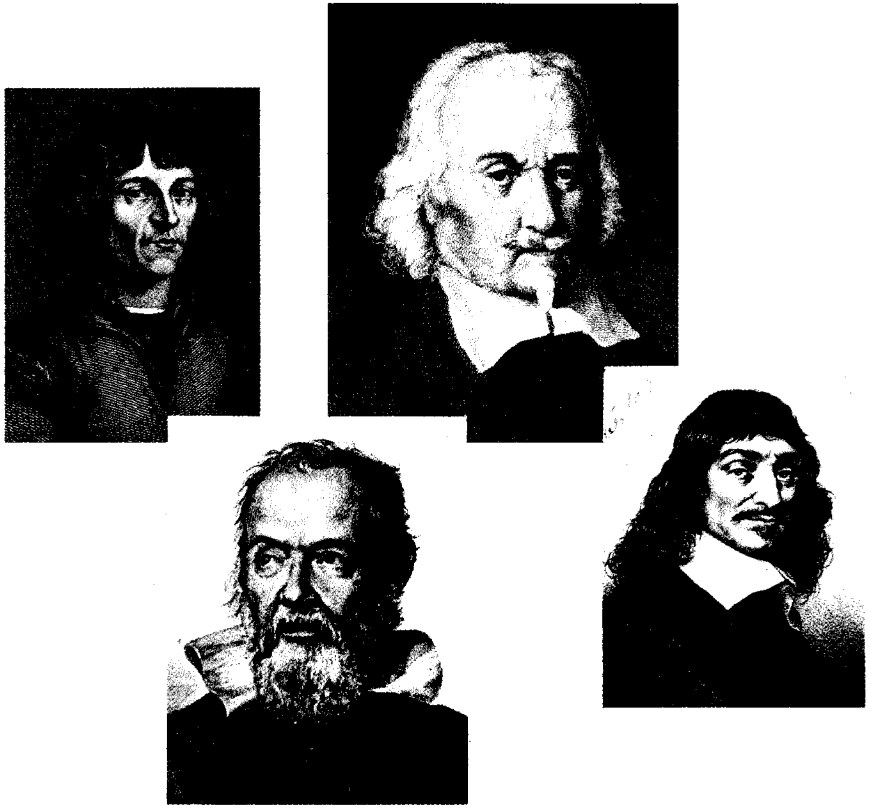


Clockwise from upper left, this page: Nicolaus Copernicus, Thomas Hobbes, René Descartes, Galileo Galilei; facing page: David Hume, Alan Turing, John von Neumann, John McCarthy, Allan Newell, Charles Babbage (center).

Introduction

Galileo instructing Milton

1 The Saga of the Modern Mind

Copernicus and the End of the Middle Ages

Our commonsense concept of "the mind" is surprisingly recent. It arose during the seventeenth century, along with modern science and modern mathematics—which is no mere coincidence. These historical roots are worth a look, not only because they're fascinating, but also because they give a richer perspective on cognitive science (which is, after all, just the latest theory in the series). Fortunately, the story itself is rather more suspenseful and intriguing than one might expect. One scholarly confession first, however: in bringing out the main contours, I have streamlined historical fact, omitting miscellaneous important people and subtle distinctions. For instance, I mention "the medievals" as if they were all alike, which is a scandal; and many of the ideas I associate with, say, Galileo or Descartes were really current at the time and were being discussed by many writers. Still, the larger drama is about right, and that's what matters.

Though the plot didn't really get rolling till the time of Hobbes and Descartes, it began building steam a century or two earlier, as the Middle Ages gave way to the Renaissance. The medieval world view was largely a Christian adaptation of ancient Greek philosophy and science, especially the works of Plato and Aristotle. The principal modification was putting God at the foundation, as creator and cause of everything (else) that exists. Prior to the creation, however, the Creator did need *ideas* of all that there would be (His plans, you might say); and these ideas played an important role for philosophers. In the first place, obviously, ordinary worldly objects were just more or less corrupt materializations of God's original, perfect ideas. (The corruptions, of course, must also have been planned by God; but medieval rationalizations for that got rather sticky and needn't concern us.)

The human intellect or soul had ideas too, something like God's; but their status and relation to objects was more problematic. One charming story (loosely Platonic in inspiration) had the thought/thing relation as really the base of a triangle, with God's intellect at the apex. Human ideas were *true* insofar as they were more or less accurate copies of God's ideas; mundane objects, in turn, were also true, though in another sense, insofar as their construction more or less conformed to those same divine ideas, now seen as designer's blueprints. Thus our thoughts could be "true of" things only via a detour through the original plans that they each imperfectly matched, in their respective ways.

A more common (and more Aristotelian) approach skipped the detour through the apex and postulated a direct relation between thought and thing at the base. But what relation could that be? By virtue of what could an idea in a mind be *of* or *about* some particular worldly object(s)? The most appealing and also standard answer was that ideas are like pictures or images: they *resemble* the objects they stand for, and they stand for those objects because they resemble them. Thus a thought could be true of a thing by "conforming" with it directly—that is, by sharing the same *form*. In a sense, the mind was just like the world (i.e., in form), only realized in a different substance (mental instead of material). This account of how thoughts could relate to things had the double advantage of intuitive plausibility and the lack of serious competition. As we shall see, however, the emergence of modern science slowly sabotaged the standard resemblance theory and eventually forced a quite different view of mental contents.

Medieval cosmology—the theory of the universe—was also basically Aristotelian. Our Earth was at the very center, surrounded in various rotating spheres by the visible heavens, which were surrounded ultimately by God's heaven, motionless and invisible.[1] The sensible world was composed of five elements: earth, water, air, fire, and the so-called quintessence (fifth element). Each of these had its "natural place" toward which it naturally tended to travel, if it were ever removed. The heavens were composed entirely of quintessence, and since this is where the quintessence

belonged, the heavens never changed (the spheres just rotated ceaselessly in place). The other four elements, however, were all jumbled up in the lowest sphere, below the moon; and hence this sphere was in constant turmoil, with things always subject to change and decay. For example, wood was flammable because the fire and air in it "tended" to go up and the earth "tended" to go down, if they ever got the chance. This same fire and air component also explained why wood floated on water while stones and ashes (which were more concentrated earth) did not float—water's natural place being above earth, but below fire and air.

All told, it was a pretty picture; but by the late middle ages there was trouble brewing. For sundry practical reasons, astronomy was the most advanced empirical science. The calendar had to be right, for determining religious holidays; and navigators, venturing ever farther from known waters, needed more accurate astronomical tables. Unfortunately, as more careful observations accumulated, it became progressively harder and more complicated to square them with the accepted geocentric (Earth-centered) theory. The situation was so serious and exasperating, especially with regard to predicting the positions of the planets, that by the thirteenth century Spain's King Alfonso X could exclaim: "If God had consulted me when creating the universe, He would have received good advice!"[2]

On the Revolution of the Spheres, published by the Polish astronomer Nicolaus Copernicus (1473–1543) in the last year of his life, turned the medieval world literally upside down. Its heliocentric (Sun-centered) theory of the universe was surely the most disorienting scientific innovation of all time—though Copernicus's successors were oddly slow in appreciating all its implications. The basic ideas were that

1. the daily motions of the heavens are just an illusion, brought about by rotation of the Earth on an internal axis; and
2. the annual circuit of the Sun through the Zodiac, as well as some of the stranger wanderings of the planets, are equally illusory, due to the Earth slowly orbiting around the Sun.

The Earth itself was reduced to the status of another planet, situated between Venus and Mars in its distance from the Sun at the center.

Not only was this proposal destined to transform astronomy, but it also (eventually) threw the rest of accepted science into disarray. For now the natural places, which were the basis of the account of all mundane movement and decay, were totally dislocated. Which directions were up and down? Why would the four elements of the universe (besides the quintessence) have their natural "tendencies" tied to some peculiar moving point inside the Earth? Where, indeed, were Heaven and Hell? The questions for the theory of motion were equally serious. Why didn't loose objects whirl off the spinning Earth and then get left behind in space as we sailed on around the Sun? Or why, at least, didn't a falling stone come down slightly "behind" (west of) its original location as the Earth rotated out from under it? These were extremely hard problems, and it was many years before the laws of inertia, force, and gravitation were all worked out to solve them.

The modern *mind* was invented ('invented' is the right word) in this same scientific ferment; and its first impulse came from the Copernican distinction between appearance and reality. It is a commonplace, of course, that things are not always what they seem; and, even in ancient times, Plato had elevated this into a philosophical principle: *never* are things *really* what they seem. The ordinary objects of perception, he maintained, are only pale imitations of true reality, like mere shadows cast on a wall by a dancing fire. (For Plato, true reality was the realm of perfect eternal "forms"—later appropriated as God's intellect, full of divine ideas.)

Copernicus, however, needed a more drastic distinction. The illusion of the Sun rising in the east is no pale shadow or imperfect mental picture of the Earth's rotation—it's totally different. The same can be said for the Sun's motion through the Zodiac, and even more so for the planets' complicated wanderings back and forth among the stars. According to the new theory, astronomical appearances weren't even similar to reality. That opened breathtaking new possibilities for the sorts of things science could conceivably discover; and at the same time, by undermining resemblance, it drove a crucial wedge between the mind and the world—a wedge that ultimately transformed our whole understanding of thinking and ourselves.

Galileo and the New Science

One final stop, before Hobbes draws the fateful computational conclusion, is the great Italian physicist Galileo Galilei (1564–1642), who enters our story in several ways. He is perhaps most famous for introducing the telescope into astronomy and thereby discovering the moons of Jupiter, the mountains on Earth's moon, the changing phases of Venus, and so on. These were all important factors in the ultimate triumph of Copernicanism. (For his trouble, Galileo was tried by the Inquisition and sentenced to house arrest for the last eleven years of his life.) But in the history of the modern mind (and also in the development of modern physics), what is most important is his application of mathematics to the problem of motion.

Galileo was convinced that the only way to understand physical nature is in terms of mathematical relationships among quantitative variables. He himself expresses the idea both colorfully and lucidly:

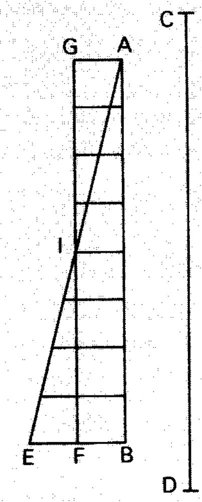
> Philosophy is written in that great book, the universe, which is always open, right before our eyes. But one cannot understand this book without first learning to understand the language and to know the characters in which it is written. It is written in the language of mathematics, and the characters are triangles, circles, and other figures. Without these, one cannot understand a single word of it, and just wanders in a dark labyrinth.[3]

Mathematics, for Galileo, was essentially geometry (and arithmetic). This is evident, for example, if you look at his proof of Theorem 1 (see box 1), where he relies on geometrical concepts and relationships to say things that we would say with algebraic equations.

What matters historically, though, is not just *that* Galileo used geometry, but *how* he used it. Traditionally, geometry was the study of figures and relations in space. But Galileo conceived of it more abstractly. So, for example, lines in his diagrams wouldn't

Box 1
Galileo's Theorem 1

The time it takes a uniformly accelerated body to cover any given distance, starting from rest, equals the time it would take the same body to cover the same distance at a constant rate of speed equal to half the maximum speed finally achieved by the accelerated body.[4]



Let line AB represent the time in which some body uniformly accelerated from rest covers the distance represented by line CD. Let line BE (intersecting AB at any angle) represent the maximum speed achieved by that body at the end. Then all the line segments parallel to BE and connecting intervening points on AB to line AE represent the gradually increasing speeds during time interval AB. Let F be the midpoint of BE, and draw lines FG and AG, parallel to BA and BF, respectively, thereby constructing parallelogram AGFB, which is equal in area to triangle AEB (with side GF bisecting AE at I); and suppose the parallels in triangle AEB are extended straight out to line IG. Then we have the aggregate of all the parallels contained in the quadrilateral equal to the aggregate of those comprised in triangle AEB; for the ones in triangle IEF exactly match those in triangle GIA, and trapezoid AIFB is common to both cases.

Now, the instants in time AB correspond one for one to the points in line AB; and, taking the parallels from all those points, the portions enclosed by triangle AEB represent the increasing speeds, while, likewise, the portions contained in the parallelogram represent nonincreasing speeds, of which there are just as many, even though they are all equal. Hence it is apparent that exactly as many momenta of speed are consumed in the accelerated motion (given by the increasing parallels in triangle AEB) as in the constant motion (the parallels in parallelogram GB); for surely the deficit of momenta at the beginning of the accelerated motion (as represented by the parallels in triangle AGI) is made up by the momenta represented by the parallels in triangle IEF. Thus two bodies will clearly travel through the same distance in the same time, when the motion of one is a uniform acceleration from rest, while that of the other is constant, with just half the momentum of the accelerated motion at its maximum speed.
Q. E. D.

always represent lines or even distances in space, but might just as well represent times, speeds, or any other interesting physical variable. Theorem 1 is a case in point: though it is about bodies traveling a given distance, no line in the diagram represents either the paths or the distance. (As if to emphasize this fact, Galileo draws CD off to the side and then never mentions it again.) Instead, the lines actually represent times and speeds. Thus point A is the "starting point," but only in time, not in space; points further down from A on the line AB represent later instants, not subsequent positions. Lines GF and AE don't represent anything; but, in effect, they determine the speeds as functions of time. That is, all the equal line segments drawn over to GF from points on AB represent equal speeds at all those times, while the gradually lengthening segments from AB to AE represent gradually increasing speeds (i.e., uniform acceleration). The distances traveled are then represented by (of all things!) the respective "aggregate" *areas*;[5] hence the proof reduces to the trivial theorem that triangle AEB encloses the same area as parallelogram AGFB.

Obviously Galileo's main contribution is not the proof itself but the abstract representation in which such a proof could be given. Discovering and validating this strange way of representing instantaneous velocity, uniform acceleration, total distance, and so on cost Galileo many years of struggle. It looks so simple or even clumsy now; but it is one of the great achievements of the human intellect. What made it really significant, though, was not any particular result but rather the fact that now all the familiar techniques of geometry could be used to establish all kinds of results. Euclid's whole deductive system could be abstracted away from geometric shapes and applied instead to motions. For example, given the empirical hypothesis that falling bodies accelerate uniformly, Galileo was able to *prove* his classic "times-squared" law;[6] and, assuming the motion of a projectile to be a combination of uniform horizontal motion and vertical free-fall (a stunning insight in itself), he could show that the actual path would be a parabola.

Like Copernicus before him, Galileo didn't "philosophize" much about the mind or soul; hence, though his dramatic new uses of geometry had important consequences for the theory of mental representation, it took Hobbes and Descartes to bring them out. Galileo did, however, draw one famous and influential conclusion about "metaphysics"—that is, about what's really real:

> I believe that for external bodies to excite in us tastes, odors, and sounds, nothing is required in those bodies themselves except size, shape, and a lot of slow or fast motions [namely, of countless "tiny particles"]. I think that if ears, tongues, and noses were taken away, then shapes, numbers, and motions would well remain, but not odors, tastes, or sounds. The latter are, I believe, nothing but names, outside of the living animal—just as tickling and titillation are nothing but names, apart from the armpit and the skin around the nose.[7]

In neighboring paragraphs Galileo also included colors and heat in the same category: qualities that aren't really present in external objects as such but arise only in our perceptions. This general idea is as old as the Greek "atomist" Democritus (fifth century B.C.), but Galileo gave it a whole new credibility. For he held that nature herself is "written in mathematical characters" (i.e., shapes, sizes, and motions) and supported that doctrine with totally unprecedented precision and detail, by actually deciphering those characters in his laws and proofs.

Most philosophers since Galileo have accepted some form of this distinction between properties that objects really have in themselves and properties that only appear in them because of the nature of our perceptual apparatus. Perhaps the best known discussion is by the English philosopher John Locke (1632–1704), who called the two groups "primary" and "secondary" qualities, respectively.[8] But the important point, from our perspective, is that the distinction drives in the Copernican wedge between how things seem and how they really are; that is, it further separates thought from the world. As we shall soon see, however, Galileo's methods of mathematical representation were destined to have an even deeper influence on the evolving modern mind.

Hobbes—The Grandfather of AI

"By RATIOCINATION, I mean *computation,*" proclaimed the English philosopher Thomas Hobbes (1588–1679), prophetically launching Artificial Intelligence in the 1650s.[9] This slogan, he explained, conveys two basic ideas. First, thinking is "mental discourse"; that is, thinking consists of *symbolic operations,* just like talking out loud or calculating with pen and paper—except, of course, that it is conducted internally. Hence thoughts are not themselves expressed in spoken or written symbols but rather in special brain tokens, which Hobbes called "phantasms" or thought "parcels." Second, thinking is at its clearest and most rational when it follows methodical rules—like accountants following the exact rules for numerical calculation. In other words, explicit ratiocination is a "mechanical" process, like operating a mental abacus: all these little parcels (which, of course, need not stand only for numbers) are being whipped back and forth exactly according to the rules of reason. Or, in cases where the rules are being ignored or bent, the person is simply confused.

Here is how Hobbes himself elaborated the point in his magnum opus, *Leviathan*:

> When a man reasoneth, he does nothing else but conceive a sum total, from addition of parcels; or conceive a remainder, from subtraction of one sum from another. . . . These operations are not incident to numbers only, but to all manner of things that can be added together, and taken one out of another. For as arithmeticians teach to add and subtract in numbers; so the geometricians teach the same in lines, figures, . . . angles, proportions, times, degrees of swiftness, force, power, and the like; the logicians teach the same in consequences of words; adding together two names to make an affirmation, and two affirmations to make a syllogism; and many syllogisms to make a demonstration.[10]

Hobbes no doubt overstretched the metaphor of "addition," but we can wink at that after all these years. More interesting is his inclusion of time, swiftness, force, etc. in the domain of geometricians; clearly he is thinking of Galileo.

Hobbes greatly admired Galileo, and in 1634 journeyed all the way to Italy just to meet him. More particularly, the great Italian's discovery that physics could be conducted with all the methodical rigor of geometric proof directly inspired Hobbes's conception of intellectual method in general. Thus his momentous suggestion that thinking *as such* is basically computation was clearly an extrapolation from Galileo's idea that the study of motion is basically geometry. And Hobbes's own main philosophical work was (wild as it seems) an attempt to do for politics what Galileo had done for physics. Needless to say, his definitions and proofs were not quite as precise and convincing as Galileo's (or we might live in a better world today); but his overall approach transformed the field—and, in recognition of that, he is often called "the father of modern political science."

Closer to our present topic, Hobbes also eagerly embraced the idea that reality itself is fundamentally "mathematical": ultimately nothing but tiny moving particles. Hence he readily agreed that so-called sensible qualities (colors, odors, tickles, and the like) are not really in objects at all but only in perceivers. At this point, however, Hobbes went his predecessor one better; for he argued, as Galileo had not, that if *all* reality is just particles in motion, then that must include the mind and its contents as well.

> All which qualities, called *sensible*, are [nothing] in the object, but so many several motions of the matter, by which it presseth our organs diversely. Neither in us that are pressed, are they any thing else, but divers motions; for motion produceth nothing but motion.[11]

Thus when I receive blue or tickling sensations from a colored feather, these sensations are really just complex patterns of tiny motions in my sense organs and brain; they are no more blue or tickling in themselves than the feather is in itself. Or rather to call them blue or tickling sensations is simply to classify them among certain repeatable patterns of movement in my body. This makes it evident, among other things, that when Hobbes speaks of putting thought parcels together and taking them apart again, he means it literally—at least to the extent that he means real physical manipulations of tiny physical symbols.

But there is a fundamental difficulty in Hobbes's whole program: to put it starkly, he cannot tell (that is, his theory cannot account for) the difference between minds and books. This is the tip of an enormous iceberg that deserves close attention, for it is profoundly relevant to the eventual plausibility of Artificial Intelligence. The basic question is: How can thought parcels *mean* anything? The analogy with spoken or written symbols is no help here, since the meanings of these are already *derivative* from the meanings of thoughts. That is, the meaningfulness of words depends on the prior meaningfulness of our thinking: if the sound (or sequence of letters) "horse" happens to mean a certain kind of animal, that's only because we (English speakers) mean that by it. Hobbes even agrees

> that the sound of this word *stone* should be the sign of a stone, cannot be understood in any sense but this, that he that hears it collects that he that pronounces it thinks of stone.[12]

Box 2
Causal Determination of Meaning

Suppose I am looking at an apple and having an experience *of* it. If we ask why that experience is of that apple (rather than of another apple or of Calcutta), the answer seems clear: that particular apple is *causing* my experience. Moreover, if I later remember the apple, that too is of that apple by virtue of a causal connection (via the initial perception). Finally, perhaps even my *concept* 'apple' gets its meaning from causal relations to various apples, such as those used to teach me the concept in the first place.

Here then is an account of meaning that is neither derivative nor based on resemblance. Does it solve the mystery? Not until questions like the following get answered:

1. **WHICH CAUSES?** Overindulgence causes splitting headaches, but splitting headaches don't *mean* (represent, stand for) overindulgence. When I see the apple, my experience is also caused by the photons entering my eye, the act of opening my eyes, and the microtexture of the apple's skin. Why do some causes generate meanings while others don't?
2. **WHICH MEANINGS?** As I regard the apple, I can see (notice, think of) an apple, a fruit, a particular gift from a student, lunch, its redness, its distance from me, and so on. These all differ in meaning (content), yet the causal connection between the apple and me seems just the same.
3. **INFORMED MEANING:** Mechanic and client alike hear the engine ping; but it "means" more to the mechanic, both in the experience (what kind of ping) and in the concept (what pinging actually is). Again there is more to meaning than is determined by the cause.
4. **NONCAUSAL OBJECTS:** I can think of the number eleven, and mean it; yet the number eleven itself has never caused anything. Likewise, I can think of the future or a possibility that didn't happen, though neither has had any effects. And what about abstractions: has the *species* raccoon caused anything not caused by individual raccoons?

None of this argues that causal factors couldn't form *part* of a more elaborate approach to original meaning; indeed, much current work is along such lines—but incorporates ideas not available to Hobbes.

Now obviously the meaningfulness of thoughts themselves cannot be explained in the same way; for that would be to say that the meanings of our thoughts derive from the meanings of our thoughts, which is circular. Hence some independent account is required.

I call this the *mystery of original meaning*. For the problem is: Where does meaningfulness *originate*? Some meanings can derive from others, but not all of them can; if the meanings of public symbols (e.g., language) derive from the meanings of internal symbols (e.g., thoughts), then the latter cannot be similarly derivative—they must be the "originals." In other words, Hobbes cannot explain thinking by saying it's *just like* talking or writing, except for being internal. There has to be some further difference that accounts for the fact that thoughts can have original meaning, while word meanings are only derivative. This "further difference" is then clearly the crux of the matter.

The standard resemblance theory did offer a kind of answer: if thoughts, unlike words, resembled or pictured their objects, then this special relation could be the ultimate source (origin) of meaningfulness. Hobbes, however, couldn't use that answer. In making discourse and computation his model of ratiocination, he effectively gave thoughts the structure of sentences or formulae, composed of distinct, arbitrary symbols. Images don't have that structure: a picture of a fat man running is not composed of three separate symbols for "fat," "man," and "running." Thus the rules for manipulating such symbols, as in proofs and derivations, don't work for images either. Therefore, in proposing a computational account of thinking, Hobbes essentially forfeited the resemblance account of meaning. Yet he had, I think, nothing else to offer—which is to say he couldn't solve the mystery of original meaning or explain the basic difference between minds and books.

Descartes

Galileo and Copernicus were outstanding physicists and first-class mathematicians, but they weren't all that much as philosophers. Hobbes was a great philosopher but an inconsequential physicist and a truly abysmal mathematician. That towering French intellect René Descartes (1596–1650) was, on the other hand, a world-shaker at everything he touched. Commonly deemed the "father of modern philosophy," he might equally be called the father of modern mathematics; and his contributions to physics, though eventually eclipsed by Newton's, were for a generation the foundation of the discipline. Perhaps his most lasting mark, however, is in the theory of the mind; for, remarkable as it seems, this is the one point at which all his work comes together.

Descartes's career and outlook began with mathematics, in the history of which he is famed for two enormous innovations: first, his development of analytic geometry, and second, his astonishing interpretation of that development. Analytic geometry is that branch of mathematics in which geometric points, lines, and relationships are represented by numbers and algebraic equations, using the system we now call "Cartesian coordinates" (= graph paper; see figure 1). Of course, parts of the idea were old; astronomers, for instance, had long used a sort of coordinate system to keep track of stars. Descartes's contribution was a systematic approach to *solving geometric problems* by *algebraic methods.* And, while he was at it, he greatly improved the algebra of his time, basically introducing the notation we still use today.

Descartes began his famous work, *La Geometrie,* by saying:

> Any problem in geometry can easily be reduced to such terms that one need only know the lengths of a few straight lines in order to construct it.[13]

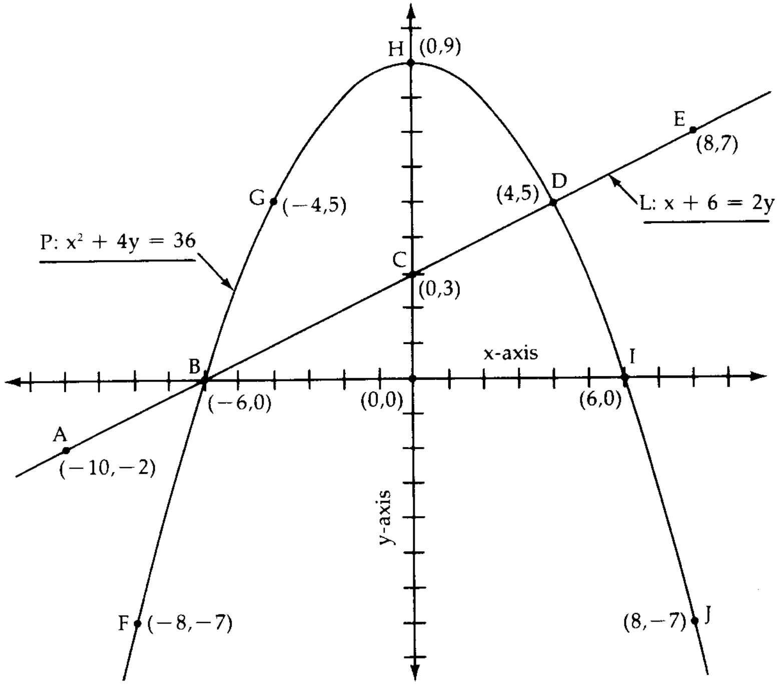


Figure 1 Cartesian Coordinates
A point in the plane corresponds to a pair of numbers, called its "X-coordinate" and "Y-coordinate," respectively. For instance, the point D above corresponds to the pair (4,5) because it is located 4 units out along the X-axis and 5 units up along the Y-axis. The values 4 and 5 (for x and y) also happen to satisfy the equation $x + 6 = 2y$; and, as you can see, the coordinates for points A, B, C, and E satisfy it too. In fact, all and only the points on line L have coordinates which satisfy that equation; hence the equation and the line "define" each other, given the coordinate system. And (first-order) equations like that are now known as "linear" equations.

In a similar way the coordinates of F, B, G, H, D, I, and J all satisfy the (second-order) equation $x^2 + 4y = 36$; and these points (plus all the others determined by the same equation) lie on the curve P, which turns out to be a parabola. (Other equations define circles, ellipses, and so on.) Notice, finally, if the above two equations are taken together and solved for x and y, the only two solutions are (−6,0) and (4,5), which correspond to points B and D, where the two curves intersect.

Most geometers in those days, Galileo and Kepler included, expressed quantitative relationships by means of geometrical proportions, such as:

Line (or area) A (in some diagram) is to B as C is to D

—which quickly got tedious and complicated. Descartes, however, saw a way to cut through all that. He regarded geometric proportions and numerical operations, like multiplication and division, as just *different forms* of a single more general relational structure. So instead of struggling with a lot of complicated proportions, Descartes calmly talked about multiplying, dividing, and extracting roots of his line segments (lengths)—and always got other line segments for answers.

Geometrically this is a bizarre idea. What could the square root of a line segment be? Moreover, the product of two lengths ought not to be another length but an *area*. Descartes's idea was to define some convenient "unit length" and then treat all those relations as special cases of proportions. Thus any given length is to its square root as the latter is to unity; or, unity is to one length as a second length is to the product of the two. In other words:

If: L is to R as R is to 1, then: $L = R^2$
If: 1 is to A as B is to C, then: $A \times B = C$.

All these relations are proportions among lengths; yet, in terms of them, the "algebraic" operations make perfect sense. The other side of the coin is that geometric proportions can themselves be re-expressed as algebraic equations, which in turn means (given the above quotation) that any geometric problem can be "translated" into algebraic notation and solved by algebraic methods.

As if unifying algebra and geometry weren't enough, Descartes also realized that physics was in the same boat. If physical laws could be represented geometrically and geometrical relationships could be expressed algebraically, then physics too could be formulated in algebraic terms. Today it seems so obvious that physical laws can be *equations* that we forget the intellectual leap required to conceive such a thing. In particular, we forget that science can be quantitative and mathematically rigorous without algebra (not to mention calculus); Galileo (and Kepler, etc.) established that, using geometry.

Inventing analytic geometry was Descartes's first great step; his second step essentially redefined the field of mathematics. He explains the basic idea in a semi-biographical passage in his *Discourse on Method*:

> Still, I had no intention of trying to learn all the particular sciences commonly called mathematics. For I saw that, although these sciences treat different objects, they all agree in considering nothing about the objects but the various relations or proportions found in them. Hence I thought it more worthwhile to examine just those proportions in general, free of any specific assumptions about the subject matter—except ones which might make understanding easier. And even then, of course, the assumptions were never binding, for my ultimate aim was a better ability to apply the same proportions to whatever other objects they fit.[14]

Galileo "abstracted" Euclid's methods away from purely spatial problems so he could apply them to physics, and now Descartes drives the principle to its radical conclusion: geometry, algebra, and physics are all equally just "applied math." Mathematics *as such* is not concerned with any specific subject matter (figures, numbers, motions, or what have you), but only with the very abstract relationships that might be found in them, or in any other objects.

Surprisingly, these same discoveries were just as revolutionary in the philosophy of mind as in mathematics. For the essential innovation is a reconception of the relation between symbols and what they symbolize—what we would now call the theory of meaning. Though mathematical notations were the model for Descartes's new ideas, he soon extended them to cover everything meaningful—especially thoughts. In other words, he regarded thoughts themselves as symbolic representations, fundamentally analogous to those used in mathematics; and, as a result, he could apply all his conclusions about representations in general to the special case of the mind. This profound maneuver, and its startling consequences, basically launched modern philosophy.

The new approach to representation has two distinct components: the negative half, rending symbol and symbolized asunder, and the positive half, getting them back together again. The negative half is clearly manifested in the mathematical work: the basic point is that algebraic formulations don't intrinsically represent numbers, and Euclidean formulations don't intrinsically represent geometric figures. Rather, algebra and geometry are two separate notations, each equally adequate for expressing numerical, spatial, kinematic (motional), or any other quantitative relationships. Nothing about either notation connects it, as such, to any particular subject matter.

Extend this negative realization to mental representations (thoughts), and you finally conclude the divorce of thought from thing, mind from world, that began so innocently in the old appearance/reality distinction. Suddenly a thought is one thing and its alleged object quite another—and there is no intrinsic relation between them. My thoughts of numbers are no more connected, as such, to numbers, than algebraic notation is connected, as such, to numbers. Those very same thoughts could equally well represent spatial facts or kinematic facts or even nothing at all (and perhaps I could never tell; see box 3). This disconcerting picture is the essence of the modern mind; we owe it all to Descartes.

The positive half—getting the symbols to symbolize again—turns out to be harder. If we read between the lines in Descartes, we can see him dividing it into two stages (with mathematics again the inspiring example):

1. What makes a notation suitable for symbolizing (representing) some subject matter?
2. What makes a suitable notation actually symbolize (represent) that subject matter?

Descartes barely answers the first question at all; but, via implication and creative anachronism, we can put good words in his

Box 3
Epistemological Skepticism

Being 'skeptical' means doubting, reserving judgment, or remaining unconvinced. 'Epistemological' means having to do with knowledge or the theory of knowledge. Epistemological skepticism is the infamous philosophical stance of doubting whether we can ever really know anything. A classic line goes like this: Given that our senses are often unreliable (we are fooled by illusions, hallucinations, dreams, etc.) and that even our most careful reasonings sometimes go astray, then how can we know for sure, at any particular moment, that we are not being fooled or irrational right then? But if we could never know *that*, we could never know anything.

Descartes used a skepticism like this to introduce one aspect of his new conception of mind (namely the divorce between thought and thing). But, as befits his genius, he first transformed the argument into something far more compelling and sinister. Suppose, he suggested, there were an "evil demon," divinely powerful but malicious, bent on deceiving me about everything. For instance, the demon might create in my mind sense impressions and even theoretical concepts that bear no relation whatever to the outside world; furthermore, he might cleverly arrange them all so that they seem to hang together and make perfect sense. Indeed, my whole life could be a single, diabolically orchestrated hallucination. How would I ever know?

How could Descartes ever dream up such a thing? After all, the illusion of the Sun rising depends directly on the real Earth turning; the ticklings around Galileo's nose are still intimately related to the shapes and motions in the feather. But the evil demon could strip us of all contact with reality whatsoever. Descartes could dream that up because he had a new vision of thoughts as mere symbols in a notational system; and he knew full well that such symbols could equally represent one subject matter or another, or none at all, and the system itself would be no different.

In my own view, epistemological skepticism has been

largely a digression within the history of philosophy and the cause of a lot of wasted effort. In any case the issue is certainly tangential in Artificial Intelligence, and we won't consider it again.

mouth. First, however, we should give a brief nod to his elaborate and sadly influential answer to the second question.

Notice that external notational systems aren't the problem. When external symbols actually represent, it's because they express thoughts, which already represent; they thereby acquire meanings, but only derivative meanings. The real issue, clearly, is those already meaningful thought-symbols; that is, it's a variation on our old friend, the mystery of original meaning. Unlike Hobbes, Descartes at least saw the problem; and he offered an amazing solution. In barest outline, he argued that (1) he could prove, just from what he found within his own mind, that a nice God exists; and (2) it wouldn't be nice to let creatures be radically misguided (especially if they worked conscientiously). Therefore, if we're conscientious, our thoughts will represent reality. Though Descartes's own version was a lot less crude, its main impact was still to send other philosophers after better answers.

Let's return to the first question: What makes a notation "suitable" for symbolizing some subject matter? It isn't just that convenient symbols can be invented for all the relevant items or variables—that, by itself, is trivial. Galileo didn't just say "the area within the triangle represents the distance traveled" and let it go at that. Nor is analytic geometry merely the clever idea of identifying geometric points with numerical coordinates (that wouldn't have cost Descartes much effort). No, what earned these men their reputations was demonstrating how, if you represented things in certain very specific ways, you could *solve problems.*

When a notation can be used to solve problems about a subject matter, then it is suitable for representing that subject matter. Galileo showed how, using his geometrical representations, he could *derive* his famous laws from a few simple assumptions; and thereby he also showed how Euclid's system can be suited to representing kinematics. Descartes did the same for algebra and geometry ("Any problem in geometry can easily be reduced . . .") and then recognized the general point: any number of different notations might be equally suitable for representing any number of different subject matters.

This point about solving problems has two subtle but important corollaries. First, it wouldn't do to get just a few scattered problems right, while giving silly (or no) results on others, for then the occasional successes would seem like mere flukes (or even hoaxes). In other words the notation-cum-problem-solving-method must form an *integrated system* that can be used systematically and reliably in a well-defined area. Second, solving problems obviously involves more than just representing them. There must also be various allowable steps that one can take in getting from (the representation of) the problem to (a representation of) the solution (for example, in derivations, proofs, etc.). Hence the integrated system must include not only notational conventions, but also *rules* specifying which steps are allowed and which are not.

Descartes himself didn't really say all this. Pioneering ideas often bubble beneath the surface in authors who can't quite articulate them. (Later writers will finally get those ideas explicit, while struggling with still newer ones.) Anyway, we can see rules and steps bubbling beneath Descartes's way of distinguishing people from "unreasoning" machines:

> For we can well imagine a machine so made that it utters words and even, in a few cases, words pertaining specifically to some actions that affect it physically. For instance, if you touch one in a certain place, it might ask what you want to say, while if you touch it in another, it might cry out that you're hurting it, and so on. However, no such machine could ever arrange its words in various different ways so as to respond to the sense of whatever is said in its presence—as even the dullest people can do.[15]

Incredible! In 1637 this man could imagine not only the loudspeaker, but also the talking Toyota; perhaps we can forgive him if he couldn't quite imagine mechanical reasoning. For to arrange words appropriately in response to the sense of a previous utterance is just to say something "reasonable," given the context. And what's reasonable (in Descartes's eyes) is determined by the rules of reasoning—that is, the rules for manipulating thought symbols, in the notational system of the mind.[16] So, essentially, Descartes is saying that machines can't think (or talk sensibly—he's anticipated Turing's test too) *because* they can't manipulate symbols rationally.

That, obviously, is hitting Artificial Intelligence where it lives. Yet the worry over "mechanical reason" is deep and challenging; it deserves a section of its own.

The Paradox of Mechanical Reason

Descartes was a *dualist*: he believed that minds and the physical universe were two entirely different kinds of substance. Minds, by their nature, can have thoughts in them—beliefs, desires, worries, decisions, and the like—all subject to the order of reason. Meanwhile, the physical universe can, by its nature, have bodies in it—physical objects and mechanisms—all subject to the order of physical law. Note the two radically different notions of 'in'. Thoughts are "in" minds, but not in the sense of being inside a three-dimensional container; minds and their "contents" have no spatial properties at all. The universe, on the other hand, essentially *is* space; all physical objects are within it spatially and always have definite sizes, shapes, and locations. It follows from this contrast that no mind can ever have a physical object in it and likewise that there can never be any thoughts in the physical universe.

Dualism actually has a strong commonsense appeal; and since it would rule out Artificial Intelligence at a stroke, we should pay attention. Suppose Frank remembers Frankfurt, or hankers for a Frankfurter. What could his thought be like *spatially* (the thought itself, not what it's about)? Is it one inch wide (or a millimeter or a mile)? Is it round like a ball (or conical or doughnut shaped)? There is something perverse about such questions; and the problem isn't just that thoughts are fuzzy or hard to measure. I can't even imagine a thought shaped like a fuzzy, one-inch doughnut. Location seems initially easier than size and shape: thoughts are "in our heads," aren't they? But if they have no size or shape, how

can they have a place? And anyway, when Frank remembers Frankfurt, exactly how far is it from his recollection to, say, his left earlobe? Of course, science may someday come up with surprising answers; but, on a common-sense level, Cartesian dualism does seem very reasonable.

There is, alas, one fundamental difficulty that no dualist has ever resolved: if thought and matter are so utterly disparate, how can they have anything to do with one another—how can they *interact*? This is the notorious *mind-body problem*, and it has driven dualists to some of philosophy's most desperate gyrations ever (see box 4 for a few samples). Here's how it goes. The laws of physics suffice to explain every movement of every material particle entirely in terms of physical forces; and these forces are all exactly determined by specified physical properties, like mass, distance, electric charge, and various quantum oddments. But if thoughts can have no size, shape, or location, they're even less likely to sport mass, charge, or queer quark quirks; hence, they can never exert any physical forces on matter (or vice versa, presumably). Consequently, all movements of material bodies can be completely explained without reference to anything mental.

So, for a dualist, the price of admitting mind–body interactions would be forfeiture of modern physics, which no mere philosopher could ever afford. On the other hand, foreswearing interactions is also rather awkward. Thus when I decide to raise my hand and then my hand goes up, it sure seems that my (mental) decision *causes* that (physical) movement. Conversely, the (physical) light rays entering my eyes sure seem to cause my (mental) visual experiences of the world. In short, mind–body interaction seems physically impossible; yet without it we could neither perceive nor act. So, despite all its intuitive appeal, dualism is a tough row to hoe.

Most alternatives to dualism turn out to be some sort of *monism*—theories that say there is really only one kind of substance instead of two. In the nineteenth century, the most popular variety of monism was *idealism*, according to which minds and ideas are the only ultimate reality; material objects were regarded either as purely illusory or as special "constructs," somehow built up out of ideas. In our own century, idealism has fallen on hard

Box 4
Dualist Desperados

INTERACTIONISM: Descartes himself actually maintained (to everybody's amazement) that mind and body *do* interact. This, of course, would be an eminently sensible view but for the minor inconvenience of contradicting everything else Descartes believed—and his reasons therefor. He did gingerly restrict the effect to subtle vapors in the pineal organ, a still mysterious body near the bottom of the brain; but somehow that didn't help much with the problem of principle.

PARALLELISM: According to this clever idea, mind and matter are related like two perfect clocks set in motion simultaneously at the creation. Each obeys its own laws and proceeds entirely independently; but, due to God's marvelous planning and workmanship, they "keep time" flawlessly and forever. Thus when the hammer crushes my thumb, and the pain instantaneously clouds my judgment, there is no causal connection, but only another "coincidence," like the noon whistle and the church bells always sounding at 12:00.

OCCASIONALISM: Though mind and body can never affect each other, God can affect anything. So another charming line has watchful Providence intervening helpfully on each "occasion" when mind and matter would interact, if only they could. For instance, that hammer doesn't really affect me (my mind) at all; but God alertly creates for me exactly the excruciation my thumb nerves signal—and then bravely forms on my carnal lips those colorful words I no sooner intend than regret.

EPIPHENOMENALISM: Here's one for agnostics. The universe is a superbly engineered machine, ticking and whirring smoothly, with everything complete and in order. Minds and conscious experiences play no role in the mechanism but are incidental by-products ("epiphenomena"), like the ticking and whirring. This peculiar approach is curiously ambivalent about

interaction: matter causes or "gives off" mind, but thought has no effect on matter. So we can watch the world go by, but we can't do anything about it (our impressions to the contrary being but a cruel hoax).

times, and the most popular monism is *materialism*, according to which (naturally) matter is the only ultimate reality. Materialists hold either that thoughts and ideas are purely illusory or else that they are special constructs, somehow built up out of matter. The difficulties that toppled idealism are not particularly relevant to AI; so we mention them only in passing and concentrate on the materialist side.

Materialism, however, has troubles of its own. For one thing, materialists find it hard to say anything terribly comforting about immortal souls; but we set that issue aside. Of more immediate concern is what to say about thoughts, if all reality is ultimately material. The crux of the issue is a deep and traditional conundrum, which I call the *paradox of mechanical reason*. Its resolution is, simultaneously, the philosophical foundation of the Artificial Intelligence boom and also the most attractive current alternative to hopeless dualism and zany idealism (not to mention vulgar behaviorism).

So what's the paradox? Reasoning (on the computational model) is the manipulation of meaningful symbols according to rational rules (in an integrated system). Hence there must be some sort of manipulator to carry out those manipulations. There seem to be two basic possibilities: either the manipulator pays attention to what the symbols and rules *mean* or it doesn't. If it does pay attention to the meanings, then it can't be entirely mechanical—because meanings (whatever exactly they are) don't exert physical forces. On the other hand, if the manipulator does not pay attention to the meanings, then the manipulations can't be instances of reasoning—because what's reasonable or not depends crucially on what the symbols mean.

In a word, if a process or system is mechanical, it can't reason; if it reasons, it can't be mechanical. That's the paradox of mechanical reason. Unfortunately, the issue is just too important to be quietly forgotten. So people have struggled with it courageously, generating a vexed and amusing history that is also somewhat enlightening (at least in retrospect).

Consider again the alternative where the meanings are taken into account. Ironically, the problem is essentially a reenactment, within monist materialism, of the basic dualist difficulty about interactions, only this time the mysterious troublemakers are meanings rather than thoughts. Materialists try to escape the interactionist trap by claiming that thoughts are really just a special kind of material object (viz., symbols), which therefore obviously can "interact" with matter. But the *meanings* of those symbols are not material objects (they are "abstract" or "conceptual" or something). The trouble is, they still have to affect the operation of the mechanism for the manipulations to be reasonable. Hence all the old embarrassments return about exerting forces without having any mass, electric charge, etc.: meanings as such simply cannot affect a physical mechanism.

But suppose this problem could be resolved. (I think many philosophers ignored it, convinced that it had to be resolvable somehow.) Would that finish the matter? Not quite; for the status of the manipulator remains disconcertingly unsettled. We're assuming, for the moment, that it manipulates thought symbols reasonably, by "paying attention" to the meanings of the symbols and the rules of reason. But how, exactly, does that work? Consider Zelda. We imagine a manipulator "reading" the symbols in her mind, figuring out what they mean, looking up various rules of reason, deciding which ones to apply, and then applying them correctly—which generally means "writing" out some new mental symbols (whatever Zelda thinks of next).

This manipulator turns out to be pretty smart. It can read and understand both the symbols it's working on and the rules it's following; it can figure things out, make decisions, apply rules to novel cases, do what it's told, and so on. But so what? Isn't Zelda the manipulator of her own thoughts, and doesn't she understand them as well as anybody? No! That cannot possibly be right. Zelda's thoughts and understandings are the symbols *being manipulated*; if carrying out the manipulations also requires thoughts and understandings, then these latter thoughts and understandings must be distinct from Zelda's. The point of the computational theory is to *explain* Zelda's thinking in terms of rational manipulations of her thought symbols. Her thinking cannot itself be employed in explaining the manipulations, on pain of rendering the account circular.

How humiliating! In order to explain thinking, the theory has to invent an inner "manipulator" who thinks, understands, decides, and acts all on its own. Philosophers have dreamt up various soothing names for this inconvenient little fellow, such as the "faculty of the will" or the lofty "transcendental ego." But the name which sticks is the one used by mocking opponents: he's the *homunculus* (which is just Latin for "little fellow"). Whatever the name, the bottom line is a simple, shattering question: If a thinking homunculus is needed to explain how Zelda can think, then what explains how the homunculus can think? A still smaller homunculus?

We should appreciate that this debacle follows directly from assuming that (rational) thought manipulations require "attention" to what the thought symbols and rules *mean*; for that's what entailed a manipulator who could understand and think on his own. In other words, sticking to the "reason" side of the paradox of mechanical reason leads to the homunculus disaster. What happens if we try the "mechanical" side?

Hume—The Mental Mechanic

Scotland's most celebrated philosopher, David Hume (1711–1776), was the first to spell out consistently the mechanical conception of thinking. The subtitle of his monumental *Treatise of Human Nature* (written in his twenties) announces the central plan of his entire philosophy: "An Attempt to introduce the experimental Method of Reasoning into Moral Subjects."[17] By "moral subjects" Hume meant not only ethics and the theory of justice but all of what he called "the science of man"—beginning with psychology. By the "experimental method of reasoning" he meant the methods of natural science, especially physics.

In other words, Hume proposed to establish a new category of human sciences, explicitly modeled on the wonderfully successful physical sciences. More particularly, he wanted to explain thinking and feeling in terms of how various mental mechanisms work; or, as he put it, to

> discover . . . the secret springs and principles by which the human mind is actuated in its operation.

In the same paragraph, Hume effectively compares his own efforts to those of the great English physicist Sir Isaac Newton (1642–1727), whom he describes as having

> determined the laws and forces by which the revolutions of the planets are governed and directed.

For he goes on to conclude:

> The like has been performed with regard to other parts of nature. And there is no reason to despair of equal success in our inquiries concerning the mental powers and economy, if prosecuted with equal capacity and caution.[18]

So Hume wants to discover the laws and forces by which ideas are governed and directed; he wants to be the Newton of the mind.

The centerpiece of Hume's theory is the famous principle of "association of ideas," which he adopted pretty much intact from the English empiricist John Locke (1632–1704), who, in turn, got a good deal of it from Hobbes. Locke is the philosopher who said the mind starts out as a *tabula rasa* (clean slate), on which experience "writes" the sense impressions basic to all knowledge; then the mind's natural "association" combines and recombines all these fundamental "ideas" into ever more complex and sophisticated science. Locke's main concern was to legitimate Newton's empirical method against the Cartesians (who still wanted physics to be "intuitive," like mathematics). He was using physical knowledge as a kind of acid test for philosophical theories: no theory of the mind could be right, in Locke's view, unless it could show why Newtonian science is good science.

Hume, on the other hand, was inspired by Newton more directly: his theory was designed not merely to accommodate Newton's physics but to imitate it. That is, Hume really put forward a "mental mechanics"; his impressions and ideas were not so much the basic evidence on which all knowledge rested but rather the basic pieces out of which all knowledge was composed—or, better yet, the basic "corpuscles" (particles) to which all the mental forces and operations applied. And the association of ideas is just the general category for all these forces; Hume himself describes it as

> a kind of ATTRACTION, which in the mental world will be found to have as extraordinary effects as in the natural, and to shew itself in as many and as various forms.[19]

Obviously he's thinking of universal gravitation (and maybe magnetism).

It is important to appreciate how different Hume is not only from Locke, but also from Hobbes. The latter shared Hume's "mechanical" outlook, even to the point of regarding thoughts as actual movements of matter in the brain. The difference is that Hobbes took thoughts on the model of numerals or words: physical symbols to be manipulated according to the rules of calculation or rational discourse. And that led, as we saw, to the question of how there could be such manipulations, for they seem to require an intelligent manipulator (the homunculus). Hume, however, had no such problem; his science of the mind was to be entirely analogous to physics, plus maybe a bit of engineering.

But, crucially, this analogy occurs at the level of explanation. Unlike Hobbes, Hume didn't claim that ideas actually *are* physical (he didn't, in fact, seem to care much about the materialism/dualism issue). Rather, he said, ideas are *like* physical particles, in that their interactions are to be explained in terms of natural forces governed by natural laws. Hence the account doesn't beg its own question by assuming some behind-the-scenes intelligence making it all work. There is no more difficulty about "how" ideas obey the laws of association than there is about "how" planets obey the law of gravity; they just do. In other words, Hume, like Newton, can say, "I frame no hypotheses."[20]

Unfortunately, however, in avoiding the homunculus pickle, Hume landed himself in another: What makes his ideas *ideas*, and what makes their interactions count as *thinking*? What, in short, is mental about the mind? Hume has so thoroughly eliminated meaning and rationality from the basis of his account that he might as well be talking about some new and strange kind of physics. In a way, of course, that was the whole point; but he now owes us an explanation of how meaning and reason get back into the picture; otherwise he hasn't produced a "mechanics of the *mind*" after all. Unfortunately, Hume has no such story to tell; indirectly, I think, he even admits it, in his own peculiar brand of "skepticism" (see box 5).

But this just brings us back to the paradox of mechanical reason: either meanings matter to the manipulations, in which case the processes aren't really mechanical (they presuppose an homunculus); or else meanings don't matter, in which case the processes aren't really rational (they're just some meaningless "machine-like" interactions). Hume is simply caught on the second horn of the dilemma. Historically, this is the point at which transcendental idealism thundered to the rescue; that is, Kant, Hegel, and their legions bravely returned to the first horn and gave up on matter. But, as mentioned earlier, this otherwise remarkable and difficult episode is really a digression within the pedigree of Artificial Intelligence. So we will skip over it and plunge back in with confident and all-conquering twentieth-century materialism.

Box 5
Semantic Skepticism

In the section of the *Treatise* entitled "Of scepticism with regard to the senses," Hume describes what he calls the "double existence" theory, according to which our perceptions and the objects they (allegedly) represent are distinct entities. Perceptions are fleeting and dependent on us, whereas objects are supposed to be durable and external; moreover, objects supposedly cause perceptions, which, in turn, resemble them. But Hume (as you might have guessed) thinks all of this is rubbish. Our minds work only with perceptions themselves; that is, we never have any "direct" experience of objects, independent of perceptions. Further, the existence of perceptions could never *logically* imply that anything else exists. Hence these "external objects" can be nothing but figments of our imaginations.

Hume has a subtle and ingenious story about why our imaginations do such a thing to us, why the confusion itself is so irresistible, and why philosophers have no hope of ever clearing it up. But the fact remains that, according to Hume's own theory, mental representation of nonmental objects is inconceivable. Thus, insofar as the meaningfulness of thought essentially involves "representational content," Hume has no place for it. So while Descartes's skeptic asks: "How can I know what I know?" Hume's skeptic grapples with the more basic question: "How can I even mean what I mean?"