion shows that nonformal physical theories admit many interpretations. Questions of completeness, uniqueness, determinacy, and computability are interdependent and surprisingly fuzzy considering the maturity of the discipline of physics.

*The Emperor's New Mind* reexamines physical, mathematical, and mental knowledge from the perspective of the late 1980s, and is unusual in considering the impact of computers on the disciplines of physics, mathematics, and cognitive science. It serves, among other things, to elevate computer science into a first-class discipline that may in time achieve the the intellectual status of physics or mathematics. In spite of questions concerning noncomputability, Penrose's book provides a baseline for scientific epistemology in the computer age.

## 5. References

Primary references cited in the body of the paper are not included here. They may be found in the secondary references [Ru] and [Ay].


[Ru] Bertrand Russell, *History of Western Philosophy*, Simon and Schuster

[Ay] A. J. Ayer, *Philosophy of the 20th Century*, Random House, 1982.

[Ku] Thomas Kuhn, *The Structure of Scientific Revolutions*, second edition, University of Chicago Press, 1970.

[T1] A. M. Turing, On Computable Numbers with an Application to the Entscheidungsproblem, *Proc London Math Society*, 1937

[T2] A. M. Turing, Computing Machinery and Intelligence, *Mind*, 1950

[Se] John R. Searle, Minds, Brains, and Computers, *Journal of Behavioral and Brain Sciences*, 1980, Reprinted in *Mind Design*, Ed John Haugeland, Bradford Books, 1981

[Cr] K. Craik, *The Nature of Explanation*, Cambridge University Press, 1943.

[Mi] Marvin Minsky, *The Society of Mind*, Simon and Schuster, 1986.

[Pe] Roger Penrose, *The Emperor's New Mind*, Oxford University Press, 1989.

[BBS] Debate between Penrose and his commentators, *Journal of Behavioral and Brain Sciences*, December 1990, Cmabridge University Press, 1990.

[Gl] James Gleick, *Chaos, Making a New Science*, Penguin Books, 1988.