CONTRADICTORY CLUES AS TO THE ORIGIN OF THE MARTIAN MOONS

JOSEPH A. BURNS
Cornell University

The meager available information that is pertinent to the origin and evolution of the Martian satellites is contradictory. The known physical properties of the Martian moons (density, albedo, color and spectral reflectivity) are similar to those of many C-type asteroids, the dark “carbonaceous” objects abundant in the outer belt but scarce near Mars; thus this line of physical evidence suggests that Phobos and Deimos are captured bodies. In contrast, calculated histories of orbital evolution due to tides in the planet and in the satellites indicate that these small craggy moons originated on nearly circular, uninclined orbits not far from their current positions; hence dynamicists prefer an origin in circum-Martian orbit. Ways are described in which these apparently contradictory viewpoints may be reconciled, although a definitive answer to the origin of the Martian satellites will almost surely have to await in situ measurements.

During the 1989 encounter of the Phobos spacecraft with its namesake satellite, the popular press was fond of writing that the Martian moon was of interest primarily because it is a captured asteroid. However, as will be described here, the scientific jury actually remains divided on the question of this satellite’s origin. Dynamicists argue that the present orbits (see Table 1) could not be produced following capture, and so the objects must have originated near Mars, where the satellites are found today, even if they do not appear as though they had been born at 1.5 AU. Virtually all other scientists maintain that every observable physical property indicates that the Martian satellites once resided in the outer asteroid belt (at ~ 3 AU), and thus the objects must be captured, although the precise mechanism whereby that occurred is often conveniently unspecified.
TABLE I
Dynamical Parameters for the Martian Satellites

<table>
<thead>
<tr>
<th></th>
<th>Phobos</th>
<th>Deimos</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$(km)</td>
<td>9378.5 (2.76 $R_{\oplus}$)</td>
<td>23459 (6.92 $R_{\oplus}$)</td>
</tr>
<tr>
<td>Period</td>
<td>7hr 39 min</td>
<td>30 hr 18 min</td>
</tr>
<tr>
<td>$n$ (deg/day)</td>
<td>1128.8446</td>
<td>285.16191</td>
</tr>
<tr>
<td>$e$</td>
<td>0.01515 (±0.00004)</td>
<td>0.000196 (±0.000034)</td>
</tr>
<tr>
<td>$i$ (deg)$^b$</td>
<td>1.068 (±0.001)</td>
<td>1.789 (±0.003)</td>
</tr>
<tr>
<td>$l$ (deg)$^c$</td>
<td>0.00934</td>
<td>0.8965</td>
</tr>
<tr>
<td>Mass (kg)</td>
<td>$1.08(±0.01) \times 10^{16}$</td>
<td>$1.80(±0.15) \times 10^{15}$</td>
</tr>
<tr>
<td>Axes (km)$^d$</td>
<td>$13.4(±.4) \times 11.2(±.2) \times 9.2(±.1)$</td>
<td>$7.5(±.3) \times 6.1(±.3) \times 5.2(±.2)$</td>
</tr>
</tbody>
</table>

$^a$Orbital parameters are taken from Chapront-Touzé (1990) but are similar to those given by Sinclair (1989) and Shor (1988). Phobos' mass comes from Avanov et al. (1989), and the semiaxes of assumed triaxial ellipsoidal shapes from Thomas (1989); cf. Duxbury and Callahan (1989); Duxbury (1989) instead fits a 6th degree and order model to Phobos' surface with a typical error of a few hundred meters. See also Burns (1986a).

$^b$Referral to the Laplace pole.

$^c$Inclination of the local Laplace pole relative to Mars' rotation pole.

$^d$Slightly revised values are given in chapter 36.

This chapter will review the lines of evidence for capture as well as for in situ origin and will discuss the inferences made therefrom. Throughout, ways will be suggested whereby the nominal models of the two camps may be modified so that, on the one hand, primordial Martian moons could exhibit properties similar to objects from the outer asteroid belt or, on the other hand, orbital evolution scenarios could be misleading. This chapter begins by reviewing ideas concerning the origin of Phobos and Deimos.

I. PROPOSED ORIGINS

Little thought has been devoted to developing detailed models for the origin of Phobos and Deimos, despite the uniqueness of the Martian moons as the only extant small satellites of the terrestrial planets (see, e.g., Stevenson et al. 1986). Nevertheless, as mentioned above, two camps have formed, one that argues for accretion in orbit and the other that maintains that the moons of Mars originated elsewhere only to be later captured by Mars.

According to the accretion school (Safronov et al. 1986), the Martian satellites, like other regular satellites (see Burns 1986b), are an unavoidable consequence of planetary formation. Circumplanetary swarms of orbiting debris result when heliocentric particles collide inelastically within a planet's sphere of influence. Once a swarm surrounds a planet, some fraction of it is gradually lost as this orbiting material systematically decays down to the planet's surface, while the swarm itself is continually replenished through ongoing collisions with other heliocentric particles. The processes that gov-
ern the agglomeration of the Martian satellites might be similar to those that produced the regular satellites of the giant planets, although in the latter case gas-dominated accretion was more probable. The small sizes of Phobos and Deimos compared to the satellites of the giant planets may be attributed to the violent bombardment of Mars’ locale by invading planetesimals from the zone of Jupiter (Safronov et al. 1986).

As the final remnant of the nebula from which Mars itself grew, the satellites, according to the accretion school, should be composed of material like Mars. They will, however, contain an especially high proportion of components that were brought in with the last heliocentric objects to be acquired by Mars; of course, similar constituents should also form a disproportionate fraction of the surface layers of Mars. As this late-arriving material might have originated from a different region of the solar nebula than bulk Mars’ material, the moons could have a composition separate from that of the planet as a whole. Extending this idea to the very last accreted material, one might think that the satellites themselves could be coated by a thin veneer having an especially distinctive make-up. However, even if this were the case initially, subsequent bombardment by stray comets and asteroids should have disrupted any objects that originally accreted about Mars (Shoemaker 1989) so that upon their re-accretion in orbit, the contemporary Martian moons may have been homogenized. However, this remixing process has not been studied and the only available evidence (from collected meteorites) shows that, at least on the regoliths of the small bodies where meteorites originated, energetic events still allow significant heterogeneity in small samples. Perhaps on the larger scale or with the re-accumulation of lost ejecta by the Martian moons due to their orbits about the planet, the satellites should be homogeneous but we simply do not know.

Permanent capture, the alternate proposal for the satellite’s origin, requires energy loss; otherwise the time reversibility of Newtonian dynamics would allow escape ultimately (Burns 1986b). Sufficient energy loss transforms a hyperbolic orbit into a bound elliptic one as long as the loss occurs somewhere within Mars’s sphere of influence, a region of some several hundred Martian radii surrounding the planet within which the planet has gravitational dominance over the Sun (cf. Hamilton and Burns 1991). Tidal capture, which not too many years ago was the only capture mode thought to be relevant to the birth of Earth’s Moon (see Burns 1977; Boss and Peale 1986), is especially unlikely for the Martian moons since their small sizes mean that relatively little energy can be drained from a hyperbolic orbit during a single pass of the planet. Hence, by elimination, aerodynamic drag has been taken as the preferred process for capture of the Martian satellites; it has also occasionally been considered as a mechanism to capture the Moon (Nakazawa et al. 1983). Usually the drag has been considered to take place in a nebula that surrounds the planet shortly after it formed. Evolution through such a nebula will essentially always cause the orbital semimajor axis to collapse,
the orbit to circularize and the inclination to decay (cf. Pollack et al. 1979a). To have a reasonable probability for capture at many planetary radii, a fairly substantial nebula is required but, in such a nebula, evolution will be rapid so that captured objects are swiftly transported to the planet’s surface. That is, there is a fundamental conundrum associated with capture by nebular drag: if capture occurs close-in, where the nebula is dense and retention more likely, then the captured body is likely to be gone after a few orbits; whereas, when the energy loss transpires at great distances from the planet, where the orbit will not evolve rapidly, the process becomes highly unlikely because the slow evolution means that there is little, if any, nebula to decelerate the particle in the first place. Unfortunately, it is not possible to go beyond these generalities to clarify the limits of plausible gas density histories that would lead to probable captures.

Burns (1978) and Pollack et al. (1979a) have described qualitatively the possible capture of a single object by aerodynamic drag in a circumplanetary envelope about Mars. Burns suggested that the current Martian configuration, where individual satellites lie on either side of synchronous orbit, might have arisen by a sequence of two events. First, a heliocentric planetoid is captured by Mars through the action of gas drag and, under the same influence, evolves to near the synchronous orbit position, where, for a sufficiently thin nebula, the evolution will effectively cease because there is little relative velocity between the captured object and the nebula. Then, that captured object is struck by another heliocentric projectile and shattered into two main fragments, the larger of which (by happenstance) lands inside synchronous orbit while the smaller falls beyond the synchronous position; these two objects then evolve tidally through the cons, with the bigger one moving faster. Given even fewer details, Shoemaker (1989) mentioned a similar scenario in which Phobos and Deimos were split apart near synchronous orbit during the heavy bombardment epoch.

Capture by an extended protoatmosphere about Mars was first considered in detail by Hunten (1979a) and more recently was pursued by Sasaki (1990). In the first of these capture scenarios, the atmosphere was initially modeled as a slowly rotating condensation of the solar nebula with a surface pressure of a few tenths of a bar, but under such circumstances capture was found to be improbable. The likelihood improved when Hunten took the atmosphere to be rapidly rotating or its density to be 1 or 2 orders of magnitude higher. Hunten considered the high-density case in some detail and argued that the evolution was halted once the upper atmosphere rapidly dissipated following the removal of the solar nebula. In this scenario, as in that of Pollack et al., many objects were captured by proto-Mars and then orbitally evolved inward and were lost to its surface: Phobos and Deimos are distinguished solely by being the final planetesimals to approach Mars while it still retained its relatively dense surrounding nebula.

Sasaki’s study addresses numerically the history of captured objects as
their orbits evolve through an extended atmosphere in the context of a full three-body problem. He argues that, at great distances from the planet where the initial capture takes place, solar tides produce important perturbations to the orbit. He considers evolution through an extended H₂-He atmosphere that has condensed from the solar nebula as such an atmosphere will have a larger scale height and gas density than would that of a degassed secondary atmosphere. In these models, the periapsis distance initially increases following capture, and that extends the evolution time so that, for a careful choice of initial conditions, it may reach $10^3$ to $10^4$ yr. Sasaki expects that this time could be lengthened more by lowering the nebular density even further to the point where the capture time could equal the lifetime of the solar nebula. It will be interesting to learn, through further numerical exploration, how narrow the windows that allow this fate.

Hartmann (1987, 1990) points out that the irregular outer satellites of Jupiter and Saturn, which are generally believed to be captured objects (see Burns 1986b), are apparently all members of spectral class C, which dominates the central part of the asteroid belt. He maintains, therefore, that these captured satellites originated in the asteroid belt and were scattered by Jupiter throughout the solar system. Since Phobos and Deimos also seem to be similar to C objects, Hartmann claims that they too were part of this scattered and captured population.

II. CLUES TO ORIGIN: PHYSICAL PROPERTIES

Essentially all known physical characteristics of the Martian satellites are consistent with the proposition that Phobos and Deimos are immigrants to the Mars region from the outer portion of the main asteroid belt. As chapter 37 describes in more detail, the limited available data on the bulk densities and on the surface properties of the Martian moons show that these characteristics of the satellites lie in the general range expected for minor planets from the outer asteroid belt, say the region from 2.8 to 3.3 AU. While this viewpoint has been generally accepted for many years, the case was strengthened considerably by several findings in the late 1980s.

The first piece of recent evidence concerns the most fundamental physical property. Estimates of the bulk density of the inner Martian satellite were narrowed by the Phobos mission’s imaging team (Avsenov et al. 1989, 1991) to 1.90 ($\pm$ 0.10) g cm$^{-3}$, which falls at the lower end of the error bound of the most recent Viking determination, 2.2 ($\pm$ 0.3) g cm$^{-3}$ (Duxbury and Callahan 1989). Since the volume of Deimos is poorly defined, this satellite’s density is not as constrained, 1.7 ($\pm$ 0.4) g cm$^{-3}$ (Duxbury and Callahan 1989), but once again is relatively small compared to most likely constituents.

These low satellite densities approach those of the most volatile-rich carbonaceous chondrites CIIs and CMs, which have the lowest densities of
any meteorites. CI's have densities as low as \( \sim 2.2 \text{ g cm}^{-3} \) whereas other classes of carbonaceous chondrites have densities in the range 2.5 to 3.5 g cm\(^{-3}\) (Mason 1962; cf. Wasson 1974). These densities are significantly less than those of the minerals out of which Mars is thought to have accumulated; accordingly they have been interpreted by some to suggest that the satellites are carbonaceous. However, the densities of the materials that most closely match Phobos's spectral reflectivity (see below), namely, anhydrous and hydrous carbonaceous chondrites or optically blackened mafic material, are not nearly as low as these numbers (Britt and Pieters 1988, Bibring et al. 1989).

If one does not wish to accept that these measured low-bulk densities require unaltered carbonaceous material, another possible interpretation of them is that, in addition to dense Martian material, the satellite interiors contain a substantial component of water ice; this volatile compound can be shielded at all depths greater than tens-to-hundreds of meters over the age of the solar system by an insulating satellite regolith having nominal properties (Fanale and Salvail 1989). The presence of some water ice in the vicinity of Mars at the time of its origin is not unreasonable (Lewis 1974a) as temperatures today in that locale are marginally low enough for its condensation and, at the time of the satellites' birth, the Sun is thought to have been cooler; moreover, the immediate neighborhood of Mars as the planet formed may have been screened from direct solar radiation by the opacity of any nebula surrounding the planet.

Another explanation for the reduced densities of the Martian moons may be that the satellites are totally unconsolidated (even at depth) and contain significant open pore space; after all, the highest internal compressive stress experienced in Phobos is \(< 100 \text{ mbar} (10^4 \text{ Pa}) \) and, in Deimos, it is even smaller (Dobrovolskis 1982). These pressures are scarcely enough to overcome the cohesion of the lunar soil and are certainly not sufficient to eliminate the pore space in unconsolidated geologic materials nor to crush such materials. A final possibility is that the surface layers of the moons are deep and especially porous, much like those suggested for Mimas (Dermott and Thomas 1988). Several comments are in order in connection with these last two ideas. First, lunar rocks that start with densities of approximately 3.0 to 3.3 g cm\(^{-3}\) develop densities of about 2.0 g cm\(^{-3}\) during the process of brecciation through impact in the lunar regolith (Britt and Pieters 1988). Second, such low densities might result as a consequence of the satellite's primordial accretion (or re-accretion out of orbiting debris following a catastrophic breakup) in the low gravity environment found in circum-Martian orbit. To reconcile the discrepancy between Phobos's measured density and those of suggested meteorite analogs requires porosities of 10 to 30% in the case of lower-density carbonaceous chondrites and 40 to 50% for the higher-density black chondrites (Hartmann 1990; Murchie et al. 1991a). The lower porosities are consistent with values of 10 to 24% found in some meteorite breccias (Wasson 1974) and measured for lunar regolith breccias of 30% or
more. However, it seems improbable that the high values required by a black chondrite composition can be achieved. Finally, the porous regolith model appears to be inconsistent with the measured amplitude of 0.8 (± 0.4) for Phobos' libration, which implies that this satellite's density is uniform with depth (Duxbury 1989). Borderies and Yoder (1990) point out that Phobos' gravity field induces a secular change in the position of the satellite's pericenter; from a recent model of the observed orbital motion, they argue that the satellite must have a nonuniform density although an even more recent orbital model (Chapront-Touzé 1990) is found to be consistent with a constant density moon. The porous regolith model also does not agree with the interpretation of radar returns from Phobos that indicates the bulk density of its surface is 2.0 (± 0.4) g cm⁻³ (Ostro et al. 1989).

So, in summary, the measured densities of the Martian satellites are surprisingly low for any meteorite class. This problem can be eliminated by mixing in some low-density volume, either pore space or ice. However, once one admits the possibility that some low-density material is present, then the densities of the satellites are no longer particularly diagnostic of composition because any intrinsically high density can be accommodated by just stirring in more pore space.

The second piece of evidence pertinent to satellite origin concerns the inferred composition of the surfaces of the Martian satellites. When interpreting this evidence, it is worthwhile to recall first that only the topmost layers are probed by remote-sensing techniques and that these layers may not indicate the true nature of the bulk interior. Spectral reflectance data for Phobos and Deimos obtained from the ground and from spacecraft are plotted in Figs. 1 and 2 (cf. Figs. 1 and 2 of chapter 37). Figure 1 compares the reflectances of Phobos and Deimos with those of several selected meteorites and the asteroid Ceres (cf. Fig. 1 of Britt and Pieters 1988, Fig. 1 of chapter 37 and Fig. 2.1.3 of Wetherill and Chapman 1988). Figure 2 shows data obtained by the Phobos spacecraft and illustrates possible matches to some meteorites (Murchie et al. 1991a).

The satellites have quite low geometric albedos (~ 0.05) and generally flat spectra in the visible, although they drop off at the shortest wavelengths. Black, neutral objects such as these are common in the outer part of the main asteroid belt; in contrast, asteroids closer to the Sun are brighter and inferred to be siliceous while materials from farther out in the solar system (e.g., the Trojan and Hilda asteroids) tend to be dark and red (Hartmann 1987). Bibring et al. (1989) argue that the very low albedo of Phobos and the satellite's weak hydration band (which is significantly shallower than that of Mars) imply that the satellite's surface composition is most similar to metamorphic grade 3 meteorites (which show no visible alteration by groundwater) rather than the water-rich C1s or C2s (cf. Clark et al. 1986); these are the most dense of the chondrites and thus do not give a match to the satellite densities.

Murchie et al. (1991a) discern at least four recognizable spectral units
on the satellite (see Fig. 2). They claim that the large color ratio exhibited by one of these spectral units, the "blue" material found on Phobos, is inconsistent with a carbonaceous chondritic composition, but is comparable to that of an assemblage of mafic minerals like those forming black chondrites, which are anhydrous, optically darkened, ordinary chondrites. From its color ratio and its ultraviolet-visible spectral properties, the "bluish gray" material also has black chondrites as its closest spectral analog. Both carbonaceous chondrites and black chondrites match the ultraviolet-visible spectra of the "reddish-gray" and the "red" units identified by Phobos investigators. Black chondrites are also anhydrous but, as previously stated, have high densities in the range of 3.4 to 3.8 g cm$^{-3}$ (Wasson 1974).

The near-infrared reflectance data obtained by the Phobos spacecraft (Bibring et al. 1989) indicate that the surface material exhibits a weak and spatially variable red slope upward between 0.8 and 3.0 μm, and demonstrates no strong water of hydration signature in the 3-μm region (Langevin et al. 1990).

Most recently, Murchie et al. (1991b) attempt to integrate results from the three multispectral detectors, aboard the Soviet spacecraft (the VSK "visible" and "near infrared" channel [Avenesov et al. 1990], the KRPM 0.3-0.6 μm instrument [Ksanfomality et al. 1991; see Fig. 2] and the ISM infrared
imaging spectrometer (Langevin et al. 1990)]. They argue that two fundamental spectral units are seen within the region where the detectors have overlapping coverage. The “bluer” of these two materials is considered to be ejecta from the large crater Stickney having a black chondrite-like composition. The “redder” unit is believed to be a mix of a red component (argued to be made up of a low-density carbonaceous chondrite material such as CM chondrites) having a weak or negligible 1-μm absorption and a gray component (averred to be like a black chondrite). Murchie et al. (1991b) maintain that this mixture reconciles the measured low density with their preferred spectral analog of black chondrites. They then go on to claim that such a mixture implies formation from more than a single source, whether planetary or asteroidal.

The spectral data on Deimos has been recently surveyed by Bell et al. (1989b) who conclude that the satellite’s reddish slope in the near infrared, flat visible continuum and strong ultraviolet absorption are seen only in a few asteroids lying between C and P classes. In their own infrared data, Bell et al. (1989b) find just a weak 3-μm band on Deimos suggesting that at least this moon’s surface is anhydrous. This observation, of course, says nothing about deeper layers which may contain hydrated clay minerals or even water ice.

The close correspondence of the spectral reflectances of the Martian satellites to various meteorites has been used to aver that the bodies are the same (see, e.g., Bell et al. 1989b), but one should be chary of such claims because existing remote-sensing data can do no more than provide nonunique identifications. While the dark neutral character of these spectra must imply some opaque component, that material could be carbon, Fe-Ni metal, troilite or something else. Furthermore, relatively small amounts of any opaque component could dominate the spectra and confound their interpretation (Clark et al. 1986). The final note to make is that, at present, it is not possible to make a one-to-one match between the spectra of the Martian satellites and those of any specific class of chondrite; all those that match the visible spectrum have a strong clay band at 3 μm, something not seen for the satellites.

Processes in the regoliths of asteroids almost certainly have altered the optical signatures of these surfaces. This has been pointed out particularly by Britt and Pieters (1988) but is contested by J. F. Bell (personal communication, 1990) who argues that meteorite breccias show that the typical surface material on the meteorite source bodies is not highly shock-blackened. In some respects, processing of the surface materials on the Martian moons will be similar to that on minor planets, but in one possibly profound way it will differ. Since Phobos and Deimos reside in Mars’s gravitational potential well, almost all the impact ejecta that is thrown off their surfaces will ultimately re-impact them (Soter 1971). This means that the constituents comprising the outer skin of the satellites have suffered innumerable collisions although most of these will occur at low impact velocities; exactly how, or even whether,
Fig. 2. Spectral reflectance data of Phobos taken by the KRKM instrument on its namesake spacecraft and compared to meteorite data. (a) 0.3–0.6 μm reflectance spectra of particular color-ratio units located in different geologic settings on Phobos, as identified at the right. All curves are normalized at 0.55 μm (square symbols) and are offset from one another by 10% for clarity. The left-hand scale only gives relative values, and the values for 0.4 to 1.0 on the scale pertain to the bottom curve. The curves were calibrated using a representative standard and a composite whole-disk spectrum of Phobos compiled from Mariner 9 UVS, Viking Lander camera and telescopic observations, and then resampled into the frequency response functions of the KRKM channels. (b) Normalized reflectance spectra of carbonaceous chondrite meteorites that are possible spectral analogs to Phobos, resampled into the KRKM bandpasses. (c) Laboratory spectra of blackened and less altered normal ordinary chondrite meteorites. (d) Laboratory spectra of differentiated meteorites, including SNCs and the eucrite assemblage; SNCs are closely related achondrites that are believed to come from Mars, while eucrites, also achondrites, are magmatic rocks. Spectra are ordered approximately by decreasing redness (figure from Murchie et al. 1991a).
these many low-speed collisions will shock-alter the surface materials in some unique way is unknown.

Besides having experienced this unusual collisional environment, materials on the surfaces of the Martian satellites will have also spent considerable time isolated and exposed in space. In contrast, only a small fraction of the material resident on the surfaces of comparable-sized asteroids will have undergone a similar history; indeed, it is suspected theoretically that most of a small asteroid's surface will be bare rock rather than accumulated layers of recycled regolith (Housen and Wilkening 1982). Thus Britt and Pieters
(1988) caution that measured optical characteristics of the Martian satellites may not reflect those of the underlying material (cf. Murchie et al. 1991a).

Among the most interesting findings of the 1989 Soviet space mission insofar as elucidating the satellite’s origin is that Phobos has a heterogeneous surface. It exhibits 10% variations on a kilometer scale in albedo and in the strength of the hydration band (Bibring et al. 1989) as well as in thermal reflectivity (Ksanfomality et al. 1989). Murchie et al. (1991a, b; Murchie 1990; cf. Bell et al. 1990a) identify lateral variations up to 45% in the surface color ratio and ultraviolet-visible (KRFM) spectra; they associate these variations with large craters and propose that some other distinct material has been brought up from depth in these regions. Viking measurements had earlier shown some variations in albedo and in thermal inertia (see Veverka and Burns 1980; chapter 37). If the spectral and albedo variation seen on Phobos indicated chemical or mineralogical differences across the surface, how such variations might occur on a satellite that is believed to have re-accreted many times over (Shoemaker 1989; see the discussion in Sec. I) is a mystery. Perhaps a nonuniform Phobos would be easier to understand if the satellite accreted inhomogeneously in the asteroid belt and was subsequently (but gently) captured intact. On the other hand, the observed variations might be caused simply by the effects of particle size which likely would differ in the neighborhood of craters (J. F. Bell, personal communication, 1990).

Essentially all of the discoveries about the Martian moons made during the 1980s have only strengthened the convictions of many planetary scientists that Phobos and Deimos are captured asteroids. However more thought should be devoted to seeing whether we are being misled by nonunique remote sensing data and incomplete information.

III. CLUES TO ORIGINS: ORBITAL HISTORIES

Sharpless (1945), using five sets of position measurements taken in the period 1877–1941, proposed that Phobos’s orbit was secularly approaching Mars. Although this interpretation may have been premature at the time it was given, subsequent observations have shown it to be true, albeit with an actual value for the secular acceleration of only about one-half that computed by Sharpless. Modern determinations of the acceleration in the longitudes of the satellites, which combine both groundbased and spacecraft observations of the Martian satellite positions relative to Mars and one another, are listed in Table II. Improvements in these values will come only gradually with refined positions from groundbased observatories and especially from spacecraft missions to Mars (cf. Kolyuka et al. 1990), as well as perhaps radar returns (Ostro et al. 1989). Phobos’ measured acceleration corresponds to an added 15° of longitudinal displacement since the satellite’s discovery in 1877.

The orbital acceleration of Phobos, compared to its Keplerian rate, is now generally acknowledged to be due to solid-body tides raised by Phobos
### TABLE II

<table>
<thead>
<tr>
<th></th>
<th>Phobos</th>
<th>Deimos</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.002625 ± 0.000056</td>
<td>0.000072 ± 0.000028</td>
<td>Shor (1988)</td>
</tr>
<tr>
<td></td>
<td>0.002474 ± 0.000034</td>
<td>0.000056 ± 0.000158</td>
<td>Sinclair (1989)</td>
</tr>
<tr>
<td></td>
<td>0.002480 ± 0.000034</td>
<td>0.000004 ± 0.000016</td>
<td>Jones et al. (1989)</td>
</tr>
<tr>
<td></td>
<td>0.002539 ± 0.000017</td>
<td>0.000101 ± 0.000013</td>
<td>Chapront-Touzé (1990)</td>
</tr>
</tbody>
</table>

*\(n\) given in deg-yr\(^{-1}\).

in Mars (see Burns 1977, 1978, 1986a; Pollack 1977). Mechanisms, including some rather bizarre ones, that earlier had been proposed to account for Phobos’ acceleration are recalled in Shklovskii and Sagan (1966) and Burns (1972). If the tidal interpretation is correct, then one can predict the future orbit and also infer its past history. We start by considering the simplest evolution model which considers only the lowest-order tidal response, assumes a circular, uninclined orbit and no changes in the dissipative properties of Mars, and ignores all resonant interactions. With this simple model, the time evolution of the orbital semimajor axis \(a\) may be written (Burns 1986a, Eq. 13)

\[
\dot{a} = 3 m k_t (G/aM_\odot)^{1/2} (R_\odot/a)^3 \sin 2\varepsilon
\]  

(1)

where \(m\) is Phobos’ mass and \(M_\odot\) that of Mars, \(G\) is the Newtonian gravitational constant, \(R_\odot\) is the Martian radius, and \(\varepsilon\) is the angular displacement (in radians) between Phobos’ longitude and the maximum of the tidal bulge on the planet (see Fig. 3). The tidal Love number \(k_t\) which depends on Mars’ internal rigidity and density distribution, has been estimated to be \(\sim 0.08\) (Ward et al. 1979). For slow motions of a linearly dissipative system, the lag angle can be expressed as \(\sin 2\varepsilon = 1/Q\), where \(1/Q\) is the specific dissipation function, which is proportional to the energy loss per cycle (see Burns 1977, 1986a). By differentiating Kepler’s third law, \(\dot{a} = -2/3(n/n)a\), where \(n\) is the mean motion of Phobos about Mars. Phobos’ measured secular acceleration \(\dot{a}\) and Eq. (1) can be employed to obtain \(Q\) for Mars, which is found to be \(\sim 100\) to within a factor of about 2; this value is comparable to that of the solid Earth (see references in Burns 1986a). We will return to discuss the future and past of Phobos but suffice it to say now that Eq. (1) predicts a collapse time in which Phobos will strike Mars of \(< 40\) Myr if the satellite were able to remain intact all the way to the planet’s surface.

The same tidal processes that inexorably drag Phobos towards Mars force Deimos away from the planet because the latter’s orbital mean motion is less than Mars’s spin rate \(\Omega\) (i.e., Deimos is beyond synchronous orbit, that orbital distance where \(n = \Omega\), but Phobos lies within this distance). This
condition means that the time delay between stress and strain that inevitably occurs during any inelastic flexing allows the maximum tidal bulge on Mars due to Deimos to rotate out from under the satellite (i.e., to precede the satellite's orbital longitude; see Fig. 3). The lack of a detectable deceleration for Deimos, as apparent in the orbital position measurements, is consistent with this tidal model since, owing to the outer satellite's smaller size and greater distance from Mars, Eq. (1) would predict that Deimos' deceleration should be < 1% of Phobos' acceleration. Note that Shoemaker (1988) does not comment on the sign of his solution for Deimos, leading one to suspect that his value, which is listed as positive and which therefore would be contrary to the theoretical prediction, contains a typographical error.

Equation (1) can be integrated to find how distant the two satellites were from Mars 4.6 Gyr ago, presuming that their orbits have always been circular. According to this straightforward but simple-minded approach, Phobos originated at 5.7 \text{ R}_d\ (\text{i.e., its orbit has undergone major change through tidal action}) while Deimos, now at 6.92 \text{ R}_d, would have moved at most a few tenths of a \text{ R}_d from Mars (i.e., tides are totally ineffective in altering its path).

Orbital eccentricity \( e \) may modify this conclusion significantly, at least for Phobos. Two kinds of tides are important in \( e \)'s evolution: tides induced in Mars due to Phobos' tug on the planet as well as tides produced in Phobos both due to its rocking (libration) about its synchronous rotation state (Yoder 1982; Burns 1986a, Fig. 1d) and due to its changing radial distance from Mars as the satellite moves along its elliptical path. Both types of tides cur-
rently circularize this satellite’s orbit and thus the small moon’s path may have been very elongate many years ago. Starting from Phobos’ present eccentricity and naively integrating into the past, Phobos’ \( e \) would grow to \( \sim 0.1 \) to 0.2 about 4.6 Gyr ago owing to planetary tides alone (Singer 1968); tidal dissipation in the satellite would account for more evolution, indicating that the original orbit had yet higher eccentricity. Furthermore, an orbit with greater eccentricity can evolve even more rapidly in \( a \) than a circular one because, if \( e \) grows swiftly enough as we integrate backwards in time, the orbit’s pericenter may approach Mars despite the orbit’s overall expansion. With a closer approach to Mars, events at pericenter dominate \( a \)’s evolution, due to the steep radial dependence of tidal effects (see Eq. 1), such that \( a \) can actually accelerate even as Phobos withdraws from the planet.

The precise historical track followed by the satellite depends, of course, on the amount of dissipation in the planet as compared to that in the satellite (since tides in the two bodies affect \( \dot{a} \) and \( \dot{\varepsilon} \) differently) and on the particular model chosen to represent each of these dissipations (e.g., linearity or nonlinearity with strain, and independence or dependence on forcing frequency). Typical orbital evolution histories resulting from tides in both Mars and the satellite under study can be found in Lambeck (1979b), Cazenave et al. (1980), Mignard (1981) and Szeto (1983); a few such paths are shown in Fig. 4. All these authors reach a similar conclusion: under the action of tides alone, Phobos’ orbit could be transformed from a primordial large elongate path to the much smaller, nearly circular one seen today. In view of this past orbital history and the previously discussed physical evidence, it then seems as though Phobos’ origin is quite simple to discern: the satellite is a captured asteroid. Unfortunately, while this inference may be correct, it cannot be reached so directly for reasons described below.

Deimos’ orbit, as already mentioned, expands hardly at all due to tidal action. For similar reasons, primarily the satellite’s remoteness from Mars, Deimos’ eccentricity, currently small, also has changed little. Thus, by the orbital criteria that were just invoked to argue that Phobos was surely a captured object, clearly Deimos is not, because it has always had an orbit much like its current one. Accordingly, the issue of the origin of the Martian moons becomes less clear-cut and, for other reasons to be mentioned presently, it is even more muddled. Before getting into these complications, let us consider the relevant topic of the past histories of the inclinations.

Orbital inclinations are most conveniently measured with respect to the pole of the local Laplace plane, that plane about which a satellite orbit precesses; for distant moons (those principally perturbed by the Sun’s gravity), the Laplace plane is very nearly the planet’s heliocentric orbital plane but, for close satellites (those mainly affected by the planetary oblateness), it is very nearly the planet’s equatorial plane (Burns 1977). The orbits of satellites maintain approximately constant inclinations even as the planet precesses under solar and planetary torques or as the orbit evolves, as long as these
Fig. 4. The Martian satellites’ orbital histories as caused by tides in Mars and the satellites. The different curves correspond to different laws for $Q$’s dependence on tidal frequency, in the planet and satellite. Szeto (1983) considered a total of 15 separate cases, of which we show only 3. Curves 1 and 4 have no dissipation in the satellite; for curve 1, Mars’ $Q$ is proportional to tidal frequency, while for curve 4, it decreases with increasing tidal frequency. Curve 7 has $Q$ in both planet and satellite increasing with frequency. (a) Phobos. Dots on curves 1 and 4 are time markers that are spaced at 1 Gyr intervals, while those on curve 7 are separated by 50 Myr. The evolution time scale is set by the absolute value of $Q$ chosen, which is fixed by the observed secular acceleration. For all cases, Phobos’ current orbit is seen to have evolved from a larger, more eccentric orbit but with little movement in inclination off the Laplace plane. (b) Deimos. Dots on curve 1 are separated by 10 Gyr, the triangle on curve 4 by 1000 Gyr; the points on the inclination history are spaced by 10 Gyr. Deimos’ orbit is basically unchanged.
changes occur slowly and as long as these inclinations are measured relative to the Laplace plane (Goldreich 1965). In the specific case of Phobos, which as noted above can develop large eccentricity under tides, Cazenave et al. (1980; see also Szeto 1983) and Mignard (1981) constructed formulations that allowed the inclination of a highly eccentric orbit to be followed relative to its Laplace plane. As expected from Goldreich's analysis, these authors all determined that the two satellites' orbits—each about a degree from the local Laplace plane today—stay close to it throughout the evolution, meaning that the satellite's currently near-equatorial tracks, when evolved back in time, switch fairly abruptly at about 13 R_e onto a trajectory near Mars' orbital plane.

Thus, if one were arguing that the Martian satellites were captured objects, the fact that their ancient orbital planes may have been close to the ecliptic, near which many solar system objects presently move, might at first seem like strong support for the capture model. However, this notion has two serious failings. First, if capture took place at 13 R_e or beyond, the amounts of energy that initially bind the protosatellites to Mars are very small. In other words, the capture process would have to be very finely tuned such that the orbits of the soon-to-be-acquired objects would have to closely duplicate Mars' path. But, if the protosatellite-cum-asteroid had a Mars-like orbit, then the original motivation for wanting to capture Phobos and Deimos (namely to explain the apparent differences between their physical properties and those of Mars) is lost. Indeed, if an object were sent along an elliptical orbit directly from the middle of the asteroid belt to the vicinity of Mars, its velocity at Mars would differ from the planet's by nearly 10 km s^{-1}, an amount that could not be removed by a thin nebula. The second failing of the intuitive idea that low inclinations imply capture is that the actual paths along which the objects would move following capture are defined by the relative velocities at which they approach Mars and not by the plane in which they travel prior to their interaction with the planet. For capture at a great distance to be a viable model, it must occur at low approach speed for reasons just mentioned above, but, for an observer on the planet, such small relative velocity vectors are rarely confined to near the planet's orbital plane but instead are almost randomly oriented. Of course all these problems may be alleviated if an unmodeled process, e.g., nebular drag, significantly modifies the orbit following capture. And it is likely that nebula drag will reduce orbital inclinations.

Earlier it was maintained that Deimos' nearly fixed orbit is persuasive evidence against the capture of the outer Martian satellite. It also weighs against Phobos once having a sizeable elongate orbit, which is the principal dynamical point in favor of the inner satellite's capture. This latter restriction occurs because, if Phobos had an orbit with both a and e large, its trajectory would have crossed Deimos's for an extended period of time prior to tidally evolving inward past it. During the interval when the orbits of the Martian
satellite pair were interlaced, mutual collisions would have been inevitable as the characteristic collision time (∼ 10⁴ to 10⁵ yr) is much briefer than any plausible orbital evolution time (Lambeck 1979b; Cazenave et al. 1980; Szeto 1983). In this way, Deimos’ orbit represents a dynamical barrier that firmly limits the total evolution of Phobos. This constraint might be avoided in several ways: (i) the orbits of both moons evolved very swiftly at the outset but then slowed down, perhaps with this evolution history being due to gas drag from a nebula that subsequently, and rapidly, dissipated; (ii) the satellites were captured sequentially with Phobos evolving out of the way before Deimos was acquired; or (iii) the Martian moons originated together as the largest fragments of a catastrophic collision (see below). None of these suggestions can be disproved but, if any were true, then there would be no possibility of following the histories and at least a few planetary scientists would be out of (part-time) work.

The conundrum that Phobos’ contemporary orbit suggests a capture that is not permitted according to Deimos’ path may be avoided altogether. It is possible that Phobos’ present orbital eccentricity (that which, although slight, drives the relatively rapid evolution) is not a tidally damped remnant of an earlier larger eccentricity but rather, at least in part, has been caused by other, more recent events. In this regard, Yoder (1982) has shown that, depending on its evolution, Phobos may have passed through three gravitational resonances at 2.9, 3.2 and 3.9 R₉, due respectively to the 3:1 resonance between the satellite’s orbital period and Mars’ spin period (meaning that the planet’s lumpy gravity is experienced at the same inertial position on every third orbit), the 2:1 resonance between the precession period of the satellite’s orbit and the planet’s orbital period, and the 2:1 resonance between the satellite’s orbital period and Mars’ spin period. On passage through each of these resonances, Phobos’ orbital eccentricity and inclination undergo jumps, whose precise values depend upon Mars’ eccentricity, obliquity and phase; related but more complex resonant interactions have been invoked by Tittermore and Wisdom (1988) and by Dermott et al. (1988) to account for Miranda’s anomalously high inclination today. If such jumps have occurred, the values of Phobos’ e and i measured currently are the tidal remnants of their values after the last resonance was passed rather than values damped from the formation epoch; as such, they should not be used as starting points for integrations into the past. Indeed, since Yoder (1982) found that the magnitudes of the jumps were sensitive functions of the phase at which the resonance was passed, and because the rate of a’s evolution at any time depends so critically on how eccentric Phobos’ orbit is at that time, the ancient histories of the Martian satellite orbits become untraceable.

Resonances have a place in Deimos’ history as well. Yoder (1982) has shown that the outer Martian moon should have an eccentricity of 0.002 if it passed through a 1:2 mean motion resonance with Phobos during the latter’s evolution. As the measured eccentricity is considerably less (0.0002), and the
time scale for tidal damping of Deimos' eccentricity is much longer than the age of the solar system, this implies that the outer moon must not have gone through the 1:2 resonance. However, if such a resonance passage did not occur, then passages by Phobos through its resonances could not be invoked to account for its eccentricity; the only ways out seem to be to claim that the Martian satellites each has distinctly different physical properties or to hypothesize that Deimos had undergone a propitious impact that just so happened to circularize its orbit. Wisdom (1987a) has suggested an alternative solution, namely that Deimos' small orbital eccentricity is a consequence of an episode of chaotic tumbling. He demonstrates that such an interval of chaotic rotation is likely for all irregularly shaped satellites and that, as a result, tidal damping will be much more effective in reducing the eccentricities of such satellites. The same process may have hastened the circularization of Phobos' path.

Much progress has been made in the last few years in understanding the orbital evolution histories of the Martian satellites. This added knowledge has reinforced the opinion of most dynamicists that the orbits of Phobos and Deimos can be explained most directly as being those of objects that originated in situ. However, recent studies of the evolution of particles through more realistic circumplanetary nebulae have the promise of allowing outer solar system objects to be captured by Mars and yet have their orbits transformed into the regular ones seen today.

IV. CONCLUSIONS

Where does this leave us? Many years ago Burns (1978) wrote that the pendulum of scientific opinion at that time seemed to be moving toward the idea that the Martian satellites were captured asteroids, but cautioned that pendula have a way of swinging back. The situation has not changed significantly since then: we have more data and better crystallized ideas but they lead in different directions. Had the Phobos spacecraft been able to complete its mission and given a firm indication of the satellite's composition, perhaps we would now know which line of evidence (crude physical measurements or incomplete dynamical theory) led to the correct past. But such was not to be and so we continue to make conjectures.

Acknowledgments. I appreciate the warm hospitality of the Lunar and Planetary Laboratory of the University of Arizona, where this chapter was written while I was on sabbatical leave from Cornell. I thank A. R. Dobrovolskis, W. K. Hartmann, H. H. Kieffer, S. J. Peale, S. Sasaki, C. W. Snyder and P. Thomas for illuminating comments on the manuscript. I especially appreciate the help that I received from J. Bell, D. Britt and S. Murchie in interpreting the spectral and mineralogical data. This research was supported by NASA's Division for Solar System Exploration.