Introduction

The question of how the Moon came to be has probably been pondered since prehistoric times. The origin of Earth, Moon, and the heavens is the subject of many creation myths. Egyptian mythology endows the Moon, as the deity Thoth, with the role of arbiter between the gods. Thoth came to be associated with wisdom and, appropriately, astronomy. Greek mythology associates the Titanian goddess Selene with the Moon and the ancient Roman culture associated Selene with the god Luna from their pantheon. Both cultures leave us with modern astronomical terms used in conjunction with the Moon, such as lunar or selenology (study of the origin of the Moon). While the ancients did come to grips with the nature of the Moon’s motions, the existence of the Moon was largely taken as a divine creation. Such thinking persists today in religious circles.

Until the European Renaissance the origin of the heavens remained a divine topic. With the emergence of new tools, thinking, and processes, explanation of the Earth and Moon’s origin became a more scientific pursuit. The subsequent explosion of science lead to a series of lunar origin hypotheses that are the topic of the following discussion.

Three Scientific Positions

The origin of the Moon remained in the realms of mythology until the 17th century. In 1610, Galileo Galilei pointed his newly constructed telescope toward the Moon and discovered a rugged world of mountain ranges, vast plains, ridges, and craters. Galileo also found similar objects orbiting Jupiter: perhaps our Moon was one among many.

This was the start of the modern scientific explosion. The solar system was becoming a Sun-centred arrangement of planets each, potentially, with moons of their own. Kepler’s insight into the nature of orbits, and Isaac Newton’s laws of motion and gravity provided the explanation of a clockwork
cosmos. From this intellectual beginning the notion that the presence of the Moon might be explained in terms of science was born.

Three scientific theories of lunar origin were proposed by philosophers and scientists during the centuries leading up to the space age: the Moon and Earth formed in unison, the Moon formed elsewhere and was captured or, the Moon was a part of Earth spun-off. These theories are expanded on in following sections.

Parallel Formation

René Descartes in his attempt at a model for the universe, *Le Monde* (1664, written circa 1630) and *Principia Philosophiae* (1644), proposed that stars, planets, and moons formed in vortices caused by the motion, and wake, of material that existed in the universe. Lumps of material could move about in the vortices in a manner governed by their weight, with lighter masses travelling faster. The material coalesced over time, with large enough pieces developing smaller vortices of their own and either forming or capturing smaller lumps that strayed too close. Over time, the material in vortices about planets would merge into moons. Descartes envisaged this as the mechanism by which the Moon formed in orbit about the Earth. Students of modern astronomy will immediately recognise a parallel, in gross terms if not physics, between this model and the solar nebula hypothesis.

The modern solar nebula hypothesis, descended from Laplace’s work, has the Sun and solar system forming from a large, cold, slowly rotating cloud of gas and dust. Collapse under gravitation and conservation of angular momentum results in a more rapid spin rate and flattening of the gas-dust mixture into a disc. The star forms in the densest, central region, and the precursors of planets in smaller gravitationally unstable regions in the outer reaches. Large planetary precursors themselves attract a disc of material from which rocky moons condense. Ignition of the central star results in a solar wind that clears much of the remaining gaseous material from the inner reaches of the solar system leaving rocky remnants to coalesce over time to form the planets and their moons. In the case of the Moon, this theory would contend that it formed in-situ, out of the same stuff as the Earth, and in the same manner.

One variant of the co-formation theory was put forward by Evgenia Ruskol during the 1960s and 70s. The hypothesis assumed that a substantial swarm of small objects, mostly less than 100 km, formed in the environs of Earth. The swarm extended some 200 Earth radii into space and was topped up by debris gathered from the region between 0.8–1.3 AU. Ruskol modelled lunar formation as accretion of these swarm particles by inelastic collisions rather than the more commonly used purely gravitational arguments.

There are weaknesses in the co-formation hypothesis that had been known for a long time. From Newton’s gravitational law, physical size,
and orbital characteristics of the Moon is it possible to determine the average density of the lunar mass. Calculations of the density gave a figure in the order of 3.3 kg m$^{-3}$ for the Moon. A comparable figure for the Earth is 5.5 kg m$^{-3}$, primarily due to the presence of a dense core beneath the Earth’s mantle. In contrast, the Moon’s density is roughly comparable to the Earth’s surface (mantle) material alone.

**Fission Theory**

During the nineteenth century Julius Robert Mayer determined that the orbital period of the Moon about Earth was very slowly increasing because of an interaction with oceanic tides (an idea that had been languishing for a century). The implication of this is that the orbit is slowly growing and, by extension, the Moon was much closer to Earth in the distant past. The reason for the Moon’s recession is related to the relative location of Earth–Moon line and the tidal bulge induced in the Earth’s body and oceans by the Moon’s presence (Figure 1). The Earth’s rotation pulls the tidal bulge forward of the Earth–Moon line imparting a gravitational torque on the Moon, lifting it into a higher orbit. The same torque is responsible for slowly retarding the rotation of the Earth. Consequently, when the Moon was closer to Earth, Earth’s rotation rate was higher.

![Figure 1: Tidally induced changes in the Moon’s orbit and Earth’s rotation.](image)

George Darwin (1845–1912), son of Charles, reasoned that by extending far enough into the past one would reach a point where the Earth and Moon were in contact. In his 1877–78 papers, Darwin postulated that the Earth may have actually been the source of the material that made up the Moon: essentially spawned from the proto-Earth. The proposal required that the early Earth be a homogeneous blob of viscous material that was spinning quickly, with a period of approximately 6 hours. Darwin’s calculated spin rate was not, of itself, sufficient to make the proto-Earth unstable. Darwin realised this point and constructed a credible trigger for the splitting of the primordial Earth into two by combining a fast spin with internal oscillations and resonant tidal effects of the Sun. The two blobs, one much smaller than the other, would go on to solidify and become the Earth and Moon at their
Figure 2: Fission Theory. (1) The early Earth is spinning rapidly and (2) flattens, (3) becomes unstable, (4) critically asymmetric, and (5) spawns a moon. Small residual droplets fall back to Earth or the Moon, (6) which moves outward to its present location.

present separation. Figure 2 shows the stages of spawning the Moon.

An article in Nature by Robert Ball (1840–1913), and the ensuing correspondence, served to popularise Darwin’s theory. The possibility that the Pacific and other ocean basins were the partially healed scars left by the spawned proto-moon (Osmond Fisher, 1882) became popular with Earth scientists opposed to Wegener’s theory of continental drift, who were quite happy to use this to argue their position. This acceptance of the theory outside of astronomy circles lent reasonable credence to the argument.

The fission theory lingered well into the twentieth century. The Russian mathematician A.M. Lyapunov (1857–1918) had gone some way to undermining one of the pillars on which Darwin’s fission theory relied, the behaviour of rotating viscous fluids. American astronomer Forest Moulton (1872–1952), in 1909, used Lyapunov’s work to argue that the fission of Moon from Earth was simply not viable [14]. By 1930, when Jeffreys’ [10] wrote, “It has always been recognized that the validity of this theory depends on a smallness of friction . . . can now be estimated; and it turns out to be sufficient to invalidate the theory.” the theory was largely ignored. Nevertheless, fission theory had a small resurgence in the years immediately preceding the space-age exploration of the lunar surface.

Wise [20] was a defender of the rejuvenated fission theory during the 1960s. Using modern Earth science Wise argued that the trigger for fission did not come from outside, as in Darwin’s original, but rather from the formation of the Earth’s core. While this position did receive support from Ringwood, Cameron, and O’Keefe there remained problems. Critics of the position continued to point out that internal sources alone were insufficient to make the Earth unstable in rotation, while others pointed out that the mass ratio of Earth–Moon was far larger than a fission approach would deliver.
Capture

The theory of lunar capture came about as an alternative to the other options in the early part of the twentieth century. The basis of this hypothesis, explored by Thomas See (1866–1962), is that the Earth captured a near fully formed object after it strayed too close (see [17], [18]). This possibility requires that the captured body form somewhere else in the solar system. See favoured formation of the Moon in the outer solar system, around the orbit of Neptune, following the thoughts of Leonhard Euler (1707–1783) who believed that orbits were constantly shrinking through resistance of passing through the æther. Other commentators suggested a closer origin of the Moon.

Regardless of the original location, the object’s orbit must be somehow perturbed into an Earth crossing orbit to allow for capture by the Earth. Apart from the planets there were no obvious candidates with substantial perturbing power. Even if an object were dislodged from an outer solar system location a plausible mechanism by which a Moon-sized body could be captured into Earth orbit would be required; none was forthcoming. With these objections, and See’s failing reputation, the capture theory was largely passed over by his peers.

Like the fission theory, the capture theory spent some time in the wilderness until it was revived in the 1950s. Gerstenkorn [7] proposed that the Moon was captured into Earth orbit from elsewhere. Unusually, Gerstenkorn’s hypothesis had initial capture into an inclined, retrograde orbit around Earth. He then provided a mechanism to increase the inclination until it flipped over to the modern prograde orbit. The possibility of lunar capture was discussed by Urey around 1960 concluding that the need to dissipate energy by tidal or collisional means made it unlikely, but that the odds of successful capture were bolstered by an abundance of candidate objects. Discussions also arose about the likely composition given the long held assumption that the lunar density indicated an iron-poor body. Urey compared the solar abundance of non-volatile elements to that of the Moon and concluded that the Moon formed elsewhere in the solar system. The capture theory remained viable if a little shaky.

Evidence Is Gathered

At the start of the space age, the three main theories for lunar origin were well established. While there were objections to each option there was a general feeling that the issue would be settled by visiting the Moon on sample return missions. Samples could not prove a case, but they could disprove two, leaving one that was still consistent with the evidence. Both the Union of Soviet Socialist Republics (USSR) and the United States (US) planned and executed lunar missions with the intent of exploring the Moon
and returning samples. Much of the activity was politically driven, collecting space ‘firsts’ at the highest priority, with science a poor second.

The US Surveyor and USSR Luna missions explored the Moon in a limited way by means of fixed and mobile robotic missions. Samples were analysed in-situ, with the exception of Luna 16, 20, and 24 which returned 301 grammes of material to Earth between 1970 and 1976. These samples were limited to regions in which it was ‘safe’ to land a robotic vehicle. The US Apollo programme resulted in six manned landings on the surface of the Moon between 1969 and 1971. Apollo 11 was severely time constrained, risk averse, and focused on political goals, but the later Apollo missions had increasing science focus culminating in Apollo 17 carrying a professional geologist (Harrisson Schmitt). The Apollo missions returned 382 kilogrammes of material from different lunar environments: mares, highlands, etc. The Apollo and Luna samples were subject to far more comprehensive analysis than was possible in a self-contained spacecraft.

The first lunar samples from Apollo were subjected to a coarse initial assessment followed by more complex and sensitive assays. The Lunar Sample Preliminary Examination Team (LSPET) was charged with the early assessments. Lunar rock, as distinct from surface dust, was found deficient in metals such as nickel and cobalt in comparison to Earth. The surface dust was treated differently as it was expected to contain meteorite debris. Some materials, notably gold, were extremely uncommon in the rocks in relative terms. Volatile metals such as sodium, potassium, and lead were also low in abundance, a fact that would be consistent with high temperature processes. The ratios of iron to nickel and cobalt were also found to be different to those of chondritic meteorites, which are understood to be good indicators of the primordial solar system composition. Later, more sensitive, assessments determined very low levels of gold, silver, bismuth, thallium, and iridium. Without exception the Moon’s rocks were found to be devoid of water and minerals that require the presence of water.

Isotopic analysis of lunar rocks showed that the $\text{^{18}O}$ and $\text{^{16}O}$ ratio was close to that of terrestrial rocks. According to Onuma, Clayton, and Mayeda [15] this indicates a formation at the same distance or a little closer to the Sun than the Earth because meteorite data indicate that $\text{^{18}O}$ to $\text{^{16}O}$ ratio declines with distance from the Sun.

Some lunar samples betrayed the presence of a surprisingly strong magnetic field some 3.7 billion years ago. This is generally consistent with a magnetic dynamo effect in a molten iron core. Nevertheless, there is no substantial magnetic field today, so if the core is present then it has ceased its dynamo activity. Apollo spacecraft were carefully monitored while in orbit to detect the presence of variations in mass concentration of the Moon: some were found, but mostly at shallow depths. Seismic sensors deployed on the surface by Apollo 12, 14, 15, and 16 revealed no large core body as they would on Earth. In fact, beyond the outermost fractured layers, the
Moon’s mantle is remarkably uniform, which is consistent with estimates of the Moon’s moment of inertia.

Much to the chagrin of planetary scientists of the time, the results from Apollo did not neatly disprove two of the three lunar origin contenders. The lunar sample analysis effectively discounts all three lunar genesis theories for one reason or another.

The bulk similarity between terrestrial and lunar rocks, coupled with the difference between lunar and primordial meteorite material effectively discounts the possibility that the Moon formed elsewhere in the solar system. Ringwood and Essene, at the Australian National University, argued that the lunar rocks showed important similarities to terrestrial rocks, based on non-volatile, oxygen-loving element abundances being within Earth-like ranges. They also concluded that the co-formation and capture hypothesis were inconsistent with the Apollo data. The lunar capture theory of See could therefore not be accurate with his preferred origin location. Formation in the region surrounding Earth was possible but still left the problem of capture.

The radical depletion of volatile materials from the lunar basalts, bulk similarities between Earth and Moon, insomuch as refractory materials are concerned makes the concept of in-situ co-formation very difficult. Two bodies forming from the same material cloud should have similar compositions and construction. That the average density of the Moon is well below that of the Earth has long been known and assumed to be the absence of a dense core. The absence of a substantial core body in data from the Apollo and later missions, such as Clementine, confirmed the belief and increased the barrier to acceptance of parallel formation.

Aside from the problem of providing the impetus for a fission event, the fission theory has problems with relative depletion of volatiles, water, and water-related minerals from the lunar crust. Volatile depletion is indicative of the presence of high temperatures that were not present on the primordial Earth. Water free geology is also consistent with high lunar temperatures. In a fission scenario the temperatures would be largely uniform between the Earth and spawned Moon.

In the wake of the lunar sample results proponents of the pre-landing theories scrambled to reconcile their favourite theory with the observational evidence. The fission theory moved from spawning a single mass to spraying a large number of globules into equatorial orbits, from which the Moon was to form. Capture theorists started looking for nearby places to form an object to capture, and reasons for the compositional differences. To varying degrees the attempts were successful but, for the most part, were not entirely convincing.
Giant Impact

While the rush to revise old theories continued, Hartmann and Davis were developing an alternative hypothesis that attempted to overcome the limitations. The lunar cratering rate, which was known to be very high in the early solar system history, provided a degree of inspiration for the work. Hartmann and Davis published a paper in 1975 [9], providing an analysis of the likely size of other planetesimals in the region around Earth-sized terrestrial planets during the period of lunar bombardment. They found that 500 to 3000 kilometres would be the likely diameter range for the next-largest bodies, with bodies around 100 kilometres numbering in the tens. The analysis continues to determine that the impact of a body of around 1200 kilometres diameter at 13 km s$^{-1}$ could put enough material to form the early Moon into orbit.

Cameron and Ward were, independent of Hartmann and Davis, working on a similar theory [5]. While sharing some characteristics with Hartmann and Ward, this theory relied on a single large impact to provide the lunar material. In Brush [3], Cameron requested that emphasis be placed on the key difference, and is quoted as, “The HD [Hartmann-Davis] idea involves many collisions and, therefore, is fatally slain by angular momentum.” Their paper is a short outline of the bulk behaviour of an impactor, of roughly Mars size, striking a tangential blow to the early Earth with sufficient momentum to account for the Earth-Moon system’s current state (Figure 3). Both the early Earth and impactor are assumed to be differentiated into mantle and
core, and possibly molten. At impact the impactor is be divided into roughly equal portions. The inner half suffers impact at approximately 11km s$^{-1}$, vaporising the majority of mantle material in the impact region and ejecting it forward at substantial velocity into orbit. The outer half of the impactor is expected to shatter into many smaller pieces travelling on ballistic trajectories that would eventually coalesce into the Earth. The metallic core of the impactor, largely unaffected by the vaporisation would plunge into the Earth and eventually merge with the metallic core. Condensation of the vaporised material ejected from the mantle regions, followed by collision-driven circularising of orbits would leave the Earth with a disc of material between 2R$_{\oplus}$ and 4R$_{\oplus}$. It is from this material that the Moon would form by accretion, outward transfer of angular momentum, and gravitational instability.

Having established the skeleton of an alternate lunar genesis, the details needed to be determined. An obvious choice was computer simulation, however, in the 1970s computing power was simply not adequate for anything but the roughest estimate. The situation has changed significantly in the intervening years, and now simulation forms a pivotal part of the the giant impact theory. By the mid 1980s realistic simulation was becoming viable, with Kipp and Melosh publishing results on 2-dimensional [11] and later 3-dimensional [12] simulation of likely impacts (but not being able to simulate gravitational attractions). Kipp and Melosh concluded that the coarse simulations generally supported the idea of a giant impact by a Mars-sized body providing sufficient material to form a moon. More complex 3-dimensional “smoothed hydrodynamic” simulation by Benz, Slattery, and Cameron [2] followed the impact by simulation of self-gravitating masses. These simulations clearly supported the impact model as a source of material for the Moon.

From the outset, the difficulties in forming a single body from a circumplanetary disc were recognised. The problem is dominated by the Roche limit of the Earth for mantle-like material ($\sim 2.9R_{\oplus}$), which a substantial portion of material from the Cameron and Ward model would fall inside. Within the Roche limit the material will not generally clump under the influence of gravity, eventually falling back to Earth. Material that made it out to the Roche limit would be free to coalesce under the influence of gravity, building larger objects, until one mass dominated and cleaned up the remaining material. In 1996, Canup and Esposito [6] determined through simulation that formation of a single, large moon under would be difficult without several contrived conditions including a particular density profile and that the largest moonlet should form in the inner reaches and migrate outward, accreting smaller moonlets as it went. They concluded that higher resolution simulation was required to better explore the issue. More recent works by Cameron and Canup [4] explored the possible range of parameters for a giant impact at higher resolution. The ratio of proto-earth to impactor was fixed at 7:3, but the total mass and angular momentum were varied.
looking for the optimal pre-lunar mass. The study concludes that the optimum condition, at least for selected mass ratio, results when the mass of the proto-earth is only half that of the modern Earth.

Discussion

A violent impact scenario can explain some of the features of lunar rock composition. The impact debris field around Earth would be hot; volatile materials would rapidly be boiled away, leaving a higher concentration of refractory materials consistent with lunar rocks. By assuming the impact event occurred after the differentiation of impactor and Earth the troublesome lack of an iron core in the Moon was neatly explained by the tendency of the impactor core to plunge into the Earth’s core. The material that would eventually form the Moon would therefore be iron deleted and of roughly the density of the Earth’s mantle material. Simulations have confirmed, to the limit of their accuracy, that the scenario is viable.

Angular momentum is a major problem in fission and capture theories: fission requires extra momentum and capture requires the dissipation of excesses. Brush [3] calls the tendency of lunar origin theories to fail on momentum grounds, “angular momentum disease.” The collision model has the distinct advantage over antecedent theories in that the impactor could carry just the correct amount of momentum into the system to account for the current angular momentum. As an added benefit, such a one-off stochastic event might also explain the axial tilt of the Earth. That said, in practice it has been difficult to construct computer simulations that provide just this scenario.

Simulations have shown that formation of a moon from vaporised impact debris is possible. One major outstanding question with the models remains: where did the impactor come from? Origin in the outer solar system, as preferred by See for his capture object, was clearly not viable on the grounds that composition, particularly oxygen isotopic composition, was not correct. Onuma et al. [15] concluded the object must have formed at or slightly inside the Earth orbit. However, anything forming physically near the Earth would likely not be stable. Belbruno and Gott [1] in a paper submitted to Astrophysical Journal provide a lengthy exploration of the possibility of a large impactor forming in the regions surrounding the co-orbital Lagrange points (L4 & L5). They conclude that, not only is formation possible, but that realistic perturbations could put the object onto a horseshoe orbit that ultimately collides with the Earth. Once the object leaves the L4 (L5) point impact occurs within a short time frame and the impact is very close to parabolic (as used in the successful simulations). Should the article survive scrutiny, the major outstanding piece of the giant impact theory may be put to rest.

The impact theory did not convince all interested parties overnight. For
example, in 1987, Stevenson [19] assessed that there were still doubts that the collision theory was strong enough to displace accretion, fission, or capture. Stevenson’s assessment was made before the majority of the simulation work was possible. Ringwood, previously a supporter of fission, published a paper in 1989 [16] disputing the single impact variation on compositional grounds and effectively siding with the many impacts variation of Hartmann and Davis. Nevertheless, the theory had a fairly solid start and has grown more solid with time as simulation and understanding have accounted for all the necessary elements.

Study into the other possibilities continues to the present day. For example, the 2001 paper of Morishima and Watanabe [13] discusses two scenarios for the co-accretion of the Moon, with particular emphasis on reduction of the angular momentum problem. Gorkavyi [8] poses another variation on the co-formation theme that attempts to account for the iron-poor lunar geology. That different views continue to exist is not surprising but they are very much a minority.

**Conclusion**

The origin of the Moon prior to the age of modern science was typically attributed to divine causes. The advent of scientific discourse lead to the formation of three competing theories over the period of approximately one hundred years. Co-formation dictated that the Moon formed in-situ from the same stuff as the Earth, capture theory had a fully-formed Moon captured in a close encounter with the Earth, and fission theory spun the Moon off from a rapidly spinning early Earth. Each theory had an initial burst of support that petered out, only to be revived when the possibility of testing the theories became reality. Each contender also had failings, generally in terms of angular momentum or composition, but these were overlooked or assumed to be solvable in some fashion.

Lunar samples and other data provided information that did not eliminate any of the three theories, but did not concretely support them either. Composition of lunar rock was similar to Earth, but sufficiently different to be used as an argument against each possibility. In light of the evidence, substantial investigation and revision of theories was undertaken with limited success. Some of this revision continues today.

Meanwhile, a new theory in which the Moon forms from impact debris out of a single impact event was being formed. Substantial effort has been expended to flesh out the detail of the theory. The success of the giant impact theory in simulation, and in explaining the compositional similarities and differences with the Moon, has made it the theory-of-choice for most planetary astronomers. It has achieved this without invoking any special circumstance or contrived initial conditions. The theory has evolved in detail
from the starting points offered by Hartmann, Davis, Cameron, and Ward but has not lost its essential characteristics.

References